On the Faithfulness of Language Model-based Data-to-Text Generation

by

Joy Mahapatra

under the supervision of

Professor Utpal Garain



A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy in Computer Science

Computer Vision and Pattern Recognition Unit
Indian Statistical Institute Kolkata

November 2026

Declaration of Authorship

I, Joy Mahapatra, declare that this thesis titled, 'On the Faithfulness of Language Model-based Data-to-Text Generation' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Joy Mahapatra.

Date: November 2026

"Stay Hungry, Stay Foolish."

Steven Paul Jobs

Abstract

Data-to-Text Generation (D2T)—the task of transforming structured or semi-structured data such as tables, graphs, and meaning representations into readable and faithful natural language—has advanced substantially with the advent of language modeling, particularly with Large Language Models (LLMs). Despite remarkable improvements in language model-based D2T systems in terms of readability, maintaining faithfulness—defined in terms of informativeness and factual consistency—remains a persistent challenge. This thesis presents a systematic exploration of faithfulness in language model-based D2T, focusing on both its evaluation and enhancement, with particular attention to influential factors such as source–reference divergence.

The thesis first introduces a structure-aware encoder-decoder framework, based on pre-LLM architectures such as Long Short-Term Memory (LSTM), which encodes input data according to its record-field hierarchy and enhances informativeness through selective gating and hierarchical encoding. It then transitions to LLM-based D2T, conducting a large-scale evaluation across multiple D2T benchmarks and LLM families. The study identifies source-reference divergence—the semantic mismatch between input data and reference text—as a primary factor influencing readability, informativeness, and overall faithfulness. To address this issue, a divergence-weighted training approach is proposed, employing an adaptive loss function that significantly improves model performance. Focusing specifically on factual consistency, the thesis presents a detailed evaluation of widely used LLMs using both automatic and human assessments, with improvements quantified through the Average Rate of Change. The findings demonstrate that source-reference divergence is a major contributor to factual inconsistency. To mitigate this effect, a divergence-aware selective generation framework is introduced, enabling models to abstain from producing factually inconsistent outputs under conditions of high divergence, thereby improving factual consistency without substantial coverage loss. Finally, empirical and mathematical analyses of scaling behavior reveal that factual inconsistency in D2T follows an exponential—rather than a power-law—scaling with respect to model size, compute, and data volume, an observation validated through a statistical framework, *VaCScal*.

Collectively, the findings of this research provide both conceptual and empirical insights into faithfulness in language model-based D2T, laying a foundation for the development of future models that are not only readable but also factually faithful.

Acknowledgements

I would like to express my deepest gratitude to my Ph.D. supervisor, Prof. Utpal Garain, for his invaluable guidance, encouragement, and unmatched support throughout my doctoral journey. His insightful mentorship and strong sense of direction have been instrumental in shaping my work and helping me identify a meaningful and practical focus for this thesis.

I am extremely thankful to all my friends in the NLP Lab for their effective collaboration and constructive discussions, which greatly enriched my research experience. I would also like to extend my warmest thanks to my friends in the CVPR Unit for their support during challenging times.

Finally, my heartfelt love and deepest appreciation go to my three sisters — BorDi, MejDi, and ChotDi — the true pillars of my life. Their boundless love, strength, and support have guided me through every step of this journey. Without them, none of this would have been possible, and I will remain forever grateful to them.

Publications and Preprints

The following publications and preprints are related to or aligned with the research presented in this thesis:

1. **Mahapatra, J.** and Garain, U., 2021. Exploring Structural Encoding for Data-to-Text Generation. *International Conference on Natural Language Generation (INLG)*, Association for Computational Linguistics.
2. **Mahapatra, J.** and Garain, U., 2026. A Comprehensive Performance Evaluation of LLMs for Data-to-Text Generation and Divergence-Weighted Training. *ACM Transactions on Intelligent Systems and Technology*, Association for Computing Machinery.
3. **Mahapatra, J.** and Garain, U., 2025. A Comprehensive Study on Factual Consistency in LLMs for Data-to-Text Generation. *Sādhanā*, Springer.
4. **Mahapatra, J.**, Roy, S. and Garain, U., 2025. Exponential Scaling of Factual Inconsistency in Data-to-Text Generation with Fine-Tuned LLMs. *Transactions on Machine Learning Research (TMLR)*.
5. **Mahapatra, J.** and Garain, U., 2025. Divergence-Aware Selective Data-to-Text Generation with LLMs for Factual Consistency. *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, Springer.

The following publications and preprints were completed during the Ph.D. period but are not related to the thesis work:

1. Saha, S., Chaturvedi, A.[†], **Mahapatra, J.**[†] and Garain, U., 2025. sudoLLM: On Multi-role Alignment of Language Models. *Empirical Methods in Natural Language Processing (EMNLP) Findings*, Association for Computational Linguistics.
2. Ghosh, K.[†], Singh, N.K.[†], **Mahapatra, J.**, Saha, S., Senapati, A. and Garain, U., 2024. SafeSpeech: A Three-module Pipeline for Hate Intensity Mitigation of Social Media Texts in Indic Languages. *Social Network Analysis and Mining*, Springer.
3. Pal, A.S., Biswas, D., **Mahapatra, J.**, Banerjee, D., Chakrabarti, P., Frangi, A.F. and Garain, U., 2024. Pathologist-like Explanations Unveiled: An Explainable Deep Learning System for White Blood Cell Classification. *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE.

[†]Equal contributions.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements iv

Publications and Preprints v

List of Figures x

List of Tables xv

Abbreviations xix

1 Introduction 1
- 1.1 Data-to-Text Generation (D2T) . . . 3
- 1.2 Language Modeling . . . 5
- 1.3 From Probabilistic D2T to Language Model-based D2T . . . 9
- 1.4 Performance Measurement of D2T Systems: Readability and Faithfulness 10
- 1.5 Evolution of D2T with Language Models and Their Performance . . . 13
- 1.6 Ensuring Faithfulness: A Key Challenge in Language Model-based D2T . 14
- 1.7 Objectives and Contributions of the Thesis . . . 14
- 1.8 Thesis Overview . . . 15

2 Enhancing Informativeness in D2T through Structural Encoding 17
- 2.1 Background . . . 17
- 2.2 Related Works . . . 18
- 2.3 Notations Used in This Work . . . 19
- 2.4 Approach . . . 19
  - 2.4.1 Structural Encoder . . . 20
    - 2.4.1.1 Field Encoding . . . 20
    - 2.4.1.2 Record Encoding . . . 23
  - 2.4.2 Decoder . . . 24
    - 2.4.2.1 State Update . . . 24

2.4.3 Copy Module . . . 26
2.4.4 Loss Function . . . 26
2.5 Experiments . . . 26
2.5.1 Baselines and Metrics . . . 26
2.5.2 Parameter Settings . . . 27
2.5.3 WikiBio Dataset and Results . . . 28
2.5.4 E2E Dataset and Results . . . 28
2.5.5 Analysis from Ablation Tests . . . 29
2.5.6 Analysis of Model with Shuffled Input . . . 31
2.5.7 Human Evaluation . . . 33
2.6 Conclusion . . . 34

3 Evaluating and Enhancing Faithfulness in LLM-based D2T 36
3.1 Background . . . 37
3.2 Related Work . . . 39
3.3 Preliminaries . . . 40
3.3.1 LLM . . . 40
3.3.2 D2T . . . 41
3.3.3 Source-Reference Divergence . . . 42
3.4 Models, Datasets and Experimental Settings . . . 43
3.4.1 Models . . . 43
3.4.2 Datasets . . . 44
3.4.3 Experimental Settings . . . 44
3.4.4 Potential Risks and Considerations of Data Contamination . . . 45
3.5 Performance Evaluations . . . 45
3.5.1 Performance Evaluations on Readability . . . 46
3.5.2 Performance Evaluations on Informativeness . . . 47
3.5.3 Performance Evaluations on Faithfulness . . . 50
3.5.4 Performance Evaluation through Human Assessment . . . 52
3.5.5 Discussion . . . 54
3.6 Robustness of Performance Evaluation with Respect to Fine-tuning and Decoding Strategies . . . 55
3.7 Performance Evaluation Amid Source-reference Divergence . . . 57
3.7.1 Results Amid Source-reference Divergence. . . . 58
3.8 Case Studies . . . 60
3.9 Source-Reference Divergence-Weighted Training . . . 61
3.10 Conclusion . . . 62

4 Evaluating Factual Consistency in LLM-Based D2T 64
4.1 Overview . . . 64
4.2 Related Work . . . 66
4.3 Preliminaries . . . 67
4.3.1 D2T . . . 67
4.3.2 Factual Consistency in D2T . . . 68
4.3.3 Source-Reference Divergence . . . 69
4.4 Experimental Setup . . . 69
4.4.1 Datasets . . . 70

4.4.2 Models . . . 70
4.4.3 Experimental Setting and Statistical Significant Testing Details . . 71
4.4.4 Average Rate of Change (AROC) . . . 71
4.5 Evaluation of Factual Consistency: Results and Discussion . . . 72
4.5.1 Results of Factual Consistency Evaluation Using SummaC-Conv 72
4.5.2 Results of Factual Consistency Evaluation Using NEOverlap . . 74
4.5.3 Results of Factual Consistency Evaluation Using AlignScore . . 75
4.5.4 Results of Factual Consistency Evaluation Using QAFactEval . 77
4.5.5 Discussion of Factual Consistency Results . . . 78
4.6 Impact of Source-reference Divergence on Factual Consistency . . . 79
4.7 Human Evaluation . . . 80
4.8 Conclusion . . . 82

5 Enhancing Factual Consistency in LLM-Based D2T 84
5.1 Background . . . 85
5.2 Quantifying Source–Reference Divergence in Data-to-Text Generation . . 86
5.3 Divergence-aware Selective Generation . . . 87
5.4 Experimental Details . . . 89
5.5 Results 1: Existence of Source-Reference Divergence in D2T . . . 89
5.6 Results 2: Factual Consistency vs. Coverage in Divergence-aware Selective Generation . . . 90
5.7 Conclusion . . . 91

6 Exponential Scaling Factual Inconsistency of LLM in D2T 92
6.1 Introduction . . . 92
6.2 Related Work . . . 94
6.3 Desiderata: Scaling Models . . . 95
6.4 VaCScal: Statistical Validation Framework . . . 95
6.5 Experimental Setup . . . 97
6.5.1 D2T Datasets . . . 97
6.5.2 LLM Families . . . 98
6.5.3 Fine-tuning and Decoding Strategies . . . 100
6.5.4 Quantification of Factual Inconsistency . . . 100
6.5.5 Training Settings and Libraries . . . 101
6.5.6 Input Linearisation and Prompt Structure . . . 101
6.6 Results . . . 102
6.6.1 Scaling Behavior Based on Model Size of Fine-Tuned (QLoRA) LLM Families . . . 102
6.6.1.1 Factual Inconsistency Measured via AlignScore . . . 102
6.6.1.2 Factual Inconsistency Measured via QAFactEval . . . 104
6.6.2 Scaling Behaviour Based on FLOPs of Fine-Tuned (QLoRA) LLM Families . . . 105
6.6.2.1 Factual Inconsistency Measured via AlignScore . . . 107
6.6.2.2 Factual Inconsistency Measured via QAFactEval . . . 108
6.7 Discussion . . . 109
6.8 A Conjecture on the Exponential Scaling of Factual Inconsistency in D2T 110
6.9 Conclusion . . . 116

7 Conclusion and Future Directions **117**
7.1 Synthesis and Key Actionable Insights . . . 119
7.2 Future Directions . . . 120
7.3 Concluding Remarks . . . 122

A Appendix: Exponential Scaling Factual Inconsistency of LLM in D2T **123**
A.1 Scaling Behaviour Based on Model Size of Fine-Tuned (QLoRA) LLM Families With Human Evaluation . . . 123
A.2 Result: Scaling Behavior Based on Model Size and FLOPs for Fine-Tuned (QLoRA) BLOOM and OPT Families . . . 126
A.3 Result: Scaling Behavior Based on Model Size of Fine-Tuned (QLoRA) LLM Families . . . 128
A.3.1 Factual Inconsistency Measured via SummaC-conv . . . 128
A.3.2 Factual Inconsistency Measured via UniEval-fact . . . 131
A.4 Results: Scaling Behavior Based on Model Size of Fine-Tuned (QLoRA) LLM Families with Nucleus Sampling . . . 133
A.5 Result: Scaling Behavior Based on Model Size of Fine-Tuned (Prefix-Tuning) LLM Families . . . 137
A.5.1 Factual Inconsistency Measured via AlignScore and QAFactEval. . . . 137
A.6 Result: Scaling Behavior Based on Fine-tune Data Size . . . 141
A.7 Scaling Experiments with Full Fine-tuning . . . 144
A.7.1 Scaling Behavior Based on Model Size of Full Fine-Tuned LLM Families . . . 145
A.7.2 Scaling Behavior Based on FLOPs of Full Fine-Tuned LLM Families 149
A.8 On the Aberrant Scaling Patterns of E2E and ViGGO . . . 152

Bibliography **153**

List of Figures

1.1 An example of D2T from the DART dataset [1], illustrating the transformation of a graph into its corresponding natural language description. . . 1
1.2 Overview of D2T with three major types: graph-to-text (left), table-to-text (middle), MR (meaning representation)-to-text (right). 3
1.3 An example of source-reference divergence, taken from the WikiTableText [2] dataset. The reference text includes two additional facts (bolded and underlined) - 'scrolling performer' and 'fun game' - that are absent in the corresponding source data. 4
1.4 Two of the most widely adopted architectures for language modeling are the encoder–decoder architecture (left) and the decoder-only architecture (right). The sequence $e_1, e_2, \ldots, e_m$ denotes the source input consisting of m tokens, while the generated textual output is represented as $w_1, w_2, \ldots, w_n$, comprising n output tokens. 8
1.5 Performance criteria for D2T, comprising two primary criteria—*readability* and *faithfulness*—along with their respective dimensions: fluency, coherence, and clarity for readability; informativeness and factual consistency for faithfulness. 10
1.6 Examples of generations exhibiting lack of informativeness and factual consistency. 12

2.1 Structure of input data and bottom-up encoding. 20
2.2 Horizontal connections in field encoding (for fields within the j-th record). 21
2.3 Hierarchical connections in field encoding (for fields within the j-th record). 23
2.4 Proposed encoder–decoder model. 25
2.5 t-SNE plots for encoder representations on shuffled input data (circled and squared points represent two data different data instances). 32
2.6 Average human rating of texts generated from the proposed model (**top left**:measures readability in E2E, **top right**: informativeness in E2E, **bottom left**: readability in WikiBio, **bottom right**: informativeness in WikiBio). α denotes the Krippendorff's alpha coefficient. 34

3.1 Three primary types of D2T: graph-to-text (left), table-to-text, and MR-to-text. 37
3.2 A schematic overview of our extensive performance evaluation of LLMs for D2T, covering multiple LLM families, datasets, and metrics, with an additional analysis considering source-reference divergence. 38
3.3 Bidirectional and unidirectional LLM architectures, with red lines indicating attention mechanisms. 41
3.4 Two popular LLM architectures: encoder-decoder and decoder-only. . . . 41

3.5 An example of source-reference divergence, taken from the WikiTableText [2] dataset. The reference text includes two additional facts (bolded and underlined) - 'scrolling performer' and 'fun game' - that are absent in the corresponding source data. 42
3.6 Model sizes of the twelve LLMs from five families: BART (blue), T5 (orange), BLOOM (green), OPT (red), and Llama 2 (purple). 43
3.7 An illustration of unfaithful generation on the E2E dataset. The generated text contains the phrase "**Highly price**", which is irrelevant with respect to the information provided in the source data. 51
3.8 Analysis of performance of T5 in context of source-reference divergence based on three datasets (E2E, WikiTableText and DART) in terms of BLEU, and PARENT, BARTSCORE. First row represents performance of T5-base, and second row represents performance of T5-large. 58
3.9 Analysis of performance of BLOOM in context of source-reference divergence based on three datasets (E2E, WikiTableText and DART) in terms of BLEU, PARENT and BARTSCORE. First row represents performance of BLOOM-3B, and second row represents performance of BLOOM-7B. . . . 59
3.10 Analysis of performance of Llama 2 in context of source-reference divergence based on three datasets (E2E, WikiTableText and DART) in terms of BLEU, PARENT and BARTSCORE. First row represents performance of Llama2-7B, and second row represents performance of Llama2-13B. 59

4.1 Three primary types of D2T: graph-to-text, meaning representation (MR)-to-text, and table-to-text. 65
4.2 Example of factual inconsistency from the DART (graph-to-text) dataset, where the generated claim introduces incorrect information—**'that comes in the moderate price range'**—not found in the source data. . 68
4.3 Illustration of source-reference divergence from the E2E [3] dataset. The reference text introduces additional information—**'is reasonably priced'**—not present in the source data. 69
4.4 All twelve LLMs across five widely used families (BART, T5, OPT, BLOOM, and Llama2) arranged by their model sizes. 70
4.5 AROC of the function $f(\cdot)$ (representing factual consistency) for the model size interval $[m_1, m_2]$ between two LLMs, $\mathcal{M}_1$ and $\mathcal{M}_2$. 72
4.6 AROC values for all LLM families across the five D2T datasets based on SUMMAC-CONV scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency. 73
4.7 AROC values for all LLM families across the five D2T datasets based on NEOVERLAP scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency. 74
4.8 AROC values for all LLM families across the five D2T datasets based on ALIGNSCORE scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency. 75
4.9 AROC values for all LLM families across the five D2T datasets based on QAFACTEVAL scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency. 76

4.10 Impact of source-reference divergence on factual consistency (measured through SummaC-Conv) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines. 77
4.11 Impact of source-reference divergence on factual consistency (measured through NEOverlap) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines. 77
4.12 Impact of source-reference divergence on factual consistency (measured through AlignScore) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines. 79
4.13 Impact of source-reference divergence on factual consistency (measured through QAFactEval) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines. 79

5.1 An illustration of graph-to-text generation, a representative form of Data-to-Text (D2T) generation. 85
5.2 An illustration of source–reference divergence. The boldfaced segments in the reference text indicate content that diverges substantially from the source data. 86
5.3 Distribution of source-reference divergence across the ViGGO, WikiTableText, and WebNLG datasets using our three proposed measures: unigram-level divergence, named entity-based divergence, and field-aware divergence. 88

6.1 An example from the DART [1] dataset illustrating a factually inconsistent output generated by an LLM (Pythia-1.4B). The generated text contains unsupported content—*"his work on Close to Ahm Delta"*—which is not grounded in the source data. 93
6.2 All three stages of *VaCScal*, the statistical validation framework for scaling models. Dashed ovals indicate outputs of each stage. 96

6.3 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by AlignScore, plotted against model size across different LLM families. . . . 103
6.4 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFactEval, plotted against model size across different LLM families. . . . 105
6.5 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by AlignScore, plotted against FLOPs across different LLM families. . . . 106
6.6 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFactEval, plotted against FLOPs across different LLM families. . . . 108
6.7 A simplified systematic view of the logits output in a LLM. . . . 113

A.1 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, based on Human evaluation, plotted against model size for QLoRA fine-tuned FLAN-T5, Mamba, OPT, and Qwen2.5 LLM families. . . . 124
A.2 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFactEval, plotted against model size across different LLM families. . . . 127
A.3 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by AlignScore, plotted against FLOPs across different LLM families. . . . 128
A.4 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by SummaC-conv, plotted against model size across different LLM families. . . . 130
A.5 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by UniEval-fact, plotted against model size across different LLM families. . . . 132
A.6 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by AlignScore, plotted against model size across different LLM families. Here we have considered nucleus sampling for text generation. . . . 134
A.7 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFactEval, plotted against model size across different LLM families. Here we have considered nucleus sampling for text generation. . . . 135
A.8 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by AlignScore, plotted against model size across different LLM families. Here we consider Prefix-Tuning for fine-tuning all LLM families. . . . 138
A.9 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFactEval, plotted against model size across different LLM families. Here we consider Prefix-Tuning for fine-tuning all LLM families. . . . 139

A.10 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by AlignScore, plotted against fine-tune data size across three datasets—DART, E2E and WikiTableText. . . . 142
A.11 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFactEval, plotted against fine-tune data size across three datasets—DART, E2E and WikiTableText. . . . 143
A.12 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by AlignScore, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families. . . . 144
A.13 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by QAFactEval, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families. . . . 146
A.14 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by SummaC-conv, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families. . . . 147
A.15 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by UniEval-fact, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families. . . . 148
A.16 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by AlignScore, plotted against FLOPs across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families. . . . 149
A.17 Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by QAFactEval, plotted against FLOPs across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families. . . . 150

List of Tables

2.1 An example of D2T generation using a model trained on the E2E restaurant-domain dataset. . . . 18
2.2 Overview of the WikiBio and E2E dataset statistics. . . . 28
2.3 Results from WikiBio dataset. . . . 28
2.4 A sample of generated texts from the proposed model on WikiBio dataset (wrong words generation, missing information) . . . 29
2.5 Results on E2E dataset. . . . 29
2.6 A sample of generated texts from the proposed model on E2E dataset (wrong words generation, missing information) . . . 30
2.7 Ablation of copy module (based on BLEU and PARENT score). . . . 30
2.8 Ablation of attention gates (based on BLEU and PARENT score). . . . 30
2.9 A qualitative analysis for the role of attention gates in our model (wrong word, word is not available in input) [E2E dataset contains several gold-standard references for a single input data, but due to space constraint only one reference is given.]. . . . 31
2.10 Ablation of encoder connections (based on BLEU and PARENT score). . . . 31
2.11 Ablation test of skip-interconnections with shuffling records in input data. 32

3.1 Key statistics for the five D2T datasets used in this study. . . . 44
3.2 Readability evaluation of LLMs using BLEU. The best-performing models are highlighted in bold. Higher values indicate better readability. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3). 46
3.3 Readability evaluation of LLMs using METEOR. The best-performing models are highlighted in bold. Higher values indicate better readability. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3). . . . 47
3.4 Informativeness evaluation of LLMs using BERTSCORE. The best-performing models are highlighted in bold. Higher values indicate better informativeness. All results are statistically significant at $p < 0.05$ (see subsection 3.4.3). . . . 48
3.5 Informativeness evaluation of LLMs using MOVERSCORE. The best-performing models are highlighted in bold. Higher values indicate better informativeness. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3). . . . 49
3.6 Informativeness evaluation of LLMs using PARENT. The best-performing models are highlighted in bold. Higher values indicate better informativeness. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3). . . . 50

3.7 Faithfulness evaluation of LLMs using BARTSCORE. The best-performing models are highlighted in bold. Higher values indicate better faithfulness. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3). . . . 51
3.8 Human evaluation of LLMs on D2T tasks. The best-performing models are highlighted in bold. The inter-annotator agreement, measured using Krippendorff's alpha (α), was 0.81, indicating moderate to high consistency among evaluators. . . . 54
3.9 Performance comparison between QLoRA and Prefix-Tuning fine-tuning strategies across all LLMs and datasets, showing consistent results and robustness in evaluation. . . . 56
3.10 Performance comparison between Beam Search and Nucleus Sampling decoding strategies across all LLMs and datasets, showing consistent results and robustness in evaluation. . . . 57
3.11 Case study on the E2E dataset with four LLMs (T5-base, T5-large, Llama2-7B, and Llama2-13B), highlighting errors by smaller models and larger models' failure to detect source-reference divergence. . . . 60
3.12 Case study on the WikiTableText dataset with four LLMs (T5-base, T5-large, Llama2-7B, and Llama2-13B), showing how larger models improve informativeness, while smaller models generate faithful but less readable content. . . . 61
3.13 Performance comparison between conventional and divergence-weighted training across LLMs and D2T datasets. Divergence-weighted training yields consistent improvements across metrics. The best results are shown in **bold**. . . . 62

4.1 Summary of key statistics, D2T types, and domains for the five incorporated datasets. The table shows the average length (in tokens), unique tokens, and total tokens for both sources (linearized to text) and references. 69
4.2 Factual consistency evaluated through SUMMAC-CONV. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3). . . . 73
4.3 Factual consistency evaluated through NEOVERLAP. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3). . . . 74
4.4 Factual consistency evaluated through ALIGNSCORE. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3). . . . 75
4.5 Factual consistency evaluated through QAFACTEVAL. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3). . . . 76
4.6 Human evaluation of factual consistency. A lower value indicates higher factual consistency. . . . 82

5.1 Factual consistency scores for FLAN-T5 under divergence-aware selective generation at varying coverage levels. Blue denotes favorable consistency gains; red indicates aberrant behavior. . . . 89
5.2 Factual consistency scores for OPT under divergence-aware selective generation at varying coverage levels. Blue denotes favorable consistency gains; red indicates aberrant behavior. . . . 90

5.3 Factual consistency scores for Qwen2.5 under divergence-aware selective generation at varying coverage levels. Blue denotes favorable consistency gains; red indicates aberrant behavior. 90

6.1 Summary of key statistics, D2T types, and domains for the five datasets used in this study. The table reports the average text length (in tokens), number of unique tokens, and total tokens for both source data (linearized to text) and reference outputs. 98
6.2 Overview of the six LLM families, their architectural paradigms, number of models, and corresponding parameter sizes (M = million, B = billion). 99
6.3 Results from *VaCScal* corresponding to Figure 6.3. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 104
6.4 Results from *VaCScal* corresponding to Figure 6.4. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 106
6.5 Results from *VaCScal* corresponding to Figure 6.5. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 107
6.6 Results from *VaCScal* corresponding to Figure 6.6. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 109

A.1 The five short sample given to each annotators for training before annotation task. 124
A.2 Results from *VaCScal* corresponding to Figure A.1. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 125
A.3 Results from *VaCScal* corresponding to Figure A.2. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 126
A.4 Results from *VaCScal* corresponding to Figure A.3. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 126
A.5 Results from *VaCScal* corresponding to Figure A.4. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 129
A.6 Results from *VaCScal* corresponding to Figure A.5. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 131
A.7 Results from *VaCScal* corresponding to Figure A.6. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 133
A.8 Results from *VaCScal* corresponding to Figure A.7. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 136
A.9 Results from *VaCScal* corresponding to Figure A.8. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 137

A.10 Results from *VaCScal* corresponding to Figure A.9. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 140
A.11 Results from *VaCScal* corresponding to Figure A.10. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 142
A.12 Results from *VaCScal* corresponding to Figure A.11. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 142
A.13 Results from *VaCScal* corresponding to Figure A.12, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 145
A.14 Results from *VaCScal* corresponding to Figure A.13. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 146
A.15 Results from *VaCScal* corresponding to Figure A.14. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 147
A.16 Results from *VaCScal* corresponding to Figure A.15. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 148
A.17 Results from *VaCScal* corresponding to Figure A.16. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 150
A.18 Results from *VaCScal* corresponding to Figure A.17. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*. 151
A.19 Type-Token Ratio (TTR) for all five D2T datasets, calculated separately for input source data and corresponding reference texts. 152

Abbreviations

D2T	**D**ata-to-**T**ext Generation
LLM	**L**arge **L**anguage **M**odel
LSTM	**L**ong **S**hort-**T**erm **M**emory

This thesis is dedicated to
the cherished memory of Maa and Baba...

Chapter 1

Introduction

Data-to-Text Generation (D2T) [4–6] is a branch of text generation that focuses on automatically transforming structured or semi-structured data—such as tables [2], graphs [1, 7], or meaning representations [8]—into human-readable natural language descriptions. Its applications span a wide range of domains, including weather forecasting [9], automated journalism [10, 11], and dialogue systems [12], among others.

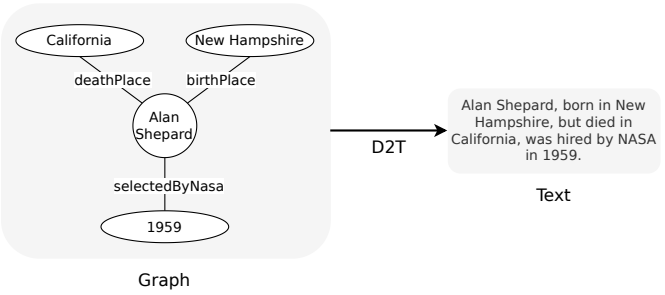


FIGURE 1.1: An example of D2T from the DART dataset [1], illustrating the transformation of a graph into its corresponding natural language description.

Language models are fundamental to all forms of text generation, including D2T, owing to their exceptional ability to produce coherent and contextually appropriate text [13–15]. Over time, D2T has evolved in parallel with advancements in language modeling [16–18]. Early progress stemmed from statistical [19–22] and recurrent neural network (RNN)-based language models [23, 24], which offered substantial improvements in fluency and coherence while enabling faster development compared to traditional rule-based D2T systems [25, 26] that were labor-intensive, time-consuming, and dependent on extensive domain-specific engineering. A major shift occurred with the advent of *large* language models (LLMs) [27–29], driven primarily by transformer architectures [30], which have redefined the state of the art in D2T. LLM-based approaches demonstrate remarkable fluency, coherence, and adaptability, often producing human-like narratives directly from structured or semi-structured data sources [31–33].

Despite their strong performance in readability—particularly in terms of fluency, clarity, and coherence—language model-based D2T systems continue to struggle with a fundamental challenge: preserving faithfulness to the source (input) data during generation [34, 35]. Several factors, including repetitive generation [36, 37], weak grounding in training data [38, 39], decoding errors [40, 41], and inherent model biases [42, 43], can lead to unfaithful outputs where generated statements are unsupported by or contradict the input data [44, 45]. Faithfulness in D2T is commonly characterized along two dimensions: informativeness, which assesses whether the generated text captures salient information from the input [46, 47], and factual consistency, which evaluates whether generated content remains correct with respect to the source data [13, 48]. Failures in either dimension may reduce the reliability and trustworthiness of D2T systems and could be particularly concerning in applications where factual accuracy is important, including domains such as healthcare and medical report generation [49–51].

Although considerable progress has been made in improving the fluency and generation capabilities of language model-based D2T systems, the reliable evaluation and enhancement of faithfulness in D2T remain open challenges [35, 52, 53]. Existing approaches for D2T often rely on imperfect automatic evaluation metrics, make assumptions regarding source-reference alignment, or overlook factors such as the structural properties of source data and source-reference divergence, i.e., discrepancies between source inputs and their corresponding human-written references [39, 54]. Consequently, understanding how these factors affect faithfulness in D2T systems remains an active research direction and may contribute to the development of more reliable data-to-text generation systems.

Motivated by these challenges, this thesis investigates faithfulness in language model-based D2T systems, with particular emphasis on large language models (LLMs), addressing both the evaluation and enhancement of faithful generation. Faithfulness is viewed as the integration of two complementary dimensions: informativeness and factual consistency. These dimensions are systematically studied using representative contemporary language models, particularly LLMs such as FLAN-T5 [55], Mamba [56], and Qwen2.5 [33], across widely adopted D2T benchmarks including E2E [8], DART [1], and WikiTableText [2], with particular attention to factors such as source structure and source-reference divergence.

In the following sections of this chapter, we begin by presenting a brief description of D2T (section 1.1). Since language models play a pivotal role in D2T, we then discuss, in section 1.2, the underlying principles of language modeling, including its core objectives, training and decoding strategies, and the prevalent paradigms of pre-training and fine-tuning. Building upon this foundation, we illustrate how language modeling integrates with D2T, leading to the emergence of probabilistic D2T approaches (section 1.3).

Subsequently, in section 1.4, we outline the key performance criteria for evaluating D2T systems, focusing on two essential qualities: *readability*—encompassing fluency, coherence, and clarity—and *faithfulness*—comprising informativeness and factual consistency. This is followed by a brief overview of the evolution of D2T with language models and their performance advancements over time (section 1.5). Next, in section 1.6, we address the challenge of ensuring faithfulness in language model-based D2T, emphasizing its significance as a central research problem. Finally, the chapter concludes by outlining the objectives of this thesis (section 1.7) and providing an overview of its overall structure (section 1.8), thereby offering a clear roadmap for the subsequent chapters.

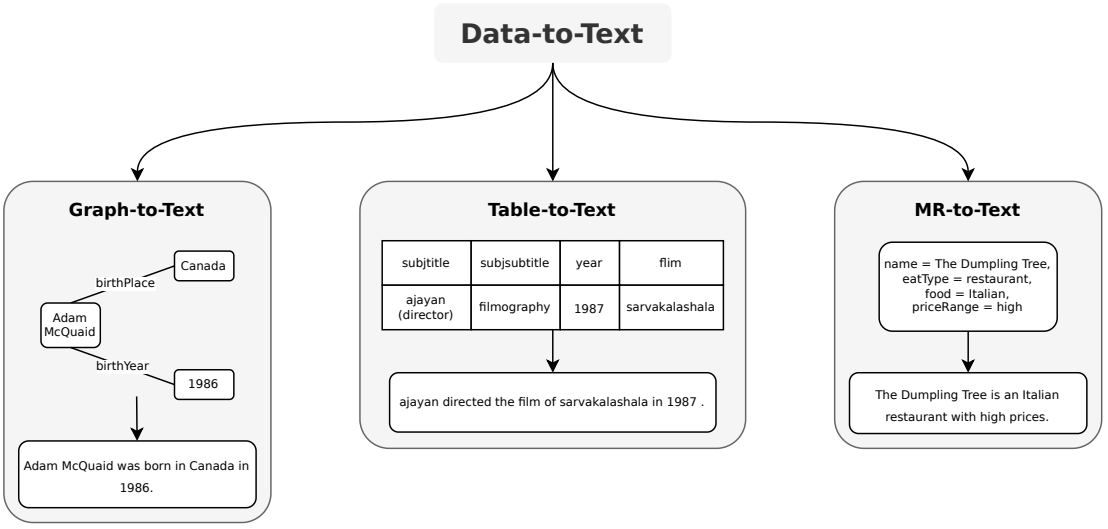


FIGURE 1.2: Overview of D2T with three major types: graph-to-text (left), table-to-text (middle), MR (meaning representation)-to-text (right).

1.1 Data-to-Text Generation (D2T)

D2T refers to the task of automatically transforming structured or semi-structured data—such as tables, graphs, or meaning representations—into coherent and contextually appropriate natural language text [57–60]. Depending on the structure of the input representation, D2T tasks are generally classified into three broad categories [35]: *graph-to-text*, *table-to-text*, and *MR-to-text* (Figure 1.2). In graph-to-text generation [1, 7], structured graph data or sets of knowledge triples are converted into fluent textual narratives, while table-to-text generation [2, 47, 61] produces natural language descriptions from tabular data. MR-to-text generation [8, 62–64] focuses on verbalizing slot–value pairs, commonly referred to as meaning representations (MRs). D2T has been applied across a variety of domains, including weather forecasting [9], automated journalism [5], and dialogue systems [65], and more recently in safety-critical contexts such as medical report generation [66], where factual accuracy and reliability are of utmost importance.

It is typically viewed as a form of *conditional text generation*, wherein the output text is conditioned on structured input data—similar in spirit to related tasks such as image captioning, machine translation, and summarization [14, 26, 52]. However, unlike these tasks, where the input is predominantly unstructured or sequential text, D2T operates on structured or semi-structured representations [35, 38]. This distinction introduces unique challenges, particularly due to the phenomenon of *source–reference divergence*, which arises when the reference text in a dataset exhibits factual or semantic discrepancies relative to its corresponding source input [39, 54, 67–69]. An example of such divergence is illustrated in Figure 1.3. Addressing these challenges requires specialized alignment strategies, training objectives, and representation techniques to ensure that models accurately interpret and faithfully verbalize the input data.

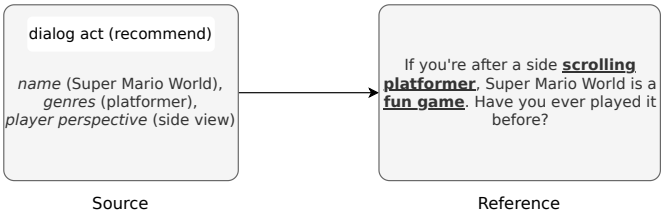


FIGURE 1.3: An example of source-reference divergence, taken from the WikiTableText [2] dataset. The reference text includes two additional facts (bolded and underlined) - 'scrolling performer' and 'fun game' - that are absent in the corresponding source data.

Although D2T has been applied across a wide range of domains, the experimental studies presented in this thesis primarily rely on widely adopted benchmark datasets to enable comparison with existing baselines and maintain consistency with established experimental settings in prior literature. Across different studies included in this thesis, experiments are conducted using the following prevalent D2T datasets: E2E, ViGGO, DART, WebNLG, WikiTableText, and WikiBio. These datasets collectively cover diverse input representations, domains, and generation characteristics. Specifically, E2E [8] is an MR-to-text dataset in the restaurant domain where structured slot–value pairs are transformed into natural language descriptions. ViGGO [62] is another MR-to-text dataset focusing on the video game domain and contains richer semantic structures with more diverse dialogue acts. DART [1] and WebNLG [7] belong to the graph-to-text category, where structured knowledge triples from multiple domains are verbalized into coherent textual descriptions. In contrast, WikiTableText [2] and WikiBio [4] are table-to-text datasets that generate descriptive text from tabular information extracted from Wikipedia. Together, these datasets provide broad coverage across graph-to-text, table-to-text, and MR-to-text D2T tasks. The findings presented in this work should therefore be interpreted within the scope of these benchmark datasets, and no direct extrapolation is made beyond them. Nevertheless, as is common in empirical research, the observations

and insights obtained from these experiments may motivate follow-up investigations and provide directions for extending the proposed ideas to broader domains and settings.

Most contemporary D2T systems are built upon a probabilistic framework, largely driven by the advancements in language models, which form their fundamental core owing to their remarkable generative capabilities [14, 70, 71]. In the following section, we discuss the principles of language modeling in greater detail.

1.2 Language Modeling

Language modeling serves as the foundational computational framework for probabilistic text generation tasks, including D2T [33, 72]. A typical language model operates through three core components: the *objective*, *training*, and *decoding* (or inference) stages. Recent advancements in language modeling have been propelled by the paradigms of *pre-training* and *fine-tuning*, which have significantly improved the generalization and adaptability of models across a wide range of generation tasks [73].

Objective. The objective of language modeling is to learn a probabilistic model $\mathcal{Q}_\phi$, parameterized by ϕ, that assigns a likelihood to a sequence of tokens [20, 28, 74], thereby capturing the underlying structure and semantics of language:

$$\mathcal{Q}_\phi : \Pr(w_1, w_2, \ldots, w_T; \phi). \tag{1.1}$$

Direct modeling of this joint probability (Equation 1.1) is intractable for long sequences; hence, it is decomposed via the chain rule of probability:

$$\Pr(w_1, w_2, \ldots, w_T; \phi) = \prod_{t=1}^{T} \Pr(w_t \mid w_1, w_2, \ldots, w_{t-1}; \phi) \tag{1.2}$$

$$= \Pr(w_1) \Pr(w_2 \mid w_1; \phi) \ldots \Pr(w_T \mid w_1, \ldots, w_{T-1}; \phi). \tag{1.3}$$

As shown in Equation 1.3, the autoregressive formulation serves as the foundation for most language models [75, 76].

Training. Given a dataset $\mathcal{D}$ comprising a collection of texts (i.e., token sequences),

$$\mathcal{D} = \left\{x^{(i)}\right\}_{i=1}^{|\mathcal{D}|}, \quad \text{where} \quad x^{(i)} = w_1^{(i)} w_2^{(i)} \dots w_{T_i}^{(i)}, \tag{1.4}$$

the training objective in language modeling is to estimate the parameters ϕ by maximizing the likelihood of the observed sequences, typically using maximum likelihood estimation (MLE) [77]:

$$\phi^* = \underset{\phi}{\operatorname{argmax}} \prod_{i=1}^{|\mathcal{D}|} \Pr(x^{(i)}; \phi) = \underset{\phi}{\operatorname{argmin}}\ \mathcal{L}(\phi). \tag{1.5}$$

Where the loss $\mathcal{L}(\phi)$ is commonly defined as the negative log-likelihood (NLL), or equivalently, the cross-entropy loss [78]:

$$\mathcal{L}(\phi) = -\sum_{i=1}^{|\mathcal{D}|} \log \Pr(x^{(i)}; \phi) = -\sum_{i=1}^{|\mathcal{D}|} \sum_{t=1}^{T_i} \log \Pr\left(w_t^{(i)} \mid w_1^{(i)}, \dots, w_{t-1}^{(i)}; \phi\right). \tag{1.6}$$

Decoding. Decoding [79], or inference, enables a trained language model $\mathcal{Q}_{\phi^*}$ to determine the next token ($\widehat{w_t}$) given the preceding context:

$$\widehat{w_t} = \underset{\text{at } t\text{-th step}}{\text{decoding}}\ \mathcal{Q}_{\phi^*}(w_1, w_2, \dots, w_{t-1}) \tag{1.7}$$

$$= \underset{\text{at } t\text{-th step}}{\text{decoding}}\ \Pr\left(\cdot \mid w_1, w_2, \dots, w_{t-1}; \phi^*\right). \tag{1.8}$$

Common decoding strategies can broadly be categorized into deterministic and stochastic approaches. Deterministic decoding methods primarily include greedy search and beam search. Greedy search generates text by selecting the token with the highest conditional probability at each decoding step, making it computationally efficient but often leading to repetitive outputs or suboptimal local decisions. Beam search extends this approach by maintaining multiple candidate sequences (beams) during generation and selecting the sequence with the highest cumulative probability, thereby exploring a larger search space and often improving generation quality. In contrast, stochastic decoding methods introduce randomness into the generation process to enhance output diversity. Common stochastic approaches include top-k sampling, pure sampling, and nucleus (top-p) sampling. In top-k sampling, the model restricts the candidate token

space to the top k most probable tokens and samples the next token from this reduced distribution, preventing selection from low-probability tokens while still preserving diversity [40]. Pure sampling selects the next token directly from the entire conditional probability distribution, allowing highly diverse outputs but potentially generating low-quality or incoherent text. To mitigate this issue, nucleus (top-p) sampling dynamically selects the smallest set of candidate tokens whose cumulative probability exceeds a threshold p, balancing generation diversity and quality by filtering out unlikely token candidates [79].

Pre-training Fine-tuning Paradigm. Language models achieve broad adaptability through a two-stage paradigm of *pre-training* and *fine-tuning* [27, 73, 78]. In this framework, pre-training focuses on learning general linguistic and semantic knowledge from large unlabeled corpora via self-supervised objectives such as next-token prediction, while fine-tuning adapts the model to specific downstream tasks using labeled data [80]. During pre-training, the model $\mathcal{Q}_\phi$ is trained on an unlabeled dataset $\mathcal{D}_{\text{pt}}$ to minimize the negative log-likelihood loss:

$$\mathcal{L}_{\text{pt}}(\phi) = -\sum_{x \in \mathcal{D}_{\text{pt}}} \sum_{t=1}^{T} \log \Pr(w_t \mid w_1, \ldots, w_{t-1}; \phi) \tag{1.9}$$

Subsequently, fine-tuning [80] adapts these parameters to a specific task using a labeled dataset $\mathcal{D}_{\text{ft}}$, optimizing the conditional likelihood of target output y given input x:

$$\mathcal{L}_{\text{ft}}(\phi) = -\sum_{(x,y) \in \mathcal{D}_{\text{ft}}} \log \Pr(y \mid x; \phi). \tag{1.10}$$

This two-stage framework allows efficient transfer of general language knowledge to task-specific applications, significantly enhancing the model's adaptability and performance across diverse language generation tasks [73, 80].

Language Model Architectures. Most Language models are implemented using one of two major architectural paradigms: encoder–decoder architecture [77, 81] and decoder-only architecture [28, 76], as illustrated in Figure 1.4. The encoder–decoder architecture consists of two components that operate jointly. The encoder first processes the input sequence $(e_1, e_2, \ldots, e_m)$ and transforms it into a latent contextual representation that captures semantic and syntactic information from the input. The decoder then

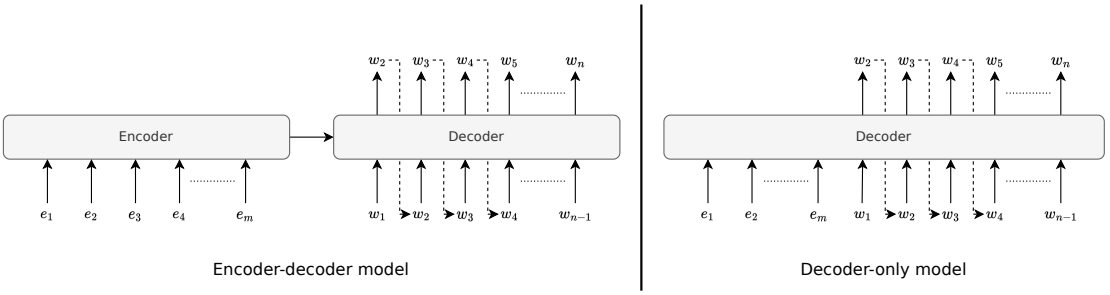


FIGURE 1.4: Two of the most widely adopted architectures for language modeling are the encoder–decoder architecture (left) and the decoder-only architecture (right). The sequence $e_1, e_2, \ldots, e_m$ denotes the source input consisting of m tokens, while the generated textual output is represented as $w_1, w_2, \ldots, w_n$, comprising n output tokens.

utilizes this encoded representation to generate the target sequence $(w_1, w_2, \ldots, w_n)$ token by token, often through an attention mechanism that allows the model to selectively focus on relevant parts of the input during generation. This architecture is particularly suitable for sequence transformation tasks such as machine translation, summarization, and D2T, where the output is conditioned on a separate input sequence. In contrast, decoder-only architectures generate text in an autoregressive manner by predicting each subsequent token $(w_1, w_2, \ldots, w_n)$ based on previously observed context $(e_1, e_2, \ldots, e_m)$. Rather than explicitly separating encoding and generation stages, these models rely on self-attention mechanisms to learn contextual dependencies directly within the input sequence. Owing to their simplicity and scalability, decoder-only models have become the foundation of many contemporary large language models and are widely used for text completion and instruction-following tasks.

From Language Models to Large Language Models (LLMs). LLMs extend the paradigm of pre-trained language models by scaling both model parameters and training corpora, often reaching billions of parameters [70, 82, 83]. They capture rich linguistic and semantic dependencies across diverse contexts and support large context windows for extended text generation [29, 32]. To efficiently adapt LLMs to specific tasks, parameter-efficient fine-tuning methods such as LoRA [84, 85] and prefix-tuning [86] update only a subset of parameters, while prompting-based techniques enable zero-shot and few-shot adaptation through instruction or example conditioning [87]. These innovations collectively make LLMs highly flexible and effective for natural language generation, leading to their widespread adoption across a broad range of generative AI applications.

1.3 From Probabilistic D2T to Language Model-based D2T

As discussed earlier, recent approaches to D2T are predominantly probabilistic in nature and leverage language models as their core generative framework [35]. Here, we establish a brief formal connection between the probabilistic formulation and language model-based implementations of D2T. A probabilistic formulation of D2T can be expressed as a conditional mapping function:

$$f_{\text{D2T}} : X \times Y \rightarrow [0, 1], \tag{1.11}$$

where $f_{\text{D2T}}(x, y)$ represents a probability distribution over the joint space of inputs X and outputs Y. Here, X denotes the space of structured or semi-structured data (such as tables, graphs, or meaning representations), and Y denotes the space of corresponding natural language texts. For a given input $x \in X$, the model assigns a probability to each possible textual realization $y \in Y$, defined as

$$f_{\text{D2T}}(x, y) = \Pr(y \mid x) \tag{1.12}$$

This conditional probability models the inherently stochastic nature of language generation, assigning higher probabilities to more fluent and faithful descriptions of the input data. In practice, f_{D2T} is parameterized by a probabilistic model $\mathcal{M}_\theta$, where θ denotes the set of trainable parameters. The goal of $\mathcal{M}_\theta$ is to approximate the true conditional distribution $\Pr(y \mid x)$ (Equation 1.12) by maximizing the likelihood of generating text y that is both semantically relevant and faithful to the input x.

Due to their exceptional generative ability, language models have become the dominant choice for implementing $\mathcal{M}_\theta$ in D2T [6, 18]. Typically, $\mathcal{M}_\theta$ is instantiated as a language model $\mathcal{Q}_\phi$, parameterized by ϕ, where $\phi \subseteq \theta$. This design reflects a common paradigm in which a pre-trained language model serves as the core generative component within a broader D2T architecture, possibly enhanced with additional modules—such as copy mechanisms or parameter-efficient fine-tuning adapters (e.g., LoRA)—for improved control and task adaptation. Under this formulation, the conditional probability of generating a text y given an input x can be expressed as:

$$\mathcal{M}_\theta(x) \approx \mathcal{Q}_\phi(x) \tag{1.13}$$

where this approximation assumes that the auxiliary/additional modules primarily guide or adapt $\mathcal{Q}_\phi$ without altering its fundamental generative objective (Equation 1.12).

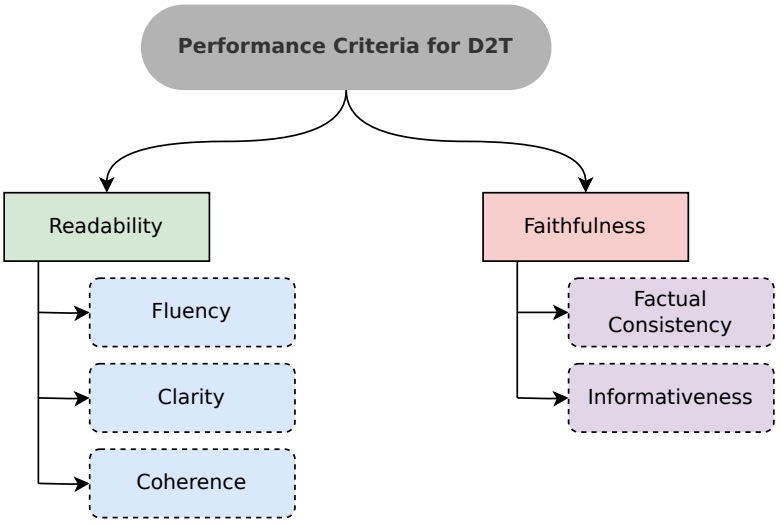


FIGURE 1.5: Performance criteria for D2T, comprising two primary criteria—*readability* and *faithfulness*—along with their respective dimensions: fluency, coherence, and clarity for readability; informativeness and factual consistency for faithfulness.

1.4 Performance Measurement of D2T Systems: Readability and Faithfulness

The performance of a D2T system is generally measured along two fundamental dimensions of quality: *readability* and *faithfulness* [35, 52]. Readability pertains to how well the generated text aligns with human linguistic standards, encompassing three key dimensions—*fluency* (the grammatical correctness and natural flow of the text) [88, 89], *coherence* (the logical organization and smooth progression of ideas) [90], and *clarity* (the precision and unambiguity with which ideas are expressed, minimizing confusion or misinterpretation) [91]. A readable output should be both syntactically well-formed and stylistically natural, ensuring that the generated text is easily comprehensible to human readers.

However, readability alone does not guarantee the reliability of a D2T system. Another crucial criterion, *faithfulness*, assesses the extent to which the generated text accurately

reflects the information contained in the input data [13, 92]. Faithfulness is commonly characterized by two dimensions: *informativeness* [38, 39] and *factual consistency* [34, 93]. These two dimensions of faithfulness in D2T capture the following objectives:

Informativeness measures whether the generated text from a D2T includes all salient and relevant information from the input source data.

Factual Consistency assesses the extent to which the generated text from a D2T remains consistent with the input source data, without introducing unsupported, contradictory, or distorted information.

Informativeness is often task-dependent and subjective, as its evaluation typically requires consideration of the user's intent or reference texts to determine whether the generated text conveys relevant information from the source data. In contrast, factual consistency is more objective, as it concerns the factual correctness of the generated text with respect to the given source data. In many D2T settings, evaluating the factual consistency of generated text with respect to the reference text alone is often sufficient, as the reference text generally conveys the same information as the input source data [53]. However, situations may arise in which the source and reference exhibit semantic or factual discrepancies, a phenomenon referred to as *source-reference divergence* [39, 48, 94]. Under these conditions, evaluating consistency with respect to both the source and the reference becomes important for obtaining a more reliable assessment. Moreover, as source data in D2T are typically structured or semi-structured, factual consistency can, in principle, be assessed directly against the source data, provided that the source information can be appropriately verbalized/linearized [65], which is often challenging in practice.

Factual consistency shares similarities with entailment in Natural Language Inference (NLI) [95, 96], where the objective is to determine whether a claim/hypothesis is supported by a given premise. In D2T, the generated text can be viewed as the claim/hypothesis and the source data as the premise. However, factual consistency in D2T differs from traditional NLI in multiple ways. Unlike NLI, where both the premise and hypothesis are textual, D2T source data are often structured or semi-structured (e.g., tables, graphs, or meaning representations), making fact extraction and representation crucial. Furthermore, using reference texts as premises may be unreliable in the presence of *source-reference divergence*, where references differ from the source data. Despite these differences, many studies have employed entailment-based measures to approximate factual consistency [35], either by assessing generated text with respect to reference texts or by verbalizing/linearizing source data into text.

It is worth noting that this thesis focuses on factual consistency in D2T with respect to the input source data. In contrast, *factuality* [97, 98], which concerns the truthfulness of generated text with respect to both the source data and broader real-world knowledge (which may include additional information that supplements or supports the source data), is beyond the scope of this thesis.

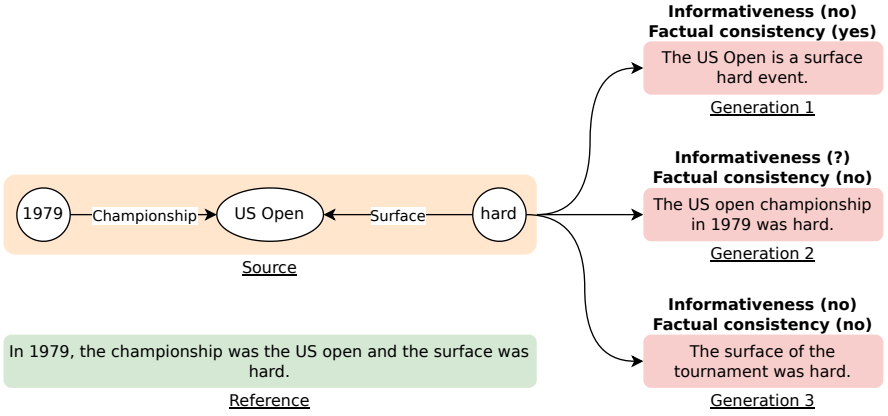


FIGURE 1.6: Examples of generations exhibiting lack of informativeness and factual consistency.

Figure 1.6 illustrates three examples of model generations that suffer from non-informativeness, factual inconsistency, or both. The examples are drawn from a graph-to-text dataset, specifically from DART [1]. The source data (a graph representing information about the 'US Open') and the corresponding gold reference text are shown alongside three erroneous generations (Generation 1, Generation 2, and Generation 3), produced by models from the BART [31], FLAN-T5 [55], and OPT [99] families, respectively. Generation 1 (BART) exhibits a lack of informativeness, as it omits a crucial detail about the year *1979*. However, it does not suffer from factual inconsistency, since the generated text contains no unsupported or contradictory facts relative to the source data. In contrast, Generation 2 (FLAN-T5) demonstrates factual inconsistency, misinterpreting the input by stating that the 'championship is hard,' whereas the source data specifies that the *surface* is hard. Finally, Generation 3 (OPT) suffers from both issues—it omits key information and introduces factual errors, resulting in a generation that is neither informative nor factually consistent with the source data.

An effective D2T system must therefore strike a balance between readability and faithfulness—producing text that is not only linguistically fluent but also factually accurate and true to the source data, thereby ensuring that the generated output is both natural and trustworthy [35, 100].

1.5 Evolution of D2T with Language Models and Their Performance

The evolution of D2T has progressed in tandem with advancements in language modeling. Early D2T systems were primarily rule-based [25], relying on manually designed, domain-specific pipelines encompassing content selection, content planning, referring expression generation, and surface realization. Although these systems offered interpretability and precise control, they required extensive manual effort and lacked scalability and generalization [22, 52, 101].

The advent of statistical language models marked an important transition toward automation in D2T [10, 22, 101, 102]. Statistical n-gram models [19, 20, 103] introduced a probabilistic framework that unified various stages of the D2T pipeline under a single data-driven paradigm. These models improved scalability and robustness but often produced text that was disfluent, repetitive, or lacking in coherence [104–106].

A significant development occurred with the emergence of neural language models, particularly those based on recurrent neural networks (RNNs) [107] and Long Short-Term Memory (LSTM) architectures [23]. These models enabled D2T systems to capture long-range dependencies and generate more fluent text [4, 108, 109]. The introduction of encoder–decoder architectures [77] further expanded these capabilities by supporting end-to-end training from structured inputs to natural language outputs [12, 110]. Despite these gains in readability, maintaining faithfulness and factual consistency remained a persistent challenge [5].

The integration of attention mechanisms [63, 111–113] and the adoption of pre-training and fine-tuning paradigms [73] gave rise to more powerful neural D2T systems [47, 61, 63, 108, 109, 114, 115], improving fluency, coherence, and contextual alignment. The introduction of the Transformer architecture [30] represented another important advancement, offering parallelized sequence modeling and improved handling of long-range dependencies, which further enhanced both readability and overall performance.

A recent phase in the evolution of D2T is driven by LLMs [27, 28, 31, 76]. Built upon transformer architectures and pre-trained on vast corpora using self-supervised objectives [27, 74, 77, 78], LLMs have achieved strong performance in D2T [14, 34, 35], achieving high level of fluency, coherence, and stylistic naturalness. Through fine-tuning and parameter-efficient adaptation [42, 116], these models often exhibit strong generalization capabilities even with limited task-specific data.

Nevertheless, despite their impressive readability and linguistic sophistication, LLM-based D2T systems continue to struggle with factual faithfulness—particularly in maintaining both the accuracy and completeness of the information conveyed [93, 100]. This limitation underscores a key ongoing challenge in language model-based D2T systems.

1.6 Ensuring Faithfulness: A Key Challenge in Language Model-based D2T

Although language model-based D2T systems—including LLMs—have achieved notable improvements over earlier approaches, particularly in terms of readability, they continue to face persistent challenges in maintaining *faithfulness* to the source data, especially with respect to *informativeness* and *factual consistency* [34, 35, 117]. Faithfulness in language model-based D2T is impeded by several factors, including exposure bias arising from teacher-forced training, which causes error accumulation during free-form generation [92]; source–reference divergence in D2T datasets, where reference texts deviate factually from the structured input, weakening data grounding [38]; stochastic decoding strategies, which introduce variability and factual inaccuracies [118]; and pre-training biases stemming from noisy or unbalanced data, which lead models to overgeneralize or rely on prior world knowledge rather than the provided input [119, 120]. A lack of faithfulness often manifests as *hallucination* [54, 121], where the model generates content that is unsupported by—or even contradictory to—the source data, thereby undermining the reliability and factual integrity of the generated text. Such behavior may pose risks in high-stakes domains such as healthcare, law, and autonomous decision-making, where factual precision and accountability are paramount [49–51]. This overarching challenge can be examined along two major directions. The first concerns the *evaluation of faithfulness* in language model-based D2T systems. Although research on faithfulness evaluation has grown substantially in recent years, several challenges remain, particularly for LLM-based D2T systems. The second direction involves the *enhancement of faithfulness*, which aims to mitigate the aforementioned sources of unfaithfulness—such as weak data grounding and model biases—that contribute to inaccurate or misleading text generation. Therefore, evaluating and enhancing faithfulness in language model-based D2T systems is a critical step toward ensuring their trustworthiness.

1.7 Objectives and Contributions of the Thesis

The challenges outlined in section 1.6 underscore the critical need to improve faithfulness in language model-based D2T. In response, this thesis pursues two complementary

objectives: (i) to systematically evaluate faithfulness and (ii) to develop methods for enhancing faithfulness in language model-based D2T systems.

In pursuit of these objectives, the thesis makes several key contributions, focusing on the two fundamental dimensions of faithfulness: *factual consistency* and *informativeness*. First, it presents a pre-LLM structure-aware encoding framework designed to enhance the informativeness aspect of faithfulness by leveraging hierarchical representations of input data and incorporating selective attention gating to refine content selection and representation. Second, it presents an empirical evaluation of state-of-the-art language models across multiple D2T benchmarks, identifying source-reference divergence as an important factor associated with model faithfulness and proposing a divergence-weighted training strategy with an adaptive loss function to mitigate its effects. Third, it provides a broad analysis of factual consistency in LLMs using both automatic and human evaluations, culminating in the development of a divergence-aware selective generation framework that improves factual consistency under the evaluated benchmark settings while retaining much of the informativeness of generated outputs. Finally, the thesis presents empirical evidence, supported by a mathematical rationale, suggesting that factual inconsistency in LLM-based D2T systems follows an *exponential* rather than a *power-law* scaling trend under the evaluated D2T settings with respect to model size, compute, and data volume, a finding further examined through a statistical framework termed *VaCScal*. Collectively, these objectives and contributions define a unified research agenda that bridges the diagnostic and corrective aspects of faithfulness in language model-based D2T. Through detailed evaluation and enhancement strategies, the thesis contributes toward the development of more reliable and trustworthy D2T systems that may be useful in applications where factual accuracy is important.

1.8 Thesis Overview

The subsequent chapters of this thesis are organized to progressively explore the evaluation and enhancement of *faithfulness* in language model-based D2T, focusing on its two core dimensions—*informativeness* and *factual consistency*.

Chapter 2 describes a structure-aware encoding framework designed to enhance informativeness in pre-LLM D2T architectures. It presents an encoder–decoder framework based on LSTM networks that encodes input data according to its inherent record–field hierarchy. The proposed model improves informativeness through selective content gating and sequential–hierarchical encoding, thereby facilitating more effective content selection and structured representation.

Chapter 3 presents an extensive performance evaluation of state-of-the-art LLMs for D2T. It examines models from five major LLM families—BART, T5, BLOOM, OPT, and Pythia—across five widely used datasets: E2E, DART, ViGGO, WikiTableText, and WebNLG, encompassing the three principal D2T paradigms of table-to-text, graph-to-text, and MR-to-text generation. The chapter evaluates LLM performance in terms of readability, informativeness, and faithfulness, and identifies *source–reference divergence*—the semantic mismatch between input data and reference text—as a primary factor undermining these dimensions. To mitigate this issue, it introduces a *divergence-weighted training* strategy with an adaptive loss function, which improves performance of LLM-based D2T systems across the evaluated datasets and metrics.

Chapter 4 reports the findings of a broad empirical study on *factual consistency*, a key component of faithfulness in D2T. It systematically evaluates factual consistency across multiple LLM families using both automatic and human assessments. To analyze the relationship between model scaling and factual consistency, the chapter employs the *Average Rate of Change* (AROC) as a robust quantitative indicator, offering insights into how factual consistency evolves with model scale.

Chapter 5 discusses approaches for mitigating factual inconsistency in LLM-based D2T. It introduces a *divergence-aware selective generation framework* that enables models to selectively abstain from generating text when input–reference divergence exceeds a defined threshold. This mechanism enhances factual consistency while maintaining satisfactory levels of coverage and informativeness.

Chapter 6 presents an empirical and mathematical investigation into the scaling behavior of factual inconsistency in LLM-based D2T. The analysis suggests that factual inconsistency follows an *exponential*, rather than a power-law, relationship with respect to model size, computational capacity, and data volume. To validate this observation, the chapter introduces a statistical validation framework, *VaCScal*, which provides statistical evidence supporting the exponential scaling trend and provides a mathematical rationale for its emergence.

Finally, Chapter 7 concludes the thesis by summarizing the key findings from all preceding chapters, discussing the major contributions of this research, and outlining potential directions for future work aimed at improving the faithfulness, factual grounding, and reliability of language model-based D2T systems.

Chapter 2

Enhancing Informativeness in D2T through Structural Encoding

This chapter[1] presents a pre-LLM, structure-aware LSTM-based encoder–decoder framework for data-to-text generation (D2T) that aims to enhance *faithfulness* in terms of *informativeness*. The proposed model encodes the input data according to its underlying structure, with the goal of producing representations that are more suitable for D2T. It further incorporates *attention gates* in the decoder to selectively attend to relevant content, which is associated with improved quality and informativeness on the evaluated benchmarks. This chapter presents the design of the proposed model and reports the results of experiments conducted on the WikiBio and E2E datasets. The experimental findings suggest that the proposed approach performs competitively and achieves improvements over the evaluated baseline and comparison systems under the considered experimental settings.

2.1 Background

Owing to their end-to-end training and the fluency of the generated text [8, 77], encoder–decoder based architectures have been commonly used in D2T systems [47, 61]. However, in such frameworks, obtaining an informative encoding of the input data remains a key challenge. Since the input data in many D2T tasks naturally exhibit a record–field structure (as illustrated in Table 2.1), incorporating *structural* encoding may provide a useful mechanism for representing this organization. Despite this, many

[1]The work presented in this chapter is based on our paper: Mahapatra, J. and Garain, U., 2021. Exploring Structural Encoding for D2T. *International Conference on Natural Language Generation (INLG)*, Association for Computational Linguistics. [115]

existing encoder–decoder models for D2T generation [61, 109] tend to emphasize attention mechanisms over the explicit structural encoding of record–field data.

TABLE 2.1: An example of D2T generation using a model trained on the E2E restaurant-domain dataset.

Data (x)	
Record 1	name: *The Punter*
Record 2	food: *English*
Record 3	price range: *high*
Text (y)	
The Punter is a restaurant with high prices.	

Selecting appropriate content from record–field structured input data—commonly referred to as *content selection*—is an important step in D2T. For instance, as shown in Table 2.1, the `name` and `price range` fields of the data are considered relevant contents with respect to the reference text y. Identifying such relevant information from encoded representations, however, remains a challenging task.

To address this, we propose an encoder–decoder model for D2T in which the encoder explicitly encodes input data according to its record–field structure. Furthermore, we introduce *structure-wise attention gates* to facilitate the selection of contextually appropriate content from the encoded data. Since the records in a given input are typically unordered (for example, in Table 2.1, no inherent order exists among the three records in x), the robustness of the proposed encoder is evaluated by shuffling these records. Comparative experiments against high-performing baseline systems on the WikiBio and E2E datasets highlight the distinct advantages of our approach. Additionally, human evaluation confirms that the proposed model achieves strong performance in terms of both readability and informativeness of the generated text.

2.2 Related Works

The research presented in this chapter is aligned with the growing body of work on data-driven D2T, where text is generated directly from structured data [4, 61, 101, 108]. Various types of data-driven text generation systems have been explored in the literature. For instance, Belz [9] employed probabilistic context-free grammars for generating text from structured data, while Chen and Mooney [10] introduced a strategic text generation technique for sportscasting in a simulated soccer game environment.

Among data-driven text generation approaches, Angeli et al. [101] was among the first to propose a domain-independent model, demonstrating its application in weather forecasting. With the introduction of the recurrent neural network language model [107], neural text generation methods have proliferated, finding success across a variety of tasks, ranging from poem generation [122] to image captioning [46, 123, 124].

Following the development of the sequence-to-sequence (seq2seq) model [77] and the introduction of various attention mechanisms [46, 112], encoder–decoder architectures have become a dominant framework in neural text generation. For selective generation tasks, where only specific portions of the input are relevant, Mei et al. [108] proposed an encoder–decoder model with an attention mechanism. In concept-to-text generation, where the goal is to produce natural language descriptions from complex conceptual inputs, encoder–decoder-based models have also demonstrated high accuracy [4, 61, 63, 109]. Moreover, data-driven approaches based on this framework have achieved notable success in dialogue systems [12, 110] and table-to-text generation tasks [2, 47].

2.3 Notations Used in This Work

We adopt a notation scheme that balances the formal definitions introduced in section 1.3 with the conventions commonly used in prominent D2T models [39, 101, 113]. The objective of this work ($\mathcal{T}$) is to generate a text representation $y = y_1 y_2 \dots y_w$, where y_i denotes the i-th word in y, given an input data instance x.

$$\mathcal{T} : \hat{y} \leftarrow \underset{y}{\operatorname{argmax}}\ p(y|x)$$

An input x consists of multiple records, denoted as $\{r_i\}_{i=1}^{n}$. Each record r_j contains a field f^j, defined by a field name n^j and its corresponding value $v^j = v_1^j v_2^j \dots v_w^j$, where v_i^j represents the i-th word in v^j.

2.4 Approach

We propose an LSTM-based [23] encoder–decoder model for the task $\mathcal{T}$, where the encoder structurally represents the input x. The decoder is implemented as a recurrent neural network consisting of two key submodules: (i) *attention gates*, which enable effective content selection, and (ii) a *copy module*, which allows the model to generate rare or unseen words by directly copying them from the input.

2.4.1 Structural Encoder

The design of the proposed encoder is motivated by the inherent structure of the input data. We assume that each input instance consists of two structural levels: (i) *fields*, representing the fine-grained structure, and (ii) *records*, representing the coarse-grained structure. For example, as illustrated in Figure 2.1, the input data x contains two records ($\{r_i\}_{i=1}^{2}$), where each record comprises two field components ($r_1 = (f_1^1, f_2^1)$ and $r_2 = (f_1^2, f_2^2)$).

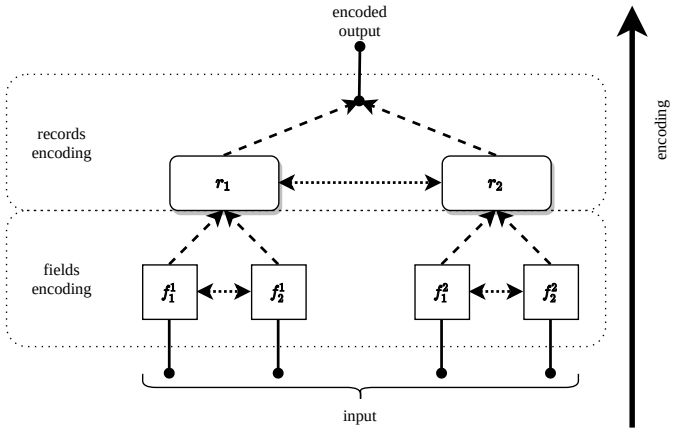


FIGURE 2.1: Structure of input data and bottom-up encoding.

The proposed encoder processes the input data based on its record–field structures (Figure 2.1) using a bottom-up approach. Each structural component (record and field) is encoded through two types of connections, as illustrated by the arrows in Figure 2.1:

- The horizontal dotted arrows represent **horizontal connections**, which capture relationships among adjacent components.
- The dashed arrows represent **hierarchical connections**, which aggregate information from similar components (either records or fields) and pass it to the next stage.

Through this structural encoding process, the proposed encoder effectively captures and encodes the hierarchical relationships inherent in the record and field structures of the input data.

2.4.1.1 Field Encoding

In the field encoding stage of our structural encoder, all field words within a record are jointly encoded into a single vector representation, which ultimately represents the

record. The embeddings of each field's value (words) and field's name are obtained using learnable embedding matrices, as defined below:

$$Z_f[f_k^j] = [E_n[n^j]; E_v[v_k^j]]$$

where $E_*[w]$ denotes the embedding of w, and $[;]$ represents concatenation. Each E_* is a learnable parameter matrix of size $(|v| \times d_{E_*})$, where $|v|$ is the vocabulary size. Note that distinct embeddings are used for field values and field names. Here, $Z_f[f_k^j]$ denotes the encoding of the k-th word in the f^j field together with its corresponding field name. This Z_f representation is then passed to the horizontal connections in the field encoding stage.

Horizontal Connections in Field Encoding. In the field encoding stage, the horizontal connections relate all field words $(Z_f[f_*^j])$ within a record (r^j). Field words (f_*^j) may appear either as a sequence or as a bag-of-words (i.e., orderless or non-sequential). For instance, in Figure 2.1, the *name* field contains two words—*"The"* and *"Punter"*—arranged in a sequence. In contrast, if the *price range* field contains two words—*"high"* and *"medium"*—to indicate that a restaurant offers both high- and medium-priced dishes, these words behave as a bag-of-words.

To accommodate both sequential and orderless characteristics, we construct the horizontal connections of field encoding in a hybrid manner. For sequential data, we employ Bi-LSTM networks [24], whereas for orderless bag-of-words data, we bypass the Bi-LSTM network using skip-interconnections [125]. An illustration of these horizontal connections is shown in Figure 2.2.

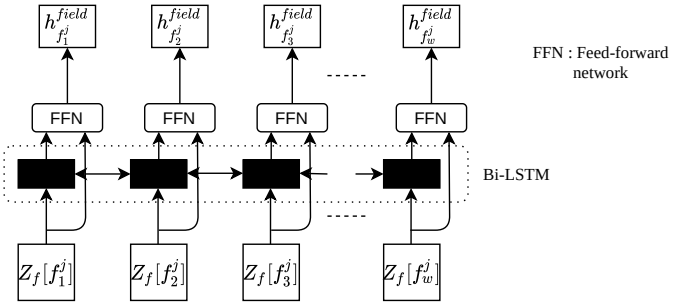


FIGURE 2.2: Horizontal connections in field encoding (for fields within the j-th record).

The outputs from the Bi-LSTM and the skip-interconnections are merged using a feed-forward network, as defined in Equation 2.1. These skip-interconnections are intended to help handle orderless or non-sequential data, a point examined in the experimental section 2.5.6.

The Bi-LSTM output for the field in the j-th record is computed as:

$$h_{f^j} = \text{BiLSTM}(Z_f[f_1^j], \ldots, Z_f[f_w^j])$$

Finally, an affine transformation is applied to the concatenation of $[h_{f_k^j}; f_k^j]$:

$$h_{f_k^j}^{field} = \tanh(\mathbf{W_{fa}}[h_{f_k^j}; f_k^j] + \mathbf{b_{fa}}) \tag{2.1}$$

where $\mathbf{W_{fa}}$ and $\mathbf{b_{fa}}$ are learnable parameters. Thus, $h_{f^j}^{field}$ (and $h_{f_k^j}^{field}$ for the k-th field) represents the output of the horizontal connections in the field encoding stage.

Hierarchical Connections in Field Encoding. The hierarchical connections in field encoding aim to aggregate all field-level information ($h_{f_k^j}^{field}$) within a record (r_j) to generate a detailed representation of the record from the perspective of its constituent fields.

$$h_j^{mp} = \text{maxpool}(h_{f^j}^{field})$$

In these hierarchical connections, we employ both a max-pooling operation and a key–query-based self-attention mechanism [126]. The use of max-pooling is motivated by the intuition that it effectively captures the most salient information of a record from its field-level representations ($h_{f_k^j}^{field}$), similar to its role in convolutional neural networks [127]. Empirically, we find that max-pooling provides a more robust summary representation than using the final state of the underlying horizontal connections, as it considers all intermediate states—an advantage particularly beneficial for long sequences.

For the self-attention mechanism applied at the field-encoding level, we use the field representations ($h_{f_*^j}^{field}$) as keys and the max-pooled record representation (h_j^{mp}) as the query. Following our earlier intuition regarding h_j^{mp}, the query vector in this attention formulation encapsulates the overall essence of the record (r_j). An illustration of these hierarchical connections is shown in Figure 2.3.

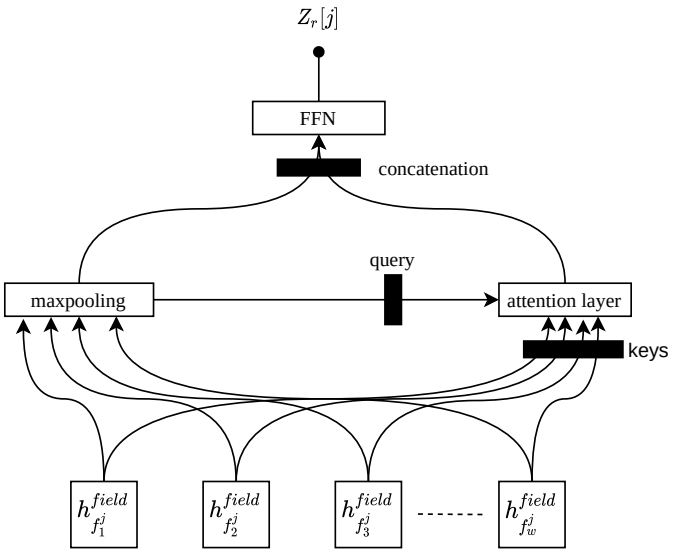


FIGURE 2.3: Hierarchical connections in field encoding (for fields within the j-th record).

For attention scoring, we adopt the widely used concatenative attention mechanism [112]:

$$score^f_{ji} = v^T_f \tanh(\mathbf{W_{fb}}[h^{mp}_j; h^{field}_{f^j_i}])$$
$$\alpha^f_{ji} = \text{softmax}(score^f_{ji})$$
$$c^f_j = \sum_{i=1}^{m} \alpha^f_{ji} \cdot h^{field}_{f^j_i}$$

Here, $score^f \in \mathbb{R}$ denotes the attention score, α^f represents the attention weight, and c^f_j is the resulting attention (or context) vector for the record r_j.

Finally, each record (r_j) is represented through an affine transformation applied to the concatenation of its max-pooled value (h^{mp}_j) and the corresponding context vector (c^f_j):

$$Z_r[j] = \mathbf{W_{fc}}[h^{mp}_j; c^f_j] + \mathbf{b_{fc}}$$

Hence, $Z_r[j]$ constitutes the output of the field encoding for the j-th record. As illustrated in Figure 2.3, the hierarchical connections in field encoding are inherently order-insensitive since both the self-attention and max-pooling operations are independent of data order.

2.4.1.2 Record Encoding

The objective of record encoding is to generate a vector representation of the input data x based on its underlying record structures. The record encoding process closely

follows the same design principles as the field encoding. It comprises both horizontal and hierarchical connections, functioning in an analogous manner to those in the field encoding stage. The output of the horizontal connections in record encoding is denoted as h_j^{record} for the j-th record. The final output of the record encoding stage is the encoded representation of the input data, expressed as $Z_d[x]$.

2.4.2 Decoder

The decoder module comprises two key components: structure-wise (i.e., record and field) attention gates and a copy module. A schematic overview of the proposed decoder architecture is illustrated in Figure 2.4.

2.4.2.1 State Update

The initial decoder state (s_0) is initialized using the output of the record encoding, $Z_d[x]$. To update the decoder state at the t-th time step (s_t), we employ a dual attention mechanism [47, 61] that integrates attention over both record and field structures. The attention operations are defined over the outputs of the horizontal connections from the record and field encoders.

$$\beta_{tj}^r = \text{attention}(s_t, r_j) \tag{2.2}$$

$$\phi_t^r = \sum_{i=1}^{n} \beta_{ti}^r \cdot h_{r_i}^{record} \tag{2.3}$$

Here, β^r denotes the attention weights for the j-th record (r_j) at decoding step t, and ϕ_t^r represents the corresponding record-level context vector. A similar formulation is used to compute the attention weights β^f for the field-level encoding.

To integrate the dual attention mechanism, we modify the effective field attention weights as $\gamma^f = \beta^f \times \beta^r$. Consequently, the effective field-level context vector (ϕ_t^f) is computed using γ^f.

Attention Gates. We introduce two attention gates—one for the record structure and one for the field structure—to control how the decoder accesses the context vectors ϕ_t^r and ϕ_t^f. These gates are defined using the current decoder state (s_t) and the encoded

data representation ($Z_d[x]$) as follows:

$$g_t^r = \sigma(\mathbf{W_{rg}} s_t + \mathbf{U_{rg}} Z_d[x] + \mathbf{b_{rg}})$$
$$g_t^f = \sigma(\mathbf{W_{fg}} s_t + \mathbf{U_{fg}} Z_d[x] + \mathbf{b_{fg}})$$

where g_t^r and g_t^f denote the attention gates for the record and field contexts, respectively. These gates play a crucial role in the model by regulating content selection from the encoded input to the decoder. The gate values dynamically vary between '0' (inhibit) and '1' (exhibit), determining the degree to which contextual information is incorporated during decoding.

The gated context information is computed as:

$$\hat{\phi}_t^r = g_t^r \odot \phi_t^r$$
$$\hat{\phi}_t^f = g_t^f \odot \phi_t^f$$

Finally, the decoder state is updated using the gated record and field context vectors as follows:

$$\tilde{s}_t = \tanh(\mathbf{W_d}[s_t; \hat{\phi}_t^{record}; \hat{\phi}_t^{field}])$$

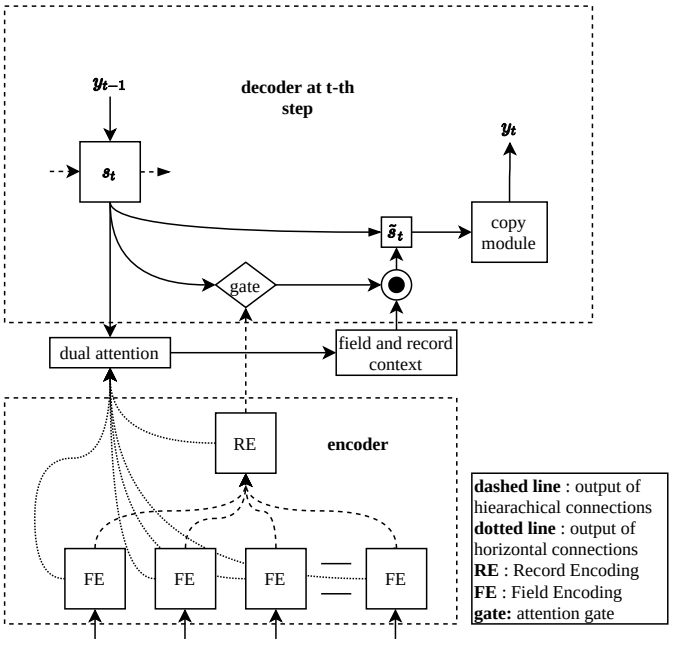


FIGURE 2.4: Proposed encoder–decoder model.

2.4.3 Copy Module

To handle rare or unseen words, we adopt the conditional copy mechanism proposed by Gülçehre et al. [36]. In our model, the binary copy-indicator variable (cp_t) in the conditional copy module is defined as follows:

$$p(cp_t = 0 \mid \tilde{s}_t, Z_d[x]) = \sigma(\mathbf{W_{cp}}[\tilde{s}_t; Z_d[x]])$$

Here, we combine the input encoding $Z_d[x]$ with the current decoder attention state $\tilde{s}_t$ to make the copy indicator cp_t sensitive to both the input context and the current decoding state.

2.4.4 Loss Function

We employ the negative log-likelihood function as the training objective for our model:

$$\text{loss} = -\sum_{i=1}^{m} \log(p(y_t^{(i)} \mid x^{(i)}, y_{<t}^{(i)})).$$

Two important implementation details are worth noting: (i) we use a $\langle$unk$\rangle$ token whenever an out-of-vocabulary word appears in a field; and (ii) we do not share embeddings between field names and field values.

2.5 Experiments

The experiments in this work are conducted on two widely used benchmark datasets for D2T: the WikiBio dataset and the E2E dataset.

2.5.1 Baselines and Metrics

We compare the proposed model against three baseline systems:

1. **Baseline 1:** A vanilla seq2seq model based on the standard sequence-to-sequence architecture [77] with a concatenative attention mechanism [112].

2. **Baseline 2:** To analyze the role of attention mechanisms in our model, this baseline is created by removing all attention components from the decoder of our proposed system.

3. **Baseline 3:** A transformer-based encoder–decoder model [30], trained from scratch for our experiments.

In addition to these three baselines, we also consider two D2T generation systems that have achieved strong performance on both the WikiBio and E2E datasets according to automatic evaluation metrics.

1. **Liu et al. [61]:** This model employs an encoder–decoder architecture with a field-gating encoder that introduces structure-aware representations and a dual attention-based decoder. Unlike our proposed encoder, which hierarchically encodes data based on record–field relationships, Liu et al. [61] encodes record and field information in parallel using separate gating structures.

2. **Nie et al. [113]:** This model introduces an encoder–decoder framework with self-attention mechanisms [30] for handling both sequential and non-sequential data. In contrast to our approach—which incorporates both hierarchical and horizontal connections—Nie et al. [113] primarily focuses on hierarchical self-attention.

To evaluate the quality of the generated text, we employ three widely used automatic evaluation metrics: BLEU [88], ROUGE [90], and METEOR [89][2]. Additionally, we use the PARENT metric [39], which assesses text quality with a particular focus on the *informativeness* aspect of *faithfulness*—especially the precision component of the metric—by jointly considering both the input data and the reference text. In contrast, BLEU, ROUGE, and METEOR evaluate generation quality solely against reference texts. Dhingra et al. [39] further showed that PARENT shows a stronger correlation with human judgments than other automatic evaluation metrics.

Since previous studies [8, 104–106] have shown that automatic evaluations alone are insufficient for reliably assessing D2T generation quality, we also conduct human evaluations on the outputs generated by our system.

2.5.2 Parameter Settings

The dimensions of both field word and field name embeddings are set to 200. Separate embeddings are used for field names and field values (i.e., vocabulary words). The model

[2]For BLEU, ROUGE, and METEOR, we use the official implementation available at: https://github.com/tuetschek/e2e-metrics

is trained using the Adam optimization technique [128] with an initial learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. All BiLSTM models have a depth of 2, and dropout is fixed at 0.5 across all layers. Most hyperparameters are tuned on predefined validation sets. During text generation, we employ beam search decoding with a beam size of 4. For Baseline 3, we use a standard Transformer architecture with a stack size of 6 layers in both the encoder and the decoder. For the WikiBio dataset, vocabulary (word type) and field type sizes are selected following the settings of Lebret et al. [4]. In most experiments, the batch size is fixed to either 32 or 64, and training continues for up to 120,000 steps. All experiments are conducted using an NVIDIA GeForce RTX 2080 Ti graphics card.

TABLE 2.2: Overview of the WikiBio and E2E dataset statistics.

dataset	instances	total words	tokens / sentence	sentences / instance	references / instance
WikiBio	728K	400K	26.1	1.0	1.0
E2E	50.5K	5.2K	14.3	1.5	8.1

TABLE 2.3: Results from WikiBio dataset.

model	BLEU	ROUGE-L	METEOR	PARENT
baseline 1	0.338	0.418	0.271	0.463
baseline 2	0.348	0.445	0.262	0.455
baseline 3	0.381	0.486	0.340	0.407
Liu et al. [61]	0.447	0.528	0.363	0.538
Nie et al. [113]	0.450	0.522	0.371	0.527
proposed method	**0.465**	**0.566**	**0.397**	**0.540**

2.5.3 WikiBio Dataset and Results

Lebret et al. [4] introduced the WikiBio dataset, which is constructed from biographical articles on Wikipedia. Table 2.2 presents the statistics of the WikiBio dataset. As shown in arg1Table 2.3, our model achieves higher scores than the evaluated baselines and previously reported comparison systems across the considered automatic evaluation metrics, including those measuring *informativeness*. Representative examples of texts generated by the proposed system are provided in Table 2.4.

2.5.4 E2E Dataset and Results

Novikova et al. [8] introduced the E2E dataset, which focuses on the restaurant domain (see Table 2.2). As shown in the comparison results in Table 2.5, our model consistently outperforms the baselines and other reported systems across nearly all evaluation metrics. The only exception is the model by Liu et al. [61], which performs slightly better

TABLE 2.4: A sample of generated texts from the proposed model on WikiBio dataset (wrong words generation, missing information)

A Sample of Generated Texts on WikiBio Dataset	
Instance 1	
input	name [pietro micheletti], birthdate [19 october 1900], deathdate [25 march 2005], birthplace [maciano di pennabilli, italy], deathplace [maciano di pennabilli, italy], allegiance [italy italy], branch [italian army], serviceyears [1917–1925], rank [major], battles [world war i, fiume's war], laterwork [farmer manager], award ["military cross", "ordine di vittorio veneto", "knight of the order of merit of the italian republic"], articletitle [pietro micheletti]
reference	pietro micheletti (19 october 1900 – 25 march 2005) was an italian military commander.
generated text	pietro micheletti (october 19, 1900 – march 25, 2005) was a italian army manager.
Instance 2	
input	name [jason buxton], nationality [canadian], occupation [film director, screenwriter], knownfor ["blackbird"], articletitle [jason buxton]
reference	jason buxton is a canadian film director and screenwriter.
generated text	jason buxton is an canadian film director and screenwriter.
Instance 3	
input	name [bert crossthwaite], fullname [herbert crossthwaite], birthdate [4 april 1887], birthplace [preston, england], deathdate [20 may 1944], deathplace [birmingham, england], position [goalkeeper], years [1906–1907, 1907–1909, 1909–1910, 1910–1914, 1914–1915], clubs [blackpool, fulham, exeter city, birmingham, stoke], caps [0, 1, 2, 49, 0], goals [0, 0, 0, 0, 0], articletitle [bert crossthwaite]
reference	herbert crossthwaite (4 april 1887 – 20 may 1944) was an english footballer who played as a goalkeeper.
generated text	herbert bert crossthwaite (4 april 1887 – 20 may 1944) was an english footballer who played as a goalkeeper.

(by approximately 1%) in terms of *informativeness*, as indicated by the PARENT metric for the E2E dataset. Representative examples of the generated text are provided in Table 2.6.

TABLE 2.5: Results on E2E dataset.

Model	BLEU	ROUGE-L	METEOR	PARENT
Baseline 1	0.517	0.520	0.344	0.569
Baseline 2	0.534	0.572	0.350	0.572
Baseline 3	0.568	0.594	0.425	0.642
Liu et al. [61]	0.653	0.614	0.428	**0.726**
Nie et al. [113]	0.662	0.612	0.399	0.663
Proposed Method	**0.675**	**0.683**	**0.442**	0.716

2.5.5 Analysis from Ablation Tests

To evaluate the effectiveness of the copy module, attention gates, and encoder in our model, we conduct a component ablation study using BLEU and PARENT scores as evaluation metrics.

Ablation of Copy Module. Table 2.7 presents the ablation results for the copy module. The reported BLEU and PARENT scores clearly indicate that the copy module plays a crucial role in improving the generated text, enhancing both *informativeness* and *fluency* across both datasets.

TABLE 2.6: A sample of generated texts from the proposed model on E2E dataset (wrong words generation, missing information)

A Sample of Generated Texts on E2E Dataset	
Instance 1	
input	name [Giraffe], eatType [restaurant], food [French], area [city centre], familyFriendly [yes], near [Raja Indian Cuisine]
reference 1	the giraffe is a restaurant - family-friendly serving french food in the city centre, near the raja indian cuisine.
reference 2	giraffe, a family-friendly restaurant in the city centre serving french food, is located near raja indian cuisine.
reference 3	giraffe is a restaurant located in the city centre near the raja indian cuisine. it offers french cuisine with a family-friendly atmosphere.
reference 4	city centre-based restaurant near raja indian cuisine, giraffe is family-friendly and serves french cuisine.
reference 5	giraffe is a restaurant that offers french food with a family-friendly atmosphere. it is located in the city centre near the raja indian cuisine.
generated text	giraffe is a french restaurant in the riverside area near raja indian cuisine. it is not family-friendly.
Instance 2	
input	name [Cocum], eatType [pub], near [The Sorrento]
reference 1	cocum is a pub by the sorrento.
generated text	cocum is a pub near the ranch.
Instance 3	
input	name [Giraffe], eatType [pub], near [The Bakers]
reference 1	giraffe is a pub found near the bakers.
reference 2	a pub named giraffe is near the bakers.
generated text	giraffe is a pub near the bakers.

TABLE 2.7: Ablation of copy module (based on BLEU and PARENT score).

	WikiBio		E2E	
	BLEU	PARENT	BLEU	PARENT
with copy module	0.465	0.540	0.675	0.716
without copy module	0.424	0.527	0.619	0.682

TABLE 2.8: Ablation of attention gates (based on BLEU and PARENT score).

	WikiBio		E2E	
	BLEU	PARENT	BLEU	PARENT
with attention gates	0.465	0.540	0.674	0.716
without attention gates	0.458	0.513	0.680	0.694

Ablation of Attention Gates. To evaluate the effectiveness of the attention gates, we perform ablation tests by removing them from the model. As shown in Table 2.8, we observe minimal to no improvement in BLEU scores. However, the inclusion of attention gates leads to a noticeable improvement in PARENT scores, indicating a noticeable improvement in *informativeness within faithfulness*. Furthermore, qualitative analysis reveals that the attention gates contribute to more informative and relevant text generation. Table 2.9 presents the results of this qualitative analysis. Although our model

TABLE 2.9: A qualitative analysis for the role of attention gates in our model (wrong word, word is not available in input) [E2E dataset contains several gold-standard references for a single input data, but due to space constraint only one reference is given.].

WikiBio	
input	name[myles wilder], birth-date[january 28, 1933], birth-place[new york city, new york], death_date[april 20, 2010 -lrb- age 77 -rrb-], death-place[temecula, california], occupation[television writer and producer], spouse[bobbe wilder -lrb- survives him -rrb-], article-title[myles wilder]
reference	myles wilder -lrb- january 28, 1933 - april 20, 2010 -rrb- was a television comedy writer and producer.
without attention gates	myles wilder -lrb- 28, 1933 april, -rrb- held a television comedy writer and.
with attention gates	myles wilder -lrb- january 28, 1933 – april 20, 2010 -rrb- was an american television writer and producer.
E2E	
input	name[Blue Spice], eatType[restaurant], food[English], area[riverside], familyFriendly[yes], near[Rainbow Vegetarian Cafe]
one reference	there is a restaurant that provides food and is children friendly, near rainbow vegetarian cafe and the riverside and is called blue spice.
without attention gates	there is a soup restaurant that provides food and good spice. friendly, near rainbow vegetarian cafe and the riverside and blue.
with attention gates	near the rainbow vegetarian café in the riverside area is a restaurant called blue spice that serves english food and is children friendly.

TABLE 2.10: Ablation of encoder connections (based on BLEU and PARENT score).

	WikiBio		E2E	
	BLEU	PARENT	BLEU	PARENT
With all connections	0.465	0.540	0.675	0.716
Without horizontal connections	0.369	0.483	0.532	0.580
Without hierarchical connections	0.414	0.498	0.581	0.673

occasionally produces minor errors regardless of whether the attention gates are used, the version with attention gates consistently generates text with fewer incorrectly inserted words, thereby confirming their positive impact on improving informativeness in faithfulness.

Ablation of Encoder Connections. Through ablation tests, we analyze the effectiveness of our encoder connections—specifically, the horizontal and hierarchical connections. Table 2.10 reports the results of these ablation experiments. The findings indicate that the proposed model performs better when both connections are present, showing clear improvements in informativeness within faithfulness as well as in fluency.

2.5.6 Analysis of Model with Shuffled Input

As discussed earlier, to enable our proposed encoder to handle both sequential and non-sequential (orderless) data, we introduced skip-interconnections to effectively manage these variations. More specifically, the horizontal connections are responsible for encoding sequential data, while the hierarchical connections play a crucial role in encoding non-sequential data. The skip-interconnections then integrate the outputs from both

the horizontal and hierarchical connections, mitigating any model bias toward record or field order.

In this section, we investigate the role of skip-interconnections by performing experiments with randomly shuffled input records. The objective of this study is to depict the robustness and effectiveness of the proposed encoder when dealing with shuffled data. The evaluation is conducted using both intrinsic and extrinsic analyses.

TABLE 2.11: Ablation test of skip-interconnections with shuffling records in input data.

Model	WikiBio		E2E	
	BLEU	PARENT	BLEU	PARENT
With skip-interconnection	0.421 ± 0.011	0.493 ± 0.017	0.637 ± 0.009	0.682 ± 0.013
Without skip-interconnection	0.413 ± 0.024	0.490 ± 0.044	0.628 ± 0.021	0.652 ± 0.036

Extrinsic Evaluation. In this experiment, we conduct an ablation study on the skip-interconnections by randomly shuffling the records in the input data for both datasets. For each dataset's test set, the record shuffling is performed five times. Table 2.11 presents the effectiveness of the proposed encoder's skip-interconnections, as reflected by the low fluctuation (standard deviation) values observed in both the PARENT and BLEU metrics.

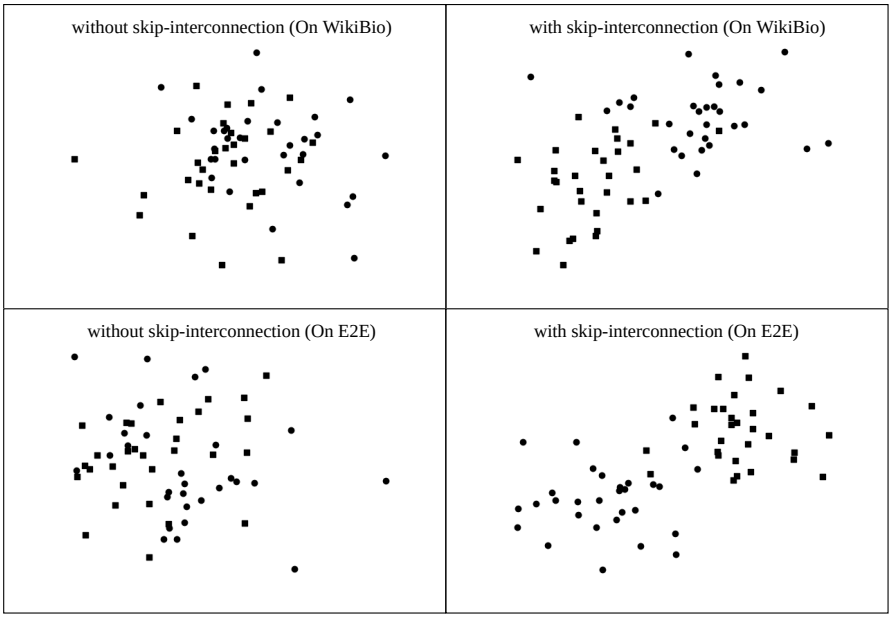


FIGURE 2.5: t-SNE plots for encoder representations on shuffled input data (circled and squared points represent two data different data instances).

Intrinsic Evaluation. To further examine the impact of skip-interconnections on our model's ability to handle shuffled input data, we visualize the encoded representations

using t-SNE plots [129]. Two random data instances are sampled from each of the two datasets (WikiBio and E2E), and each instance is shuffled into approximately 30 different arrangements. We then generate t-SNE plots of the encoded representations of these shuffled inputs produced by our encoder. As shown in Figure 2.5, the well-disentangled encoded representations of the shuffled data depict the effectiveness of the skip-interconnections in preserving structural consistency.

2.5.7 Human Evaluation

For the human-based evaluation, four university graduate students from different academic backgrounds were selected as annotators. A sample of 150 generated texts from the proposed model was chosen for each of the two datasets (E2E and WikiBio). Along with each generated text, the corresponding input data and reference text were also provided to the annotators. Each instance was evaluated by at least three human annotators. In this evaluation, we focus on two key qualities of the generated texts—*informativeness* (representing the *informativeness within faithfulness*) and *readability* (or fluency) [52]. *informativeness* assesses whether the generated text accurately and completely conveys the relevant content from the input data, while *readability* measures the fluency, clarity, and overall linguistic quality of the generated output. Annotators rated each text on a 5-point scale for both dimensions, where a score of 5 indicates the highest quality. The results of the human evaluation are presented in Figure 2.6, along with inter-annotator agreement measured using Krippendorff's alpha coefficient[3]. The evaluation shows that the texts generated for the WikiBio dataset achieved higher scores in both readability and informativeness (within faithfulness) compared to those from the E2E dataset. The instructions provided to the annotators for conducting the human evaluation are presented in the text box below.

Instructions to Annotators

Participation in this annotation task is entirely voluntary. By proceeding with the evaluation, you acknowledge that you have read and understood the instructions and agree to participate in the study.

For each instance, you will be provided with the input data, a reference text, and a generated text. Your task is to evaluate the generated text on the following two dimensions:

—

[3] https://www.nltk.org/api/nltk.metrics.html

- **Informativeness**: Assess whether the generated text accurately and completely conveys the information contained in the input data without introducing unsupported facts.
- **Readability**: Assess the fluency, grammatical correctness, coherence, and overall clarity of the generated text.

Assign a score from 1 to 5 for each dimension, where 1 indicates very poor quality and 5 indicates excellent quality. Evaluate each dimension independently. For informativeness, focus primarily on the input data when determining whether the generated text is factually accurate and complete. In addition to the numerical rating, provide a brief justification for your assessment (in the `comment` box), highlighting any important observations regarding *informativeness* and *readability*.

Please make your judgments independently and rely on your own assessment. Do not discuss the evaluation instances with other annotators or alter your ratings based on others' opinions.

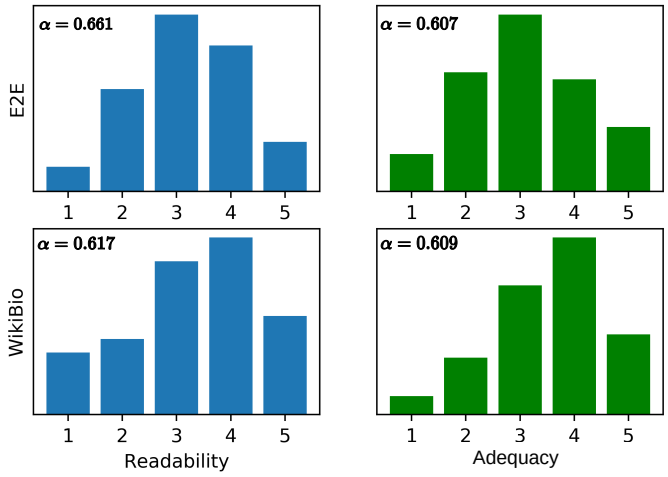


FIGURE 2.6: Average human rating of texts generated from the proposed model (**top left**:measures readability in E2E, **top right**: informativeness in E2E, **bottom left**: readability in WikiBio, **bottom right**: informativeness in WikiBio). α denotes the Krippendorff's alpha coefficient.

2.6 Conclusion

In this chapter, we presented a structural encoder for an encoder–decoder-based, pre-LLM D2T model that effectively encodes record–field structured input data. Experiments on WikiBio and E2E suggest that the proposed encoder is effective for both

the sequential and non-sequential data representations considered in this study. By integrating attention gates, the model is able to selectively extract and utilize relevant content from the encoded representations. Evaluation on the WikiBio and E2E datasets clearly shows that the proposed structural encoding–based pre-LLM model improves the *informativeness within faithfulness* of generated texts, confirming its effectiveness in producing accurate and high-quality data-driven narratives.

Chapter 3

Evaluating and Enhancing Faithfulness in LLM-based D2T

In the previous chapter, we discussed structural encoding for record–field structured input data using pre-LLM models based on an LSTM-based encoder–decoder architecture. From this chapter onward, the focus shifts to LLM-based D2T. This chapter[1] presents a broad empirical evaluation of Large Language Models (LLMs) for Data-to-Text generation (D2T), with a particular focus on assessing and enhancing *faithfulness*, with special attention to *informativeness*. It systematically examines twelve LLMs from five major open-source families—BART, T5, BLOOM, OPT, and Llama 2—across five widely used D2T datasets, evaluating their performance across three key dimensions: readability, informativeness, and faithfulness. Through both automatic and human evaluations, the results suggest that, within the evaluated models and datasets, larger model sizes are generally associated with improved readability and informativeness, although corresponding improvements in faithfulness are not consistently observed. To address this limitation, the chapter introduces a *source-reference divergence-weighted training* strategy that adaptively reweights training instances based on their divergence, yielding improvements across the evaluated dimensions in the reported experimental settings. Overall, this work contributes empirical insights into LLM behavior in D2T and proposes strategies that improve faithfulness under the evaluated experimental conditions.

[1] The work presented in this chapter is based on our paper: Mahapatra, J. and Garain, U., 2025. A Comprehensive Performance Evaluation of LLMs for Data-to-Text Generation and Divergence-Weighted Training. *ACM Transactions on Intelligent Systems and Technology*, Association for Computing Machinery. [94]

3.1 Background

The recent progress in D2T can be largely attributed to the advancements in LLMs, which have demonstrated remarkable generative capabilities [75, 130]. While LLMs have become widely adopted as foundational models for D2T [6, 18], the literature appears to contain relatively few studies that systematically examine their performance across multiple D2T datasets, model families, and evaluation dimensions. Such an evaluation may help inform future research and development efforts in LLM-based D2T.

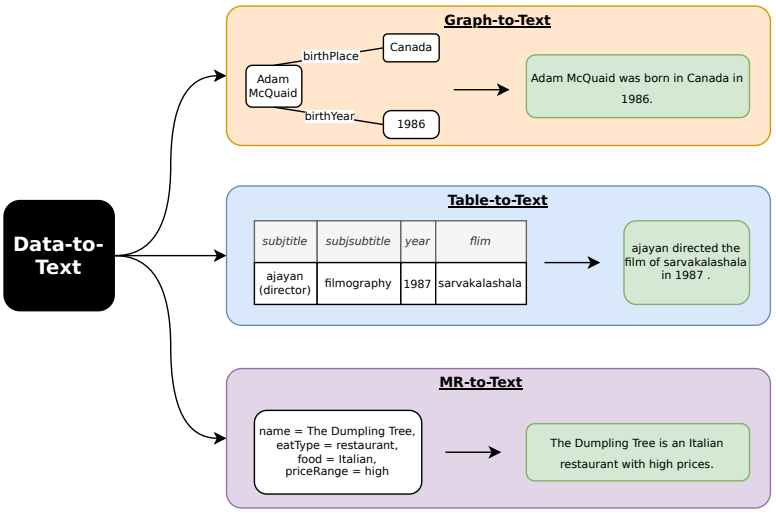


FIGURE 3.1: Three primary types of D2T: graph-to-text (left), table-to-text, and MR-to-text.

The chapter presents a broad performance evaluation of LLMs for D2T, addressing an underexplored aspect of existing research. To capture a diverse range of LLM behavior, the study includes twelve models from five widely adopted open-source families: BART [31], T5 [75], OPT [99], BLOOM [72], and Llama 2 [130]. The evaluation spans all three primary D2T paradigms, utilizing five popular datasets—E2E [8] and ViGGo [62] (MR-to-text), WikiTableText [2] (table-to-text), and DART [1] and WebNLG [7] (graph-to-text).

The study examines three fundamental qualities for assessing D2T performance: *readability* (fluency and coherence), *informativeness* (effective information transformation), and *faithfulness* (factual accuracy) [54]. Six commonly used automatic evaluation metrics are employed: BLEU [88] and METEOR [89] for readability; BERTSCORE [131], MOVERSCORE [132], and PARENT [39] for informativeness; and BARTSCORE [133] for faithfulness. These evaluations are further complemented by human assessments, which are commonly regarded as an important reference for evaluating text quality.

The chapter also discusses the robustness of LLM performance across different fine-tuning and decoding strategies, and explores a critical phenomenon in D2T—*source-reference divergence*, where input data semantically diverges from its reference text. To mitigate the negative impact of such divergence, a *source-reference divergence-weighted training* strategy is introduced. A schematic overview of the evaluation scope is illustrated in Figure 3.2.

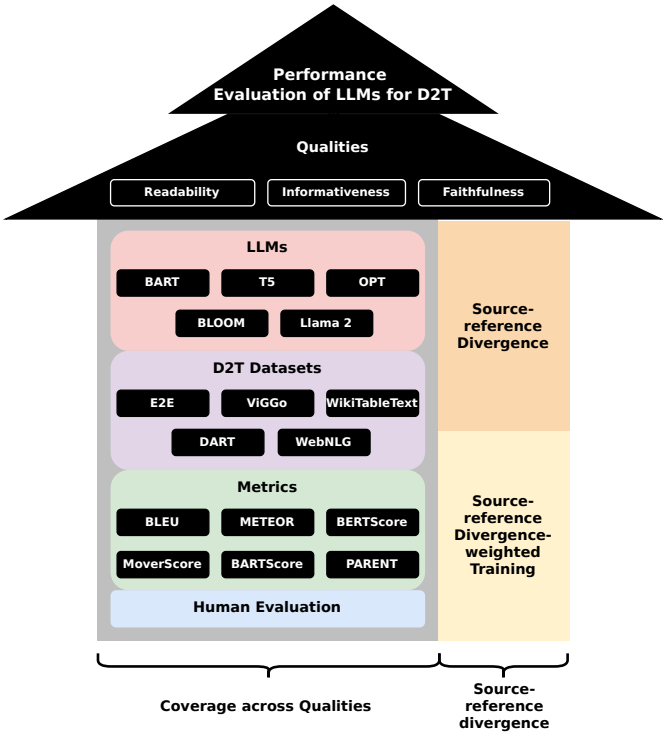


FIGURE 3.2: A schematic overview of our extensive performance evaluation of LLMs for D2T, covering multiple LLM families, datasets, and metrics, with an additional analysis considering source-reference divergence.

The empirical findings presented in this chapter reveal several important insights. It is observed that increasing the model size (i.e., the number of parameters) of LLMs generally enhances *readability* and *informativeness*, with the Llama 2 family outperforming all others across these dimensions. However, *faithfulness*, as measured by BARTSCORE, does not consistently improve with larger model size and, in some cases, even declines—highlighting the need for caution when prioritizing factual accuracy. No single LLM family demonstrates consistent superiority in faithfulness across all D2T datasets, although human evaluations consistently favor larger models in terms of readability, informativeness, and faithfulness. The evaluation indicates that the observed trends remain largely consistent across the examined fine-tuning and decoding strategies. Additionally, the results indicate that as *source-reference divergence* increases, performance in all three key qualities—readability, informativeness, and faithfulness—declines across

all model families, with model size offering limited mitigation. To address this limitation, the proposed *source-reference divergence-weighted training* strategy improves LLM performance in D2T across these evaluation dimensions.

The remainder of this chapter is organized as follows. Section 3.2 discusses the current representative in evaluation benchmarks for LLMs, highlighting recent developments and related research efforts in this area. Section 3.3 provides an overview of LLMs, D2T tasks, and the concept of source-reference divergence. Section 3.4 describes the models, datasets, and experimental configurations used in the study. Section 3.5 presents a detailed performance evaluation of LLMs on D2T tasks, encompassing the three key qualities along with human evaluation results. Section 3.6 depicts the robustness of the evaluation with respect to different fine-tuning and decoding strategies. Section 3.7 discusses the impact of varying levels of source-reference divergence on LLM performance. Section 3.8 provides two case studies that illustrate the observed evaluation outcomes. Section 3.9 details the proposed source-reference divergence-weighted training approach. Finally, Section 3.10 concludes the chapter with a summary of the key findings and potential directions for future research.

3.2 Related Work

Performance evaluation of LLMs has become increasingly critical as these models are widely applied across various language generation tasks. To meet this need, several benchmarks have been developed to assess diverse aspects of LLM behavior, including reasoning ability, consistency, faithfulness, and hallucination. Among the earliest large-scale evaluation frameworks, HELM (Holistic Evaluation of Language Models) [134] introduced a unified platform for assessing LLMs across multiple scenarios, emphasizing evaluation metrics such as accuracy, robustness, fairness, and efficiency. Subsequently, MT-Bench [135] was proposed to evaluate multi-turn conversational capabilities in LLMs using both human and GPT-based assessments, providing detailed insights into dialogue coherence and reasoning. HaluEval [136] focuses specifically on hallucination detection in LLM-generated text across different domains and tasks, thereby offering a benchmark for measuring factual consistency. Similarly, TruthfulQA [97] evaluates whether models generate truthful responses to questions designed to elicit common misconceptions, thus serving as a valuable measure of factual accuracy and hallucination resistance. While these benchmarks have significantly advanced the evaluation of LLM reliability, most focus primarily on question answering, summarization, or dialogue tasks. In contrast, structured-to-text generation—an equally important area within natural language generation—has received comparatively limited attention in existing LLM evaluation

frameworks. Recent efforts have begun to explore how LLMs perform when required to generate coherent, faithful, and informative narratives directly from structured inputs. Building upon this direction, the current study seeks to contribute toward addressing this gap by presenting an extensive performance evaluation of LLMs on D2T tasks. The work evaluates LLMs across multiple datasets and key evaluation dimensions—including *readability*, *informativeness*, and *faithfulness*—to provide additional empirical insights into how these models handle structured data transformation in comparison to conventional natural language generation tasks.

3.3 Preliminaries

3.3.1 LLM

As discussed in Chapter 1, LLMs have substantially influenced recent developments in language modeling. For brevity and reference, this section revisits key concepts and provides a few additional details relevant to the present discussion. LLMs fundamentally differ from earlier predictive language modeling approaches [107] due to their massive parameterization and extensive self-supervised pre-training on large-scale text corpora, enabling them to capture rich linguistic, semantic, and contextual knowledge. These characteristics are considered potential factors underlying the ability of LLMs to generalize across diverse tasks and application domains, including D2T [93, 130]. Recent advancements in parameter-efficient fine-tuning techniques, such as LoRA and QLoRA [85], have further increased the practical applicability of LLMs for text generation tasks by reducing computational requirements in many settings. LLMs can be broadly categorized into two major types, as illustrated in Figure 3.3: *bidirectional* and *unidirectional* LLMs [14]. In bidirectional models, predictions leverage both left and right contextual information, allowing for a thorough understanding of surrounding words. The primary objective of a bidirectional LLM is to predict a masked token (w_i) based on its masked context, formally defined as:

$$w_i \leftarrow \underset{w}{\text{argmax}} \ \Pr(w| \underbrace{w_1 \dots w_{i-1} w_{\text{mask}} w_{i+1} \dots w_n}_{\text{masked context}})$$

In contrast, unidirectional LLMs [99] operate in a left-to-right manner, generating text sequentially by conditioning each prediction on the preceding context. The primary objective of such models is to predict the next word (w_n) given the prior sequence of tokens, thereby modeling the forward dependency structure essential for natural language

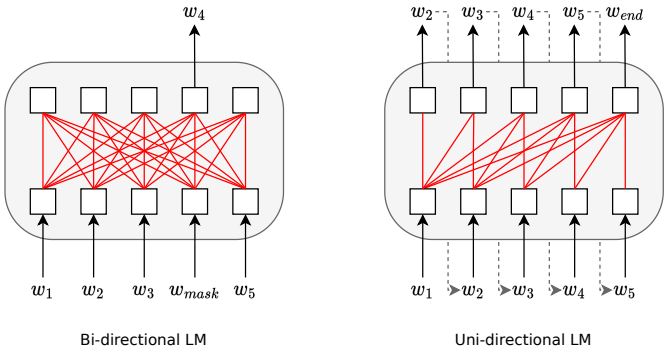


FIGURE 3.3: Bidirectional and unidirectional LLM architectures, with red lines indicating attention mechanisms.

generation. This approach can be formally expressed as follows:

$$w_n \leftarrow \underset{w}{\operatorname{argmax}} \ \Pr(w|\underbrace{w_1 w_2 \dots w_{n-1}}_{\text{prior context}})$$

LLMs are primarily built upon two fundamental architectural paradigms, as illustrated in Figure 3.4: the *encoder–decoder* architecture [31] and the *decoder-only* architecture [99]. In the encoder–decoder framework, the encoder is designed to process and represent the input sequence, while the decoder autoregressively generates the corresponding output text. This architectural distinction plays an important role in determining how contextual information is encoded and utilized during generation.

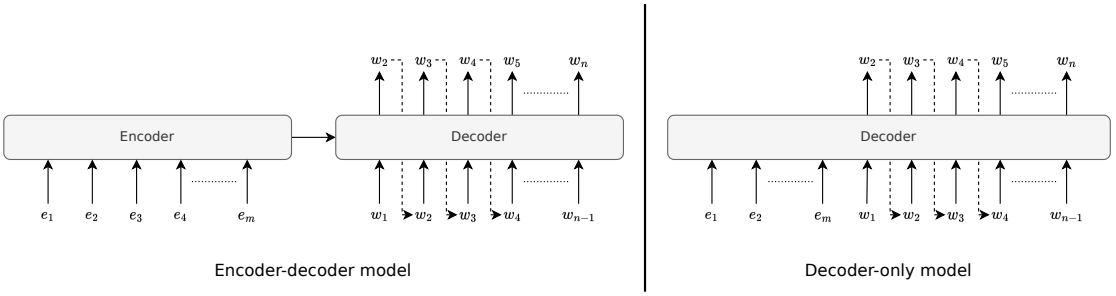


FIGURE 3.4: Two popular LLM architectures: encoder-decoder and decoder-only.

3.3.2 D2T

For the sake of completeness, this section revisits some key details of D2T; for a more broad discussion, readers are referred to Chapter 1. D2T [35] aims to transform semi-structured data—commonly referred to as source data—into coherent natural language text. Typically, a D2T model is trained on a dataset composed of source–reference pairs, $\mathcal{D} = \{(s, r)\}_{\in \mathcal{D}}$, where s denotes the semi-structured source input, and $r = r_1 r_2, \dots, r_{|r|}$

represents the corresponding reference description, with each token r_i belonging to the vocabulary set $\mathcal{V}$, i.e., $r_i \in \mathcal{V}$ [39].

Any predictive D2T model (including LLM-based systems) generally involves three main phases: *modelling*, *training* (or fine-tuning), and *inference*. During the modelling phase, a predictive model $\mathcal{M}_\theta$—parameterized by θ—is defined to capture the mapping between source data and target text, i.e., $\mathcal{V}^* \sim \mathcal{M}_\theta(s)$. In the training or fine-tuning phase, the model $\mathcal{M}_\theta$ is optimized on the dataset $\mathcal{D}$, typically using maximum likelihood estimation [14, 35], as shown in Equation 3.1 [15, 16]:

$$\theta^* \leftarrow \underset{\theta}{\operatorname{argmax}} \sum_{(s,r)\in\mathcal{D}} \mathcal{L}(\mathcal{M}_\theta(s), r) \tag{3.1}$$

Here, $\mathcal{L}$ denotes the loss function, most commonly the cross-entropy loss. During inference, the trained model $\mathcal{M}_{\theta^*}$ generates an output text g—an approximation of r—from the input s using a suitable decoding strategy [79], as defined in Equation 3.2:

$$g \leftarrow \underset{\mathcal{V}^*}{\text{decoding}}\ \mathcal{M}_{\theta^*}(s) \tag{3.2}$$

3.3.3 Source-Reference Divergence

Several studies [39, 68] have identified issues related to information disparity between source data (s) and corresponding reference data (r) across various D2T tasks. These issues, referred to as source-reference divergence [39, 54], are illustrated in Figure 3.5. The source-reference divergence often arises from data collection processes and the nature of the tasks [7, 58]. Such divergence is common and difficult to avoid, especially in D2T scenarios [39], making it important to evaluate model performance in its presence.

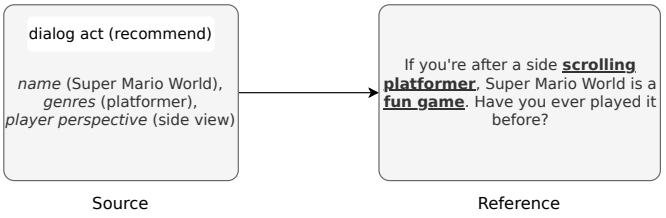


FIGURE 3.5: An example of source-reference divergence, taken from the WikiTableText [2] dataset. The reference text includes two additional facts (bolded and underlined) - 'scrolling performer' and 'fun game' - that are absent in the corresponding source data.

3.4 Models, Datasets and Experimental Settings

This section outlines the models, D2T datasets, and experimental settings employed in our study.

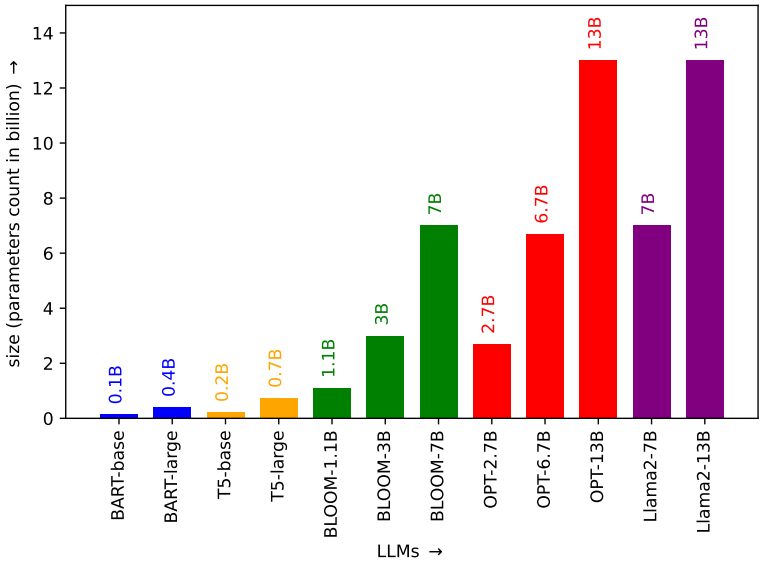


FIGURE 3.6: Model sizes of the twelve LLMs from five families: BART (blue), T5 (orange), BLOOM (green), OPT (red), and Llama 2 (purple).

3.4.1 Models

We include twelve popular LLMs from five well-known families: BART, T5, BLOOM, OPT, and Llama 2 (shown in Figure 3.6), all models are downloaded from Hugging Face [137]. BART [31] is an encoder-decoder model pre-trained with text denoising techniques, and we use BART-base (139M parameters) and BART-large (406M parameters). T5 [75], also an encoder-decoder model, aims to unify multiple NLP tasks and is pre-trained using span-corruption on the C4 dataset; we test with T5-base (223M parameters) and T5-large (738M parameters). BLOOM [72] is a decoder-only model trained on the ROOT dataset; we use BLOOM-1B (1B parameters), BLOOM-2.7B (2.7B parameters), and BLOOM-6.7B (6.7B parameters). OPT [99], another decoder-based model trained on the RoBERTa dataset, the Pile, and Reddit; we test with OPT-2.7B (2.7B parameters), OPT-6.7B (6.7B parameters), and OPT-13B (13B parameters), whose performance comparable to GPT-2 and GPT-3. Lastly, Llama 2 [130], the successor to Llama, is trained through pre-training, fine-tuning, and reinforcement learning, and we use Llama2-7B (7B parameters) and Llama2-13B (13B parameters). We selected twelve LLMs from five model families for two main reasons. First, each family is trained on

different pretraining data and exhibits variations, making them well-suited for a comprehensive evaluation. Second, relying on a single LLM is inadequate for model size comparisons; thus, we chose families with at least two models each.

3.4.2 Datasets

We included five popular D2T datasets, each representing key task types: E2E and ViGGO for MR-to-text, WikiTableText for table-to-text, and DART and WebNLG for graph-to-text. Table 3.1 summarizes key statistics and characteristics of each datasets. All datasets are downloaded from Tabgenie [138] and HuggingFace [137]. The E2E dataset [8] consists of slot-value pairs from the restaurant domain, while the ViGGO dataset [62] offers diverse dialogue acts from video games. WikiTableText [2] provides table-to-text pairs derived from Wikipedia tables. DART [1] generates text from knowledge graph triplets, and WebNLG [7] focuses on RDF-to-text generation from DBpedia triplets, known for its diverse, crowdsourced outputs. These five datasets were chosen for their diversity among themselves and collective representation of all three primary types of D2T. As shown in Table 3.1, they exhibit significant differences in lexical richness, with E2E having the lowest and WikiTableText the highest ratio of unique to total tokens. Additionally, they provide a balanced representation of open-domain and closed-domain knowledge.

TABLE 3.1: Key statistics for the five D2T datasets used in this study.

dataset	D2T types	domain	dataset size	source (linearized)			reference		
			(# instances)	average length	unique tokens	total tokens	average length	unique tokens	total tokens
E2E	mr-to-text	closed	36,856	27.3	125	1M	20.8	4.5K	885K
ViGGo	mr-to-text	closed	6,900	29.9	618	206K	21.5	4.4K	148K
WikiTableText	table-to-text	open	13,318	35.2	29K	469K	13.9	24K	185K
DART	graph-to-text	open	70,524	34.8	44K	2.5M	19.3	45K	1.5M
WebNLG	graph-to-text	open	38,872	30.4	7K	1.2M	20.1	19K	905K

3.4.3 Experimental Settings

In our experiments, we fine-tuned all LLMs for D2T tasks using the QLoRA (Quantized version of Low-Rank Adapter) technique [85], a parameter-efficient fine-tuning approach with a learning rate of $1e^{-04}$. All models were quantized to 4-bit precision. Both QLoRA and quantization configurations are detailed in code listing 1. Each LLM was fine-tuned on all five datasets for 3 epochs, maintaining a consistent text length of 256 for both source data and generated text. For optimization, we employed the AdamW optimizer [139], operating at `torch.bfloat16` precision in QLoRA setups. Text generation utilized a beam-search decoding strategy with a beam size of 5. All experiments were conducted on a local system equipped with an NVIDIA RTX A6000 (48 GB) GPU

to ensure sufficient processing power. We assess the statistical significance of all results of LLM performance using Welch's t-test [140], comparing each LLM to the best performer within its family. A significance level of $p < 0.05$ is applied, with each test based on a sample size of 6. Input linearization and prompt structure are important components of D2T, particularly in the context of LLM-based generation. The prompt design and input linearization templates used in this work are described in detail in subsection 6.5.6.

Listing 1 QLoRA and model quantization configuration

```
from peft import LoraConfig
# QLoRA configuration, implemented through PEFT
qlora_config = LoraConfig(task_type=TaskType.SEQ_2_SEQ_LM|TaskType.CAUSAL_LM, r=16,
↪  lora_alpha=16, lora_dropout=0.05)

from transformers import BitsAndBytesConfig
# Quantitation configuration
bnb_config = BitsAndBytesConfig(load_in_4bit=True, bnb_4bit_use_double_quant=True,
↪  bnb_4bit_quant_type="nf4", bnb_4bit_compute_dtype=torch.bfloat16)
```

3.4.4 Potential Risks and Considerations of Data Contamination

A potential concern in LLM evaluation is data contamination, where benchmark datasets may partially or fully overlap with the pretraining corpora of some models [141]. Since the D2T datasets used in the experiments are publicly available, it cannot be entirely ruled out that some samples may have been encountered during pretraining. However, scientifically determining such contamination remains practically impossible because the exact contents of LLM pretraining data are generally unavailable. Knowledge-cutoff information alone cannot reliably establish contamination, as it only indicates a temporal boundary and does not reveal the actual data used during training [142]. Nevertheless, the impact of potential contamination is expected to be limited in the present study. The analyses focus primarily on comparative trends, factual consistency (or faithfulness) behavior, and model characteristics across multiple datasets and experimental settings, rather than on isolated benchmark scores. In addition, D2T requires accurate grounding between structured inputs and generated text, where memorization of surface patterns alone does not necessarily imply faithful generation [35, 82]. Therefore, while possible data overlap cannot be completely excluded, it is unlikely to substantially affect the overall observations and conclusions.

3.5 Performance Evaluations

In this section, We evaluated LLM performance for D2T across three key metrics: readability, informativeness, and faithfulness.

3.5.1 Performance Evaluations on Readability

The readability of a D2T model, encompassing fluency and coherence [52], evaluates grammatical correctness and content organization. We assess readability using BLEU and METEOR.

TABLE 3.2: Readability evaluation of LLMs using BLEU. The best-performing models are highlighted in bold. Higher values indicate better readability. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3).

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.1	0.365	0.168	0.338	0.372	0.408
	BART-large	0.4	**0.381**	**0.181**	**0.367**	**0.444**	**0.520**
T5	T5-base	0.2	0.380	0.158	0.365	0.404	0.449
	T5-large	0.7	**0.398**	**0.193**	**0.390**	**0.442**	**0.535**
OPT	OPT-2.7B	2.7	0.350	0.262	0.421	0.441	0.521
	OPT-6.7B	6.7	**0.369**	**0.269**	**0.426**	0.448	0.538
	OPT-13B	13.0	0.347	0.269	0.412	**0.463**	**0.549**
BLOOM	BLOOM-1.1B	1.1	0.374	0.255	0.411	0.437	0.491
	BLOOM-3B	3.0	**0.380**	0.260	0.396	**0.446**	0.520
	BLOOM-7B	7.0	0.379	**0.274**	**0.423**	0.444	**0.530**
Llama 2	Llama2-7B	7.0	**0.419**	0.248	0.436	0.494	0.532
	Llama2-13B	13.0	0.408	**0.288**	**0.451**	**0.51**	**0.563**

- BLEU. BLEU [88] assesses readability by measuring n-gram overlap between generated and reference texts through calculating the geometric mean of n-gram precision (Precision_n) and applying a brevity penalty (β) to handle length differences. We can represent BLEU as the following equation:

$$\text{BLEU} = \beta \cdot \exp\left(\sum_n \log(w_n \cdot \text{Precision}_n)\right)$$

  where w_n denotes the weight associated with the corresponding precision. We use SacreBLEU (BLEU-4) [143] implementation to obtain our blue results. Despite criticisms regarding its effectiveness in measuring informativeness and faithfulness [105], BLEU remains a valuable metric for evaluating readability and fluency [35]. Therefore, we included BLEU in our assessment of readability.

- METEOR. METEOR [89], unlike BLEU, uses unigram matching and considers both precision and recall. It stands out as a weighted version of the F1-score, represented by the following equation:

$$\text{METEOR}_{\text{F1-measure}} = \frac{10 \cdot \text{Precison} \cdot \text{Recall}}{\text{Recall} + 9\text{Precison}}$$

Where, Precision and Recall denote precision and recall based on unigram-based matching. Unlike BLEU, METEOR incorporates synonym matching (using WordNet synsets) and word stemming (via Porter's stemmer) to better capture textual similarity. Thus, we believe that conjunction METEOR with BLEU for evaluating readability offers a systematic assessment.

TABLE 3.3: Readability evaluation of LLMs using METEOR. The best-performing models are highlighted in bold. Higher values indicate better readability. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3).

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.1	0.354	**0.315**	0.320	0.334	0.337
	BART-large	0.4	**0.358**	0.265	**0.337**	**0.375**	**0.397**
T5	T5-base	0.2	0.358	0.247	0.334	0.359	0.374
	T5-large	0.7	**0.368**	**0.276**	**0.352**	**0.379**	**0.402**
OPT	OPT-2.7B	2.7	0.355	0.307	0.366	0.382	0.403
	OPT-6.7B	6.7	**0.366**	0.305	**0.371**	0.387	0.406
	OPT-13B	13.0	0.355	**0.308**	0.361	**0.391**	**0.411**
BLOOM	BLOOM-1.1B	1.1	0.369	0.304	0.367	0.379	0.386
	BLOOM-3B	3.0	**0.371**	0.305	0.356	**0.386**	0.403
	BLOOM-7B	7.0	0.369	**0.314**	**0.375**	0.385	**0.412**
Llama 2	Llama2-7B	7.0	**0.376**	0.307	0.375	0.403	0.412
	Llama2-13B	13.0	0.37	**0.318**	**0.38**	**0.407**	**0.419**

Takeaways from Bleu (Table 3.2) and Meteor (Table 3.3) Results. Our empirical results on readability, as shown in Tables 3.2 and 3.3, reveal two key insights. First, the size of LLMs within all five families plays a significant role in enhancing readability across D2T tasks. Increasing model size consistently improves readability metrics, as measured by both BLEU and METEOR (with only a few discrepancies observed in LLM families like OPT by very small margin). These results suggest that, within the evaluated models and datasets, readability generally increases with model size in D2T tasks. In addition, among the evaluated models, Llama 2 achieves the highest readability scores on both metrics. However, smaller models like BART and T5 remain competitive, showing strong BLEU and METEOR results compared to larger models BLOOM and OPT, specially for E2E, WebNLG and DART datasets. The lower number of training instances in ViGGO and WikiTableText may explain their exclusion.

3.5.2 Performance Evaluations on Informativeness

Informativeness in D2T models refers to the content similarity or preservation between the generated and reference texts, indicating how well the model conveys useful information [35, 52]. We assess informativeness using BERTSCORE, MOVERSCORE and PARENT metrics.

- BERTSCORE. BERTSCORE [131] leverages BERT's contextual representations to assess text similarity, unlike n-gram-based models that struggle with capturing

contextual nuances. BERT generates contextual word representations from its hidden or output layers, and BERTSCORE uses these representations, applying cosine similarity to estimate text similarity. It calculates an F1-score based on the contextual alignment between the generated text (g) and the reference text (r). The main reason for choosing BERTSCORE to evaluate informativeness is its ability to assess the inherent (semantic) similarity between texts rather than just their syntactic similarity. We believe that emphasizing semantic similarity offers a more accurate representation of informativeness, setting it apart from aspects such as fluency or readability.

TABLE 3.4: Informativeness evaluation of LLMs using BERTSCORE. The best-performing models are highlighted in bold. Higher values indicate better informativeness. All results are statistically significant at $p < 0.05$ (see subsection 3.4.3).

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.1	0.915	0.839	0.876	0.903	0.913
	BART-large	0.4	**0.916**	**0.843**	**0.879**	**0.919**	**0.933**
T5	T5-base	0.2	0.916	0.828	0.882	0.913	0.921
	T5-large	0.7	0.916	**0.848**	**0.892**	**0.922**	**0.934**
OPT	OPT-2.7B	2.7	0.908	0.879	0.896	0.919	0.927
	OPT-6.7B	6.7	**0.910**	0.874	**0.898**	0.919	0.927
	OPT-13B	13.0	0.908	**0.879**	0.897	**0.920**	**0.928**
BLOOM	BLOOM-1.1B	1.1	0.911	0.884	0.895	0.915	0.917
	BLOOM-3B	3.0	**0.912**	0.883	0.887	0.916	0.925
	BLOOM-7B	7.0	0.911	**0.886**	**0.896**	**0.917**	**0.928**
Llama 2	Llama2-7B	7.0	0.918	0.882	0.897	0.931	0.934
	Llama2-13B	13.0	**0.918**	**0.889**	**0.900**	**0.931**	**0.938**

- MOVERSCORE. Similar to BERTSCORE, MOVERSCORE [132] emphasizes contextual similarity between texts. It measures semantic similarity using contextual embeddings derived from all words in the text together with a power-law normalization scheme. To compute the semantic distance between two texts, MOVERSCORE employs Word Mover's Distance (WMD) over the underlying n-gram representations of the texts, as shown in Equation 3.3:

$$\text{WMD}(g, r) = \min_{F \in \mathbb{R}^{|g| \times |r|}} \sum C \odot F \tag{3.3}$$

where C and F denote the cost matrix and flow matrix, respectively. In WMD, each entry of the cost matrix is computed using the contextual representations of the n-grams obtained after applying the power-law transformation. The flow matrix is determined by assigning weights to each n-gram based on inverse document frequency (IDF). Unlike BERTSCORE, MOVERSCORE additionally incorporates structural and syntactic information through the word-mover distance formulation. Therefore, using both MOVERSCORE and BERTSCORE provides a more comprehensive evaluation of informativeness.

TABLE 3.5: Informativeness evaluation of LLMs using MOVERSCORE. The best-performing models are highlighted in bold. Higher values indicate better informativeness. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3).

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.1	0.656	0.597	0.654	0.650	0.657
	BART-large	0.4	**0.666**	**0.601**	**0.667**	**0.676**	**0.688**
T5	T5-base	0.2	0.663	0.592	0.659	0.669	0.671
	T5-large	0.7	**0.665**	**0.608**	**0.670**	**0.684**	**0.687**
OPT	OPT-2.7B	2.7	0.655	0.645	0.682	0.673	0..676
	OPT-6.7B	6.7	**0.659**	0.639	**0.687**	0.673	0.678
	OPT-13B	13.0	0.654	**0.646**	0.684	**0.676**	**0.679**
BLOOM	BLOOM-1.1B	1.1	0.658	0.649	0.676	0.669	0.663
	BLOOM-3B	3.0	0.658	0.651	0.670	0.670	0.673
	BLOOM-7B	7.0	**0.658**	**0.655**	**0.678**	**0.671**	**0.679**
Llama 2	Llama2-7B	7.0	**0.671**	0.651	0.686	0.691	0.689
	Llama2-13B	13.0	0.668	**0.66**	**0.69**	**0.692**	**0.691**

- PARENT. Unlike BERTSCORE and MOVERSCORE, PARENT [39] evaluates informativeness by considering both the reference text and the source data. This enables PARENT to assess whether the generated text is not only similar to the reference but also grounded in the source information, making it particularly useful in the presence of source–reference divergence. Consequently, PARENT has been shown to correlate strongly with human judgments in D2T evaluation. The metric computes an F1-score based on entailed precision and entailed recall using lexical entailment techniques such as word overlap and co-occurrence. Let the lexical entailment of an n-gram g_i of the generated text g (i.e., $g_i \in \text{n-gram}(g)$) with respect to the source data s and the reference text r be denoted by $E_n(g_i, s)$ and $E_n(g_i, r)$, respectively. Then, the entailed precision of the generated text with respect to both the source data and the reference text is defined as:

$$\text{Precision}_{E_n}(g, s, r) \propto \frac{\sum_{g_i \in \text{n-gram}(g)} E_n(g_i, r) + (1 - E_n(g_i, r))\, E_n(g_i, s)}{|g_i \in \text{n-gram}(g)|}$$

  Similarly, the entailed recall of g with respect to both s and r is defined as:

$$\text{Recall}_{E_n}(g, s, r) = \text{Recall}_{E_n}(g, s)^{\lambda} \cdot \text{Recall}_{E_n}(g, r)^{1-\lambda}$$

  where λ denotes the weighting factor between the two recall components. Here, $\text{Recall}_{E_n}(g, r)$ is computed through lexical entailment with the reference text, while $\text{Recall}_{E_n}(g, s)$ is computed using longest common subsequence matching, since the source data s is often semi-structured in D2T tasks. Finally, the PARENT score is calculated as the F1-measure of Precision_{E_n} and Recall_{E_n}.

Takeaways from BertScore (Table 3.4), MoverScore (Table 3.5) and Parent (Table 3.6) Results. The results from both BERTSCORE (Table 3.4) and MOVERSCORE (Table 3.5) strongly indicate that increasing the model size of LLMs enhances

TABLE 3.6: Informativeness evaluation of LLMs using PARENT. The best-performing models are highlighted in bold. Higher values indicate better informativeness. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3).

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.1	**0.604**	0.353	0.440	0.551	0.590
	BART-large	0.4	0.592	**0.355**	**0.462**	**0.597**	**0.621**
T5	T5-base	0.2	0.590	0.344	0.472	0.577	0.592
	T5-large	0.7	**0.604**	**0.369**	**0.512**	**0.606**	**0.637**
OPT	OPT-2.7B	2.7	0.528	0.395	**0.510**	0.557	0.589
	OPT-6.7B	6.7	**0.549**	**0.401**	0.507	0.559	0.600
	OPT-13B	13.0	0.529	0.399	0.499	**0.565**	**0.612**
BLOOM	BLOOM-1.1B	1.1	0.565	**0.405**	0.502	**0.561**	0.566
	BLOOM-3B	3.0	**0.569**	0.400	0.476	0.556	0.585
	BLOOM-7B	7.0	0.565	0.404	**0.504**	0.557	**0.594**
Llama 2	Llama2-7B	7.0	**0.602**	0.390	0.531	0.621	0.621
	Llama2-13B	13.0	0.599	**0.418**	**0.531**	**0.622**	**0.640**

their informativeness, particularly when information alignment is calculated between reference and generated texts. Across all evaluated datasets, Llama 2 outperforms all other LLM families in terms of informativeness as measured by these two metrics. However, T5 and BART show competitive performance against BLOOM and OPT. Notably, T5 and BART achieve higher scores on the E2E, WebNLG, and DART datasets compared to BLOOM and OPT. Analysis of the PARENT scores (Table 3.6) suggests that, within the evaluated model families and parameter ranges, larger model sizes were generally associated with improvements in informativeness and readability, although some exceptions were observed for the BLOOM and OPT families. While Llama 2 consistently performs well across all LLM families in PARENT scores, T5 outperforms Llama 2 on the E2E dataset, and DART also shows better results. Additionally, T5 and BART surpass OPT and BLOOM on the E2E, WebNLG, and DART datasets. Overall, Table 3.6 illustrates that larger LLM families like OPT, BLOOM, and Llama 2 may struggle to capture informativeness effectively when considering information alignment between source and reference data. In contrast, smaller LLM families can achieve impressive results on certain datasets.

3.5.3 Performance Evaluations on Faithfulness

Faithfulness is crucial in D2T models [34, 44], as it evaluates the factual accuracy of generated text against the source data, focusing on the tendency to generate incorrect and irrelevant facts. An example of unfaithful generation from the E2E dataset is shown in Figure 3.7. We measure faithfulness using the BARTSCORE metric.

- BARTSCORE. BARTSCORE[133] evaluates the faithfulness between two texts by estimating the likelihood of text generation based on a sequence-to-sequence model like BART. BART's strong capability to capture faithfulness makes it a reliable

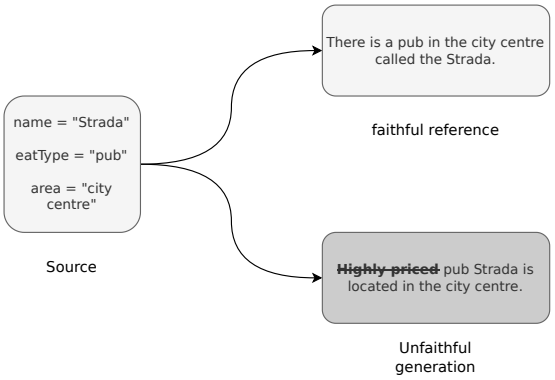


FIGURE 3.7: An illustration of unfaithful generation on the E2E dataset. The generated text contains the phrase "**Highly price**", which is irrelevant with respect to the information provided in the source data.

metric for identifying discrepancies between generated text and source content. For this reason, we incorporated BARTSCORE in our experiments to provide a robust assessment of faithfulness in the generated outputs. The metric is based on the principle that if two pieces of text are factually equivalent, the likelihood of generating one from the other will be higher. Faithfulness in BARTSCORE is determined by the probability (equation3.4) of transforming from the source text (s) to the generated text (g).

$$\text{Faithfulness}(s, g) = \Pr(\text{BART}(s \rightarrow g)) \tag{3.4}$$

BARTSCORE has been reported to correlate more strongly with human judgments than several earlier evaluation metrics.

TABLE 3.7: Faithfulness evaluation of LLMs using BARTSCORE. The best-performing models are highlighted in bold. Higher values indicate better faithfulness. All results are statistically significant at $p < 0.05$ (mentioned in subsection 3.4.3).

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.1	0.086	**0.065**	**0.068**	0.066	**0.119**
	BART-large	0.4	**0.087**	0.064	**0.071**	**0.069**	0.110
T5	T5-base	0.2	0.087	**0.074**	0.070	**0.072**	**0.104**
	T5-large	0.7	**0.093**	0.063	**0.077**	0.070	0.101
OPT	OPT-2.7B	2.7	**0.057**	0.045	**0.068**	**0.063**	**0.089**
	OPT-6.7B	6.7	0.053	**0.046**	0.066	0.062	0.087
	OPT-13B	13.0	0.050	0.043	0.065	0.063	0.087
BLOOM	BLOOM-1.1B	1.1	**0.089**	**0.042**	**0.057**	**0.064**	**0.090**
	BLOOM-3B	3.0	0.083	0.041	0.054	0.064	0.089
	BLOOM-7B	7.0	0.086	0.040	0.056	0.063	0.088
Llama 2	Llama2-7B	7.0	**0.094**	**0.05**	**0.062**	0.073	**0.106**
	Llama2-13B	13.0	0.089	0.048	0.061	**0.074**	0.105

Takeaways from BartScore (Table 3.7) Results. Our analysis of the empirical results from Table 3.7 reveals that increasing model size does not necessarily enhance the faithfulness of LLMs within a family for D2T tasks. In many cases, increasing the

number of parameters may even degrade performance on faithfulness, particularly for OPT, BLOOM, and Llama 2. T5 also experiences a loss of faithfulness as model size increases on the ViGGo and DART datasets. Even when increasing model size leads to slight improvements in faithfulness as measured by BARTSCORE, these enhancements are minimal. Therefore, the results indicate that larger LLMs may still exhibit reduced faithfulness performance despite improvements in other generation qualities. Moreover, no LLM family emerges as a clear winner for faithfulness based on BARTSCORE. However, Llama 2 suggests better faithfulness on the E2E and DART datasets, while T5 excels in ViGGo and WikiTableText, and BART performs well in WebNLG.

3.5.4 Performance Evaluation through Human Assessment

Human evaluation remains a crucial component in assessing the performance of D2T models, particularly LLMs [105, 106]. To complement the automatic evaluation, we conducted a human study to analyze the perceived quality of LLM-generated texts across multiple datasets. For the evaluation process, four graduate students from diverse academic backgrounds participated to reduce potential bias. Participation was voluntary, and all necessary instructions and resources were provided. Prior to the main evaluation, each annotator completed a short training phase consisting of five example input–output pairs with gold-standard annotations supplied by the authors. This familiarized them with the evaluation criteria and task objectives. Annotators were instructed to independently assess each generated text along three key qualities: *readability*, *informativeness*, and *faithfulness*. Ratings were recorded on a five-point Likert scale ranging from 1 (lowest) to 5 (highest). During the main evaluation phase, each annotator rated 15 generated outputs for every combination of LLM and dataset. The detailed instructions provided to the annotators for conducting the evaluation are presented below.

Instructions for Human Evaluation

Participation in this annotation task is entirely voluntary. By proceeding with the evaluation, you acknowledge that you have read and understood the instructions and agree to participate in the study.

Please carefully read the provided input data (source), reference text, and generated output before assigning a score. Your task is to evaluate the overall quality of the generated text by jointly considering the following three aspects:

- **Readability.** Determine whether the generated text is fluent, grammatically correct, coherent, and naturally written.

- **Informativeness.** Determine whether the generated text includes important and relevant information from the provided input data and avoids significant omissions.
- **Faithfulness.** Determine whether the generated text is supported by the provided input data and reference text (while being mindful of possible source–reference divergence) without introducing unsupported, contradictory, or hallucinated information.

Do not assign separate scores for individual criteria. Instead, consider all three aspects together and provide a single overall quality score using the following five-point scale:

- **Score 1 (Very Poor):** The generated text has severe issues in multiple aspects. It may contain major factual consistency errors, significant omissions, poor fluency, or incoherent language.
- **Score 2 (Poor):** The generated text has noticeable problems in one or more aspects, such as factual inaccuracies, missing important information, or readability issues.
- **Score 3 (Moderate):** The generated text is acceptable overall but contains several limitations or deficiencies.
- **Score 4 (Good):** The generated text is generally satisfactory with only minor issues affecting quality.
- **Score 5 (Excellent):** The generated text is fluent, informative, and factually consistent with the provided information.

Please base your judgment on the overall balance among these characteristics rather than focusing excessively on any single aspect. Evaluate each sample independently and avoid allowing previous evaluations to influence subsequent judgments. In addition to the numerical rating, please provide a short explanatory comment describing the main reasons behind your assigned judgment, including notable factual inconsistencies, readability issues, lack of informativeness, or other important observations that influenced the score.

Table 3.8 summarizes the human evaluation results across all LLMs and datasets. Overall, the trends observed in human assessments are broadly consistent with those found in automatic metrics, particularly regarding *readability* and *informativeness*. However, unlike the automatic metrics, the human evaluations revealed a slightly different trend

TABLE 3.8: Human evaluation of LLMs on D2T tasks. The best-performing models are highlighted in bold. The inter-annotator agreement, measured using Krippendorff's alpha (α), was 0.81, indicating moderate to high consistency among evaluators.

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.139	0.26	0.32	0.36	**0.28**	0.48
	BART-large	0.406	**0.42**	**0.45**	**0.40**	**0.44**	**0.58**
T5	T5-base	0.223	0.44	0.38	0.52	0.48	0.55
	T5-large	0.738	**0.45**	**0.48**	**0.54**	**0.52**	**0.64**
OPT	OPT-2.7B	2.7	0.32	0.28	0.45	0.40	0.48
	OPT-6.7B	6.7	0.32	0.32	**0.48**	0.42	0.52
	OPT-13B	13	**0.44**	**0.42**	0.45	**0.48**	**0.52**
BLOOM	BLOOM-1.1B	1.1	0.25	0.22	0.35	0.40	0.50
	BLOOM-3B	3	0.32	0.30	0.36	0.38	0.55
	BLOOM-7B	7	**0.42**	**0.36**	**0.52**	**0.56**	**0.55**
Llama 2	Llama2-7B	7	**0.45**	0.35	0.45	0.55	0.48
	Llama2-13B	13	0.42	0.35	0.45	**0.56**	**0.60**

in *faithfulness*, where larger LLMs were often rated higher across all three evaluation qualities. These findings suggest that, according to human judgment, overall generation quality in D2T tasks is influenced by a balanced combination of readability, informativeness, and faithfulness—rather than faithfulness alone.

3.5.5 Discussion

From the empirical results (Table 3.2, 3.3, 3.4, 3.5, 3.6, and 3.7) obtained from six automatic metrics across three key qualities—*readability*, *informativeness*, and *faithfulness*—we observe that increasing the number of parameters within an LLM family generally enhances readability and informativeness, while it may adversely affect faithfulness in D2T tasks. Interestingly, our human evaluation (Table 3.8) reveals a more balanced perspective: larger LLMs tend to receive higher ratings across all three qualities, suggesting that their improved linguistic fluency and contextual understanding may partially offset faithfulness detected by automatic metrics. This divergence highlights that automatic metrics may over-penalize minor errors on faithfulness while underestimating perceived quality from a human standpoint. Therefore, in applications where faithfulness is of utmost importance—for example, in safety-critical settings such as medical report generation [49]—a smaller, more controllable LLM within the same family may be preferable. Conversely, in scenarios that prioritize readability, informativeness, and faithfulness from an overall human perspective—such as creative or user-facing D2T applications—using larger LLMs can be beneficial due to their enhanced linguistic expressiveness and coherence.

3.6 Robustness of Performance Evaluation with Respect to Fine-tuning and Decoding Strategies

In this section, we examine the robustness of our performance evaluation with respect to different fine-tuning and decoding strategies. We aim to verify whether the observed trends in LLM performance on D2T remain consistent across alternative model fine-tuning and decoding methods. For this robustness analysis, we consider three representative LLM families—T5, BLOOM, and LLAMA 2—encompassing six individual models: T5-base, T5-large, BLOOM-3B, BLOOM-7B, Llama2-7B, and Llama2-13B. To ensure coverage of different D2T paradigms, we include one dataset for each type: E2E for MR-to-Text, WIKITABLETEXT for Table-to-Text, and DART for Graph-to-Text. Model performance is evaluated along three key qualities of D2T: *readability* (measured by BLEU), *informativeness* (measured by PARENT), and *faithfulness* (measured by BARTSCORE).

Listing 2 Prefix-Tuning configuration

```
from peft import PrefixTuningConfig
# Prefix-Tuning configuration, implemented through PEFT
pfx_config = PrefixTuningConfig(task_type=TaskType.SEQ_2_SEQ_LM|TaskType.CAUSAL_LM,
↪  num_virtual_tokens=32)
```

Fine-tuning Strategies. To assess robustness under different fine-tuning approaches, we compare our primary QLoRA fine-tuning strategy with Prefix-Tuning [86]. Unlike QLoRA, which introduces low-rank adapters into specific weight matrices of the LLM, Prefix-Tuning prepends a sequence of trainable virtual tokens to the model's input embeddings (code listing 2). These trainable vectors act as continuous prompts that guide the model toward task-specific behavior without modifying the original model parameters. Let assume the input sequence be represented as $X = [x_1, x_2, \ldots, x_n]$, where x_i denotes the embedding of the i-th token. Prefix-Tuning introduces a trainable prefix sequence $P = [p_1, p_2, \ldots, p_m]$, where m represents the prefix length and $p_i \in \mathbb{R}^d$ denotes the embedding of a virtual token. The modified input can therefore be expressed as $X' = [P; X]$, where $[;]$ denotes sequence concatenation. Within the transformer architecture, these prefix representations are incorporated into the self-attention mechanism by augmenting the key and value representations as $K' = [K_P; K]$ and $V' = [V_P; V]$, where K_P and V_P are learnable prefix-associated key and value vectors. Consequently, the attention operation becomes $\text{Attention}(Q, K', V') = \text{softmax}\left(\frac{QK'^T}{\sqrt{d_k}}\right)V'$, allowing the model to attend to both the original input and task-specific prefix information.

Through this mechanism, Prefix-Tuning can effectively adapt model behavior while updating only a small number of additional parameters, making it a computationally efficient fine-tuning approach.

As shown in Table 3.9, Prefix-Tuning achieves performance comparable to QLoRA across most metrics, with only marginal degradation observed in a few cases. We attribute these minor differences to the fixed number of tunable parameters in Prefix-Tuning, which may not optimally scale with model size across different LLM families. Nevertheless, the overall consistency between both fine-tuning methods suggests that our evaluation findings remain stable and robust across fine-tuning paradigms.

TABLE 3.9: Performance comparison between QLoRA and Prefix-Tuning fine-tuning strategies across all LLMs and datasets, showing consistent results and robustness in evaluation.

metric	family	model	size (number of parameters in billion)	QLoRA			Prefix-Tuning		
				E2E	WikiTableText	DART	E2E	WikiTableText	DART
BLEU	T5	T5-base	0.2	0.380	0.365	0.404	0.377	0.351	0.411
		T5-large	0.7	0.398	0.390	0.442	0.393	0.380	0.432
	BLOOM	BLOOM-3B	3.0	0.380	0.396	0.446	0.389	0.375	0.452
		BLOOM-7B	7.0	0.379	0.423	0.444	0.377	0.420	0.447
	Llama 2	Llama2-7B	7.0	0.419	0.436	0.494	0.435	0.446	0.474
		Llama2-13B	13.0	0.408	0.451	0.510	0.423	0.452	0.528
PARENT	T5	T5-base	0.2	0.590	0.472	0.577	0.591	0.474	0.597
		T5-large	0.7	0.604	0.512	0.606	0.624	0.492	0.573
	BLOOM	BLOOM-3B	3.0	0.569	0.476	0.556	0.570	0.477	0.489
		BLOOM-7B	7.0	0.565	0.504	0.557	0.563	0.449	0.534
	Llama 2	Llama2-7B	7.0	0.602	0.531	0.621	0.578	0.439	0.641
		Llama2-13B	13.0	0.599	0.531	0.622	0.559	0.532	0.533
BARTSCORE	T5	T5-base	0.2	0.087	0.070	0.072	0.088	0.069	0.077
		T5-large	0.7	0.093	0.077	0.070	0.095	0.069	0.078
	BLOOM	BLOOM-3B	3.0	0.083	0.054	0.064	0.085	0.072	0.088
		BLOOM-7B	7.0	0.086	0.056	0.063	0.087	0.067	0.067
	Llama 2	Llama2-7B	7.0	0.094	0.062	0.073	0.093	0.081	0.078
		Llama2-13B	13.0	0.089	0.061	0.074	0.079	0.068	0.078

Decoding Strategies. To examine robustness under different decoding strategies, we compare the deterministic Beam Search decoding (used in our main experiments) with the stochastic Nucleus Sampling [79] technique, also known as Top-p sampling. For Nucleus Sampling, we set $p = 0.95$, meaning that tokens are sampled from the smallest possible set whose cumulative probability exceeds 0.95. As shown in Table 3.10, the results from Nucleus Sampling are largely consistent with those obtained using Beam Search, indicating that similar trends are observed across the evaluated decoding approaches. A slight drop is observed in the *faithfulness* (measured via BARTSCORE), which can be attributed to the inherent stochasticity of sampling-based decoding. Apart from this minor variation, both decoding methods produce highly similar results, providing evidence that the observed evaluation trends are largely consistent across the examined decoding strategies.

TABLE 3.10: Performance comparison between Beam Search and Nucleus Sampling decoding strategies across all LLMs and datasets, showing consistent results and robustness in evaluation.

metric	family	model	size (number of parameters in billion)	beam search			nucleus sampling		
				E2E	WikiTableText	DART	E2E	WikiTableText	DART
BLEU	T5	T5-base	0.2	0.380	0.365	0.404	0.394	0.396	0.416
		T5-large	0.7	0.398	0.390	0.442	0.423	0.426	0.439
	BLOOM	BLOOM-3B	3.0	0.380	0.396	0.446	0.428	0.463	0.471
		BLOOM-7B	7.0	0.379	0.423	0.444	0.419	0.492	0.479
	Llama 2	Llama2-7B	7.0	0.419	0.436	0.494	0.471	0.454	0.463
		Llama2-13B	13.0	0.408	0.451	0.51	0.485	0.495	0.548
PARENT	T5	T5-base	0.2	0.590	0.472	0.577	0.611	0.477	0.587
		T5-large	0.7	0.604	0.512	0.606	0.504	0.521	0.584
	BLOOM	BLOOM-3B	3.0	0.569	0.476	0.556	0.571	0.429	0.522
		BLOOM-7B	7.0	0.565	0.504	0.557	0.554	0.537	0.556
	Llama 2	Llama2-7B	7.0	0.602	0.531	0.621	0.613	0.535	0.628
		Llama2-13B	13.0	0.599	0.531	0.622	0.605	0.524	0.634
BARTSCORE	T5	T5-base	0.2	0.087	0.070	0.072	0.086	0.063	0.079
		T5-large	0.7	0.093	0.077	0.070	0.068	0.064	0.064
	BLOOM	BLOOM-3B	3.0	0.083	0.054	0.064	0.086	0.050	0.063
		BLOOM-7B	7.0	0.086	0.056	0.063	0.074	0.055	0.062
	Llama 2	Llama2-7B	7.0	0.094	0.062	0.073	0.098	0.057	0.072
		Llama2-13B	13.0	0.089	0.061	0.074	0.090	0.055	0.072

3.7 Performance Evaluation Amid Source-reference Divergence

As discussed in Subsection 3.3.3, source-reference divergence is a common and unavoidable issue in D2T [39, 68]. So we here we evaluated how LLMs perform in D2T tasks when source-reference divergence is present. Consistent with the robustness analysis setup, we employ the same LLMs, D2T datasets, and evaluation metrics to examine the characteristics of source–reference divergence. To provide a clearer illustration of the impact of source–reference divergence, we categorize the test partition of each dataset into three equally sized groups based on the average source–reference divergence scores: `low` (instances with relatively low source–reference divergence), `medium`, and `high` (instances with relatively high source–reference divergence). Each group contains approximately one-third of the total test instances. We assess source-reference divergence ($\text{div}(s, r)$) based on the differences between the source text (s) and the reference text (r) at the unigram level, as follows:

$$\text{div}(s, r) = 1 - \frac{\left|\text{unigram}(s) \cap \text{unigram}(r)\right|}{\left|\text{unigram}(s) \cup \text{unigram}(r)\right|}$$

Where $|x|$ denotes the cardinality and 'unigram(x)' represents the set of all unigrams of the text x. We opt for unigram-based matching to calculate divergence rather than higher-order n-gram matching because the source data S often consists of semi-structured non-textual data, where higher-order grams may overlook significant portions of information [48, 67].

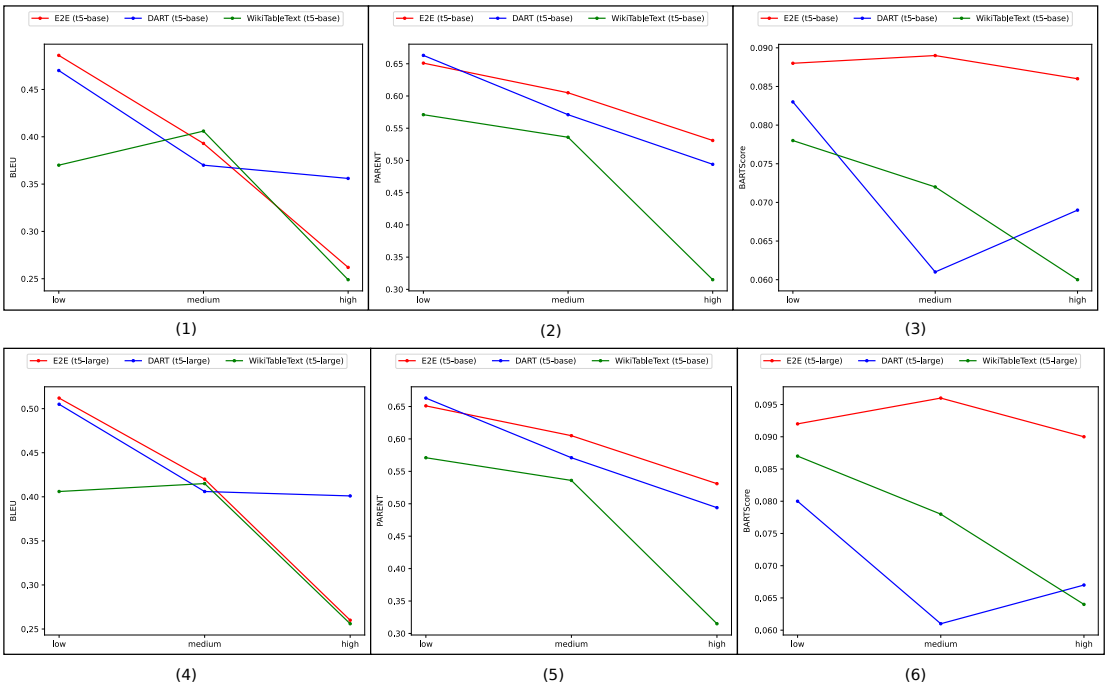


FIGURE 3.8: Analysis of performance of T5 in context of source-reference divergence based on three datasets (E2E, WikiTableText and DART) in terms of BLEU, and PARENT, BARTSCORE. First row represents performance of T5-base, and second row represents performance of T5-large.

3.7.1 Results Amid Source-reference Divergence.

Figures 3.8, 3.9, and 3.10 illustrate the performance of T5, BLOOM, and Llama 2 across three D2T datasets, specifically focusing on source-reference divergence. The results highlight a significant influence of source-reference divergence on the performance of LLMs across all three families. In the `low` divergence group, characterized by minimal discrepancies between source and reference texts, all three LLM families consistently achieve higher scores across the evaluated metrics (BLEU, PARENT, and BARTSCORE). This indicates that when source-reference divergence is low, LLMs tend to excel in terms of readability, informativeness, and faithfulness. Conversely, in the `high` source-reference divergence group, where divergences are more pronounced, all LLMs exhibit notably lower performance across all three qualities compared to the `low` and `medium` groups. This underscores the substantial impact of source-reference divergence on LLM performance, with higher levels of source-reference divergence leading to decreased scores in readability, informativeness, and faithfulness. Another key finding from Figures 3.8, 3.9, and 3.10 is that changes in model size within each LLM family do not significantly affect performance across the three qualities when considering source-reference divergence. Thus, we conclude that the model size of LLMs within a family is largely agnostic to their performance concerning source-reference divergence.

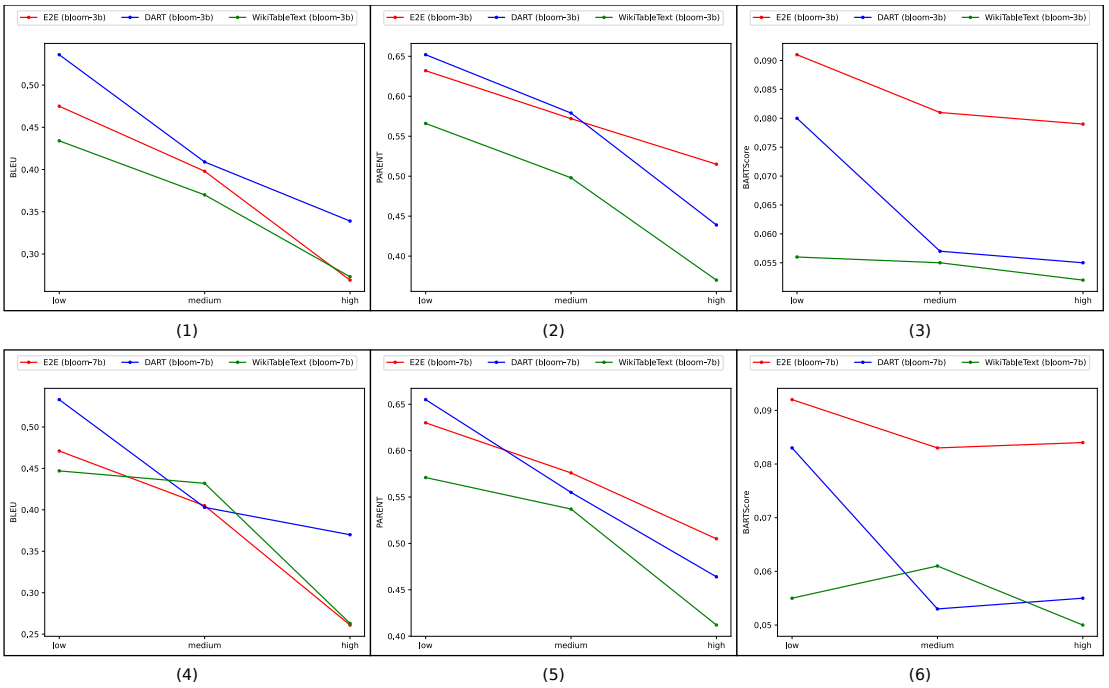


FIGURE 3.9: Analysis of performance of BLOOM in context of source-reference divergence based on three datasets (E2E, WikiTableText and DART) in terms of BLEU, PARENT and BARTSCORE. First row represents performance of BLOOM-3B, and second row represents performance of BLOOM-7B.

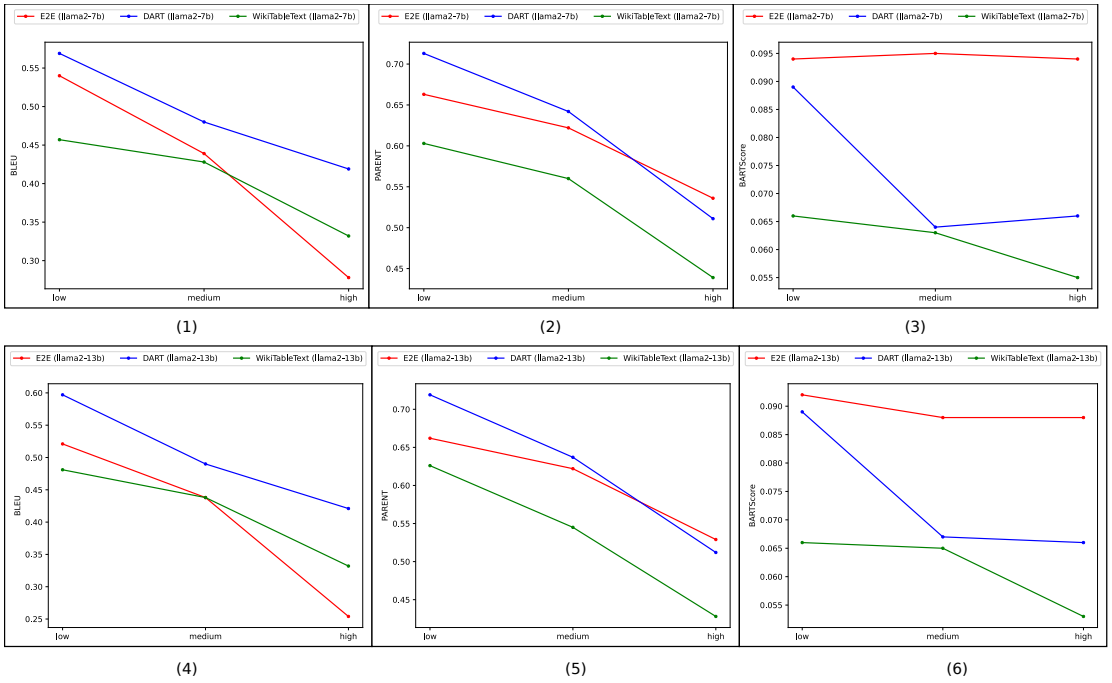


FIGURE 3.10: Analysis of performance of Llama 2 in context of source-reference divergence based on three datasets (E2E, WikiTableText and DART) in terms of BLEU, PARENT and BARTSCORE. First row represents performance of Llama2-7B, and second row represents performance of Llama2-13B.

3.8 Case Studies

We present two case studies (Table 3.11 and 3.12) conducted on the E2E and WikiTableText datasets, each showcasing two samples and their corresponding generations from four LLMs: T5-base, T5-large, Llama2-7B, and Llama2-13B. In the first case (Table 3.11) study using the E2E dataset, one example illustrates how smaller LLMs like T5-base may fail to accurately convey information. Another example highlights a source-reference divergence phenomenon, where larger LLMs like Llama2-13B fail to capture the divergence, while other LLMs successfully handle it. Table 3.11 illustrates examples in which smaller LLMs exhibit lower informativeness, while larger LLMs sometimes struggle under conditions of source-reference divergence. The second case study (Table 3.12) on the WikiTableText dataset suggests how increasing LLM size enhances informativeness as a D2T model. Additionally, it reveals that smaller LLMs such as T5-base and T5-large tend to prioritize faithfulness, while larger LLMs prioritize readability. Table 3.12 reveals that smaller-sized LLMs prioritize faithfulness, rendering them more suitable for applications requiring safety-critical considerations.

TABLE 3.11: Case study on the E2E dataset with four LLMs (T5-base, T5-large, Llama2-7B, and Llama2-13B), highlighting errors by smaller models and larger models' failure to detect source-reference divergence.

source	reference	LLM	generated text	remarks
name=The Vaults; eatType = pub; food=French; priceRange=less than £20; area=city centre; familyFriendly=yes	Near Raja Cuisine, in the city centre is a pub offering French food. The Vaults is cheap and family-friendly.	T5-base	The Vaults is a family-friendly French pub in the city centre with a price range of less than £20.	All LLMs are successfully generated similar texts, however T5-base transformed with 'food=French' into 'French pub', mistakenly.
		T5-large	The Vaults is a family-friendly pub in the city centre that serves French food for less than £20.	
		Llama2-7B	The Vaults is a family-friendly pub that serves French food. It is located in the city centre and has a price range of less than £20.	
		Llama2-13B	The Vaults is a family-friendly pub that serves French food for less than £20. It is located in the city centre.	
name=The Waterman; eatType=restaurant; priceRange=high; area=riverside; familyFriendly=yes; near=Raja Indian Cuisine	The Waterman is a restaurant near to Raja Indian Cuisine that is child-friendly. It has a high price range but is located in riverside.	T5-base	The Waterman is a high priced, child friendly restaurant in the riverside area near Raja Indian Cuisine.	In this example, there is a discrepancy between the source data and the reference. The source data mentions familyFriendly=yes, while the reference uses the term 'child-friendly'. Interestingly, all three LLMs, except Llama2-13B, successfully identify this discrepancy. This case highlights the possibility that larger LLMs may overlook certain divergences in the data.
		T5-large	The Waterman is a child friendly restaurant in the riverside area near Raja Indian Cuisine. It has a high price range.	
		Llama2-7B	The Waterman is a high priced restaurant in the riverside area near Raja Indian Cuisine. It is child friendly.	
		Llama2-13B	The Waterman is a high priced family friendly restaurant in the riverside area near Raja Indian Cuisine.	

TABLE 3.12: Case study on the WikiTableText dataset with four LLMs (T5-base, T5-large, Llama2-7B, and Llama2-13B), showing how larger models improve informativeness, while smaller models generate faithful but less readable content.

source	reference	LLM	generated text	remarks
[['subjtitle', 'history of the jews in malaysia'], ['subjsubtitle', 'history'], ['year', '1881'], ['males', '14'], ['females', '9'], ['boys', '5'], ['girls', '4'], ['total', '32']]	in 1988 , there were 32 the jews in malaysia in which there were 14 males , 9 females , 5 boys and 4 girls .	T5-base	in 1881, there were 32 jews in malaysia.	All LLMs are experiencing issues with *informativeness* as they struggle to generate all the source content accurately. In this example, larger LLMs show improved informativeness, with Llama2-13B achieving comparatively better results.
		T5-large	in 1881, there were 14 males and 9 females in malaysia.	
		Llama2-7B	there were 14 males and 9 females in the jews in malaysia in 1881 .	
		Llama2-13B	there were 14 males , 9 females , 5 boys and 4 girls in the jews in malaysia in 1881 .	
[['subjtitle', 'bath rugby'], ['subjsubtitle', 'current squad'], ['player', 'jack wilson'], ['position', 'wing'], ['union', 'new zealand']]	in bath rugby , wing jack wilson was a member of new zealand union .	T5-base	jack wilson was a wing in new zealand.	While T5-base and T5-large provide faithful information, Llama2-7B and Llama2-13B outperform them in terms of readability.
		T5-large	jack wilson was from union new zealand.	
		Llama2-7B	jack wilson was from new zealand .	
		Llama2-13B	jack wilson was from new zealand .	

3.9 Source-Reference Divergence-Weighted Training

From the previous discussion (section 3.7), we have observed that source-reference divergence plays an important role in determining the performance of LLMs on D2T tasks. To explicitly incorporate this insight into the training process, we propose a *Source-Reference Divergence-Weighted Training* strategy (Equation 3.5), which extends the conventional Maximum Likelihood Estimation (MLE) loss ($\mathcal{L}$) formulation (Equation 3.1) by introducing a source–reference divergence-weighted loss function ($\mathcal{L}_{\text{weighted}}$) for each training instance.

$$\theta^* \leftarrow \underset{\theta}{\operatorname{argmax}} \sum_{(s,r)\in\mathcal{D}} \underbrace{w(s,r)\cdot\mathcal{L}\big(\mathcal{M}_\theta(s),r\big)}_{\mathcal{L}_{\text{weighted}}\left(\mathcal{M}_\theta(s),r\right)} \tag{3.5}$$

Here, s and r denote the source and reference text, respectively, and $w(s,r)$ represents the weight assigned to each instance. We define the weight as the inverse of the source–reference divergence:

$$w(s,r) = 1 - \operatorname{div}(s,r) \tag{3.6}$$

where $\operatorname{div}(s,r) \in [0,1]$ quantifies the semantic divergence between the source and reference.

Intuitively, this weighting mechanism functions as a form of *regularization*, encouraging the model to place greater emphasis on instances with low divergence (i.e., semantically well-aligned source-reference pairs), while reducing the influence of highly divergent examples that may introduce unfaithfulness during training. Thus, the modified loss

can be interpreted as:

$$\mathcal{L}_{\text{weighted}}\big(\mathcal{M}_\theta(s), r\big) = \sum_{(s,r)\in\mathcal{D}} \big[\, \mathcal{L}(\mathcal{M}_\theta(s), r) - \lambda \cdot \text{div}(s, r) \cdot \mathcal{L}(\mathcal{M}_\theta(s), r)\big], \tag{3.7}$$

where λ controls the regularization strength implicitly determined by $w(s, r)$.

TABLE 3.13: Performance comparison between conventional and divergence-weighted training across LLMs and D2T datasets. Divergence-weighted training yields consistent improvements across metrics. The best results are shown in **bold**.

metric	family	model	size (number of parameters in billion)	conventional training			Weighted training		
				E2E	WikiTableText	DART	E2E	WikiTableText	DART
BLEU	T5	T5-base	0.2	0.380	0.365	0.404	**0.387**	**0.367**	**0.446**
		T5-large	0.7	0.398	**0.390**	0.442	**0.418**	0.388	**0.462**
	BLOOM	BLOOM-3B	3.0	0.380	0.396	0.446	**0.388**	**0.427**	0.442
		BLOOM-7B	7.0	0.379	0.423	0.444	**0.411**	**0.438**	**0.469**
	Llama 2	Llama2-7B	7.0	0.419	0.436	**0.494**	**0.451**	**0.449**	0.439
		Llama2-13B	13.0	0.408	0.451	0.510	**0.448**	**0.516**	**0.528**
PARENT	T5	T5-base	0.2	0.590	**0.472**	0.577	**0.602**	0.470	**0.582**
		T5-large	0.7	0.604	0.512	0.606	**0.620**	**0.545**	0.573
	BLOOM	BLOOM-3B	3.0	0.569	**0.476**	0.556	**0.572**	0.463	**0.582**
		BLOOM-7B	7.0	0.565	0.504	0.557	**0.573**	**0.556**	**0.591**
	Llama 2	Llama2-7B	7.0	0.602	0.531	**0.621**	**0.616**	**0.557**	0.608
		Llama2-13B	13.0	0.599	0.531	0.622	**0.612**	**0.558**	**0.632**
BARTSCORE	T5	T5-base	0.2	0.087	0.070	0.072	**0.088**	**0.075**	**0.081**
		T5-large	0.7	0.093	**0.077**	0.070	**0.104**	0.072	**0.086**
	BLOOM	BLOOM-3B	3.0	0.083	0.054	0.064	**0.085**	**0.061**	**0.076**
		BLOOM-7B	7.0	0.086	0.056	0.063	**0.097**	**0.094**	0.063
	Llama 2	Llama2-7B	7.0	0.094	0.062	0.073	**0.096**	**0.075**	**0.078**
		Llama2-13B	13.0	**0.089**	0.061	0.074	0.088	**0.086**	**0.082**

We adopt the same experimental setup as in the robustness analysis, employing the same set of LLMs, D2T datasets, and evaluation metrics. Table 3.13 presents the comparison between standard MLE training and divergence-weighted training. Across all three evaluation metrics—BLEU, PARENT, and BARTSCORE—the proposed weighted training consistently outperforms the conventional approach across multiple LLM families and datasets. These results empirically substantiate the effectiveness of our divergence-weighted method in enhancing D2T quality.

3.10 Conclusion

The chapter presents a broad evaluation of LLMs for D2T, encompassing twelve models from five popular LLM families, evaluated across multiple datasets and metrics. The study integrates automatic and human evaluations, robustness analyses, and introduces a *source-reference divergence-weighted training* strategy to provide a broad understanding of LLM behavior in D2T tasks. The findings suggest that increasing model size generally enhances readability and informativeness, with Llama 2 exhibiting the strongest overall performance in these aspects. However, larger models do not consistently improve faithfulness and may even lead to its degradation. Human evaluations reveal a more nuanced trend—larger models often receive higher ratings across all three

qualities, suggesting that their superior readability and contextual expressiveness can compensate for minor unfaithfulness that is frequently over-penalized by automatic metrics. Robustness analyses further indicate that these trends remain stable across different fine-tuning strategies (QLoRA vs. Prefix-Tuning) and decoding methods (Beam Search vs. Nucleus Sampling), thereby supporting the consistency of the evaluation framework. Moreover, experiments examining source-reference divergence suggest that model performance degrades notably as divergence increases, with model size having a limited effect in mitigating this challenge. To address this limitation, the proposed *source-reference divergence-weighted training* strategy adaptively adjusts instance-level loss based on the degree of divergence between source and reference data. This approach improves performance across readability, informativeness, and faithfulness under the evaluated settings, suggesting that it may serve as a useful regularization mechanism.

Chapter 4

Evaluating Factual Consistency in LLM-Based D2T

Building upon the findings of the previous chapter, which examined the overall performance of LLMs for data-to-text generation (D2T) and proposed the *source-reference divergence-weighted training* strategy, this chapter[1] presents a broad empirical evaluation of *faithfulness* in terms of *factual consistency*. It describes an empirical investigation of factual consistency across five representative LLM families (T5, BART, OPT, BLOOM, and Llama 2) and five widely used D2T datasets (E2E, ViGGo, WikiTableText, DART, and WebNLG). In this work, the primary goal is to evaluate faithfulness by quantifying the factual alignment between generated text and its source data, thereby examining associations between factual consistency and factors such as model scale, dataset diversity, and source-reference divergence. Overall, this chapter reports the findings of a multi-dataset empirical study that provides additional empirical insights into factual consistency in LLM-based D2T and informs future strategies for enhancing model faithfulness.

4.1 Overview

The rapid advancement of LLMs has influenced developments in D2T [6, 144]. LLMs have often achieved higher scores than earlier D2T architectures on several evaluated benchmarks on several evaluation dimensions, particularly in terms of *readability*—capturing fluency and coherence—and *informativeness*—reflecting the extent to

[1]The work presented in this chapter is based on our paper: Mahapatra, J. and Garain, U., 2025. A Comprehensive Study on Factual Consistency in LLMs for Data-to-Text Generation. *Sādhanā*, Springer. [117]

which generated text preserves source information [35]. Despite these notable gains, generating *factually consistent* text in D2T remains a persistent challenge for LLMs.

Factual consistency in D2T refers to the degree to which the generated text accurately reflects the factual content of the input data [54]. Deficiencies in factual consistency may reduce users' confidence in D2T outputs, particularly in applications where factual accuracy is important, such as medical report generation [50] and financial reporting [138]. Accordingly, evaluating factual consistency is important for understanding the faithfulness of LLMs in various applications. However, despite the increasing adoption of LLMs, relatively few studies have examined factual consistency across multiple datasets and model families in a unified evaluation setting.

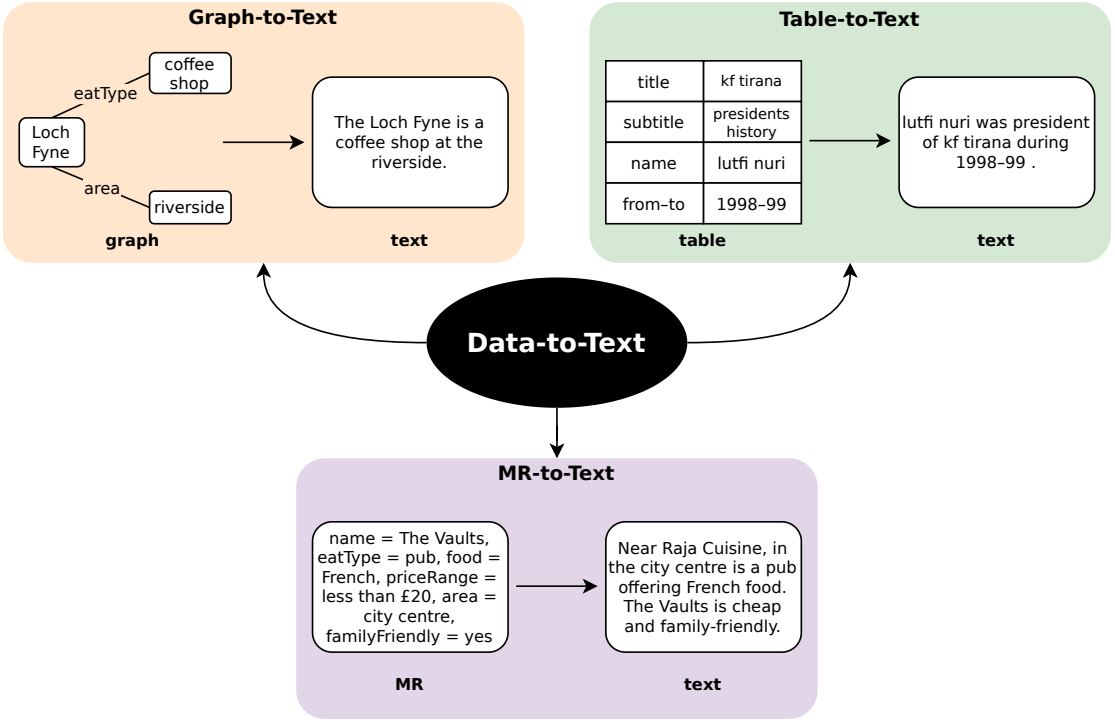


FIGURE 4.1: Three primary types of D2T: graph-to-text, meaning representation (MR)-to-text, and table-to-text.

This chapter contributes toward addressing this gap by presenting a multi-dataset evaluation of the factual consistency of LLMs in D2T. The study encompasses three principal D2T task types—*table-to-text*, *graph-to-text*, and *meaning representation (MR)-to-text*—as illustrated in Figure 4.1. It evaluates five widely used D2T datasets: E2E [3] (MR-to-text), ViGGo [62] (MR-to-text), WikiTableText [2] (table-to-text), DART [1] (graph-to-text), and WebNLG [7] (graph-to-text). To thoroughly analyze model behavior, twelve LLMs of varying scales are examined, representing five widely used open-source model families: BART [31], T5 [75], OPT [99], BLOOM [72], and Llama 2 [130]. To the best of our knowledge, few prior studies have evaluated factual consistency across a similarly broad combination of D2T datasets and LLM families.

Factual consistency is assessed through a combination of *automatic metrics* and *human evaluations*. Four widely used automatic factual consistency metrics are employed: SummaC-Conv, a natural language inference-based metric; NEOverlap, a named-entity-based metric; AlignScore, an information alignment-based metric; and QAFactEval, a question-answering-based metric. Additionally, to investigate how factual consistency evolves with increasing model size (in terms of the number of trainable parameters), the *average rate of change (AROC)* is computed for each LLM family.

The empirical analysis presented in this chapter suggests three key findings regarding the factual consistency of LLMs in D2T:

1. **Model Performance:** *Llama 2* generally attains the highest factual consistency among the evaluated LLM families. However, smaller models such as T5 and BART exhibit competitive performance on larger and lexically less diverse datasets (see section 4.5).

2. **Impact of Model Scale:** AROC-based analyses across multiple automatic metrics indicate that larger models within the evaluated LLM families are generally associated with higher factual consistency scores (see section 4.5). These trends are broadly consistent with the human evaluation results (see section 4.7).

3. **Effect of Source-Reference Divergence:** The results presented in section 4.6 suggest that source–reference divergence—a common phenomenon in D2T in which the reference text semantically deviates from the source data [39, 54]—is associated with reduced factual consistency in LLM-generated outputs.

4.2 Related Work

Factual consistency is a critical requirement for D2T systems [35, 52]. Early research by Tian et al. [13] identified factual inconsistency—termed *intrinsic hallucination*—as a fundamental challenge in D2T, while Xiao and Wang [121] further linked such inconsistencies to predictive uncertainty in generation models. Later, Li et al. [54] emphasized that factual consistency remains a persistent limitation, often overshadowing other dimensions such as readability and informativeness. Subsequent analyses [54, 93] have attributed factual inconsistency to multiple factors, including model architecture, model scale, and source-reference divergence. The advent of Transformer-based architectures [30, 111] has contributed to advances in D2T, leading to the emergence of LLMs that have demonstrated strong generalization across a variety of domains and languages [130]. Despite their strengths, several studies [34, 100, 116] report that LLMs,

while often achieving higher *readability* and *informativeness*, continue to struggle with maintaining *factual consistency*. Prior surveys [35, 54] highlight the prevalence of factual errors in LLM-generated D2T outputs, underscoring the need for more systematic evaluation frameworks. However, existing efforts have generally been limited in scope, focusing on a narrow range of models or datasets. To address this gap, the present study conducts a large-scale, systematic evaluation of factual consistency across multiple LLM families and D2T benchmarks, aiming to advance the understanding of faithfulness and inform the development of more factually reliable D2T systems.

4.3 Preliminaries

Although most of these concepts have been discussed in detail in Chapter 1, they are briefly revisited here for completeness. This section introduces the key foundational concepts that support our study, including an overview of D2T, the definition of factual consistency, and the concept of source-reference divergence. As the fundamentals of D2T and LLMs have already been presented in Chapter 1, the discussion here is limited to background information directly relevant to the evaluation framework employed in this chapter.

4.3.1 D2T

D2T [35] aims to generate coherent natural language text $r = r_1 r_2 \dots r_{|r|}$ from structured or semi-structured input data s (e.g., tables, graphs, or meaning representations). A D2T dataset typically consists of pairs $\mathcal{D} = \{(s, r)\}$, where each source s is associated with a reference text r drawn from the vocabulary $\mathcal{V}$, i.e., $r_i \in \mathcal{V}$ [16]. A model $\mathcal{M}_\theta$, parameterized by θ, is trained on $\mathcal{D}$ to maximize the likelihood of producing r given s as shown in Equation 4.1.

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{(s,r)\in\mathcal{D}} \mathcal{L}(\mathcal{M}_\theta(s), r) \tag{4.1}$$

Here, $\mathcal{L}$ represents the empirical loss function, typically the cross-entropy loss. During inference, the trained model $\mathcal{M}_{\theta^*}$ generates an output g from source data s by applying a decoding strategy [54], as expressed in Equation 4.2.

$$g \leftarrow \underset{\mathcal{V}^*}{\text{decoding}}\ \mathcal{M}_{\theta^*}(s) \tag{4.2}$$

Advancement in transformer-based architectures and large-scale pretraining have made LLMs the de facto predictive models for D2T, achieving notable improvements in fluency and content preservation. However, as discussed earlier, these models still face substantial challenges in maintaining factual consistency.

4.3.2 Factual Consistency in D2T

Factual consistency assesses whether the generated text (*claim*) accurately reflects the factual information in the source or reference data (*context*). In D2T, this involves verifying whether all facts present in the input data are correctly expressed in the generated text, and whether no extraneous or incorrect facts are introduced. A lack of factual consistency, as illustrated in Figure 4.2, can lead to the inclusion of misleading or unsupported claims, thereby undermining model reliability.

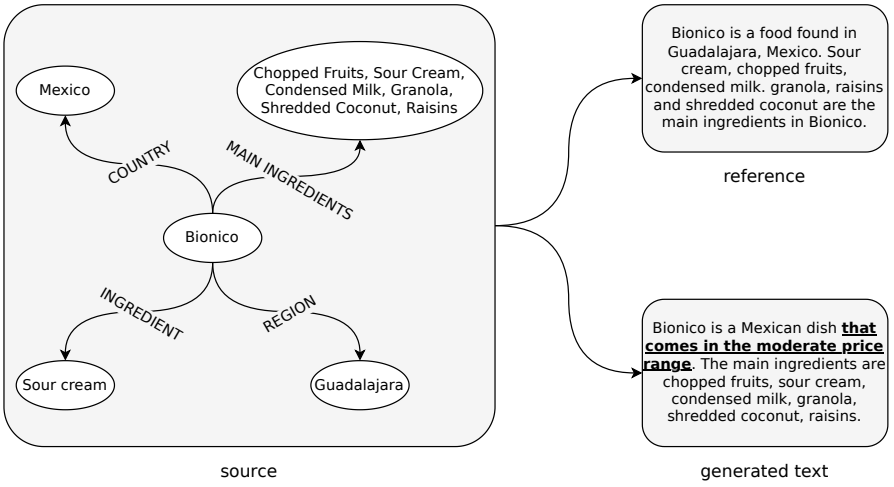


FIGURE 4.2: Example of factual inconsistency from the DART (graph-to-text) dataset, where the generated claim introduces incorrect information—**'that comes in the moderate price range'**—not found in the source data.

Evaluating factual consistency is important for determining a D2T model's faithfulness. Two major evaluation strategies exist: (i) *human assessment*, which provides direct but expensive and subjective evaluations, and (ii) *automatic metric-based evaluation*, which offers scalable and repeatable assessments. Studies [145, 146] have proposed automatic factual consistency metrics (e.g., entailment-based and question-answering-based models) that exhibit strong correlations with human judgments, making them practical tools for large-scale factual consistency evaluation.

4.3.3 Source-Reference Divergence

As previously discussed in Chapter 3 (subsection 3.3.3), this section revisits the concept of source-reference divergence for completeness. Source-reference divergence is a well-documented phenomenon in D2T [13, 39, 54, 68], referring to a semantic or factual mismatch between the structured source data and its corresponding reference text. An illustrative example is shown in Figure 4.3, where the reference text introduces an additional fact—**'is reasonably priced'**—that is not present in the source data.

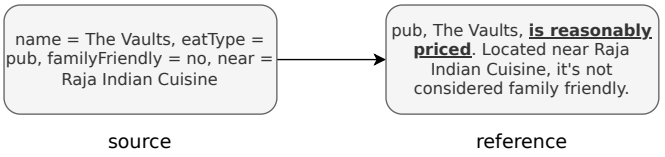


FIGURE 4.3: Illustration of source-reference divergence from the E2E [3] dataset. The reference text introduces additional information—**'is reasonably priced'**—not present in the source data.

This divergence typically arises from two main factors. First, D2T reference texts often include implicit or inferred information to improve naturalness and fluency, leading to deviations from the explicit content of the source. Second, many D2T datasets are constructed through heuristic pairing methods, where references and sources are not always perfectly aligned [54]. Such mismatched pairings are difficult to eliminate and are prevalent across benchmark datasets. Consequently, evaluating factual consistency under conditions of source-reference divergence is critical for understanding how LLMs handle imperfect or noisy supervision—an important consideration in the development and evaluation of D2T systems.

4.4 Experimental Setup

This section describes the datasets, LLMs, and experimental settings used in our experiments, along with statistical significance tests to ensure reproducibility.

TABLE 4.1: Summary of key statistics, D2T types, and domains for the five incorporated datasets. The table shows the average length (in tokens), unique tokens, and total tokens for both sources (linearized to text) and references.

dataset	D2T types	domain	size	source (linearized)			reference		
			(# instances)	type-token ratio (TTR)	unique tokens	total tokens	type-token ratio (TTR)	unique tokens	total tokens
E2E	mr-to-text	closed	36,856	0.0001	125	1M	0.0057	4.5K	885K
ViGGo	mr-to-text	closed	6,900	0.0029	618	206K	0.0294	4.4K	148K
WikiTableText	table-to-text	open	13,318	0.0621	29K	469K	0.1302	24K	185K
DART	graph-to-text	open	70,524	0.0178	44K	2.5M	0.0328	45K	1.5M
WebNLG	graph-to-text	open	38,872	0.0057	7K	1.2M	0.0228	19K	905K

4.4.1 Datasets

We use five well-known D2T datasets for three primary task types: DART and WebNLG for graph-to-text, WikiTableText for table-to-text, and E2E and ViGGO for MR-to-text. The key statistics for these datasets are in Table 4.1, all obtained from Kasner et al. [138] and Transformers [137]. E2E [3] is a restaurant-domain MR-to-text dataset, and ViGGO [65] focuses on dialogue MR-to-text in the video game domain. WikiTableText [2] provides table-to-text pairs from Wikipedia with high lexical diversity (or lexical richness) (as TTR is higher for both source and reference). DART [1] provides text from knowledge graph triplets, while WebNLG [7] focuses on RDF-to-text from DBPedia triplets. We acknowledge that the incorporated datasets may pose a potential risk of data contamination in the LLM evaluation setup; however, this limitation has been addressed by considering the nature of the D2T task and the adopted evaluation paradigm, as discussed in subsection 3.4.4.

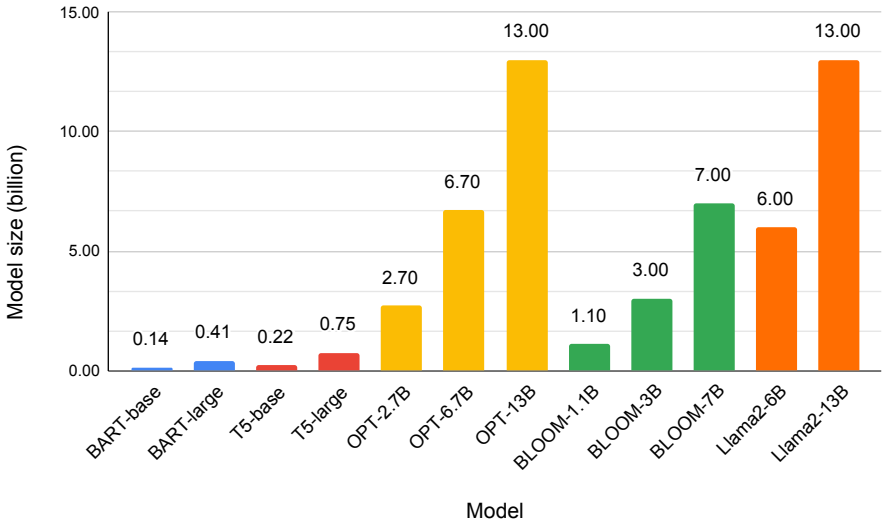


FIGURE 4.4: All twelve LLMs across five widely used families (BART, T5, OPT, BLOOM, and Llama2) arranged by their model sizes.

4.4.2 Models

We include twelve LLMs from five families—BART, T5, BLOOM, OPT, and Llama 2—downloaded from Hugging Face [137]. All these twelve LLMs are shown in figure 4.4, along with their sizes. For the BART, T5, OPT, BLOOM, and Llama 2 families, we select two models from BART and T5, three models from OPT and BLOOM, and two models from Llama 2. BART [31] and T5 [75] are encoder-decoder models, with BART pre-trained using text denoising and T5 with unified NLP tasks. BLOOM [72] and OPT [99] are decoder-only models, trained on ROOT and Reddit/Pile/RoBERTa,

respectively. Llama 2 [130] is trained with pre-training, fine-tuning, and reinforcement learning. These LLM families were selected for their varied pretraining dataset and sizes, enabling a broad evaluation of factual consistency across different D2T tasks.

4.4.3 Experimental Setting and Statistical Significant Testing Details

In our experiments, we fine-tuned all LLMs for D2T tasks using QLoRA (Quantized Low-Rank Adapter) [85], a parameter-efficient fine-tuning method with a learning rate of $1e^{-04}$ and 4-bit quantization. Each model was fine-tuned for three epochs on all five datasets, with source and generated text lengths fixed at 256 tokens. We used the AdamW optimizer [139] at `torch.bfloat16` precision for QLoRA. Text generation was performed through popular nucleus sampling decoding approach with nucleus size ($p = 0.95$). All experiments were run on a local system with an NVIDIA RTX A6000 (48 GB) GPU. We assessed statistical significance using Welch's t-test [140], comparing each LLM to the best performer in its family, with a significance level of $p < 0.05$ and a sample size of 6. Since input linearization and prompt design play a crucial role in LLM-based D2T generation, the complete templates and prompting strategies adopted in this work are provided in detail in subsection 6.5.6.

4.4.4 Average Rate of Change (AROC)

Increasing model size within an LLM family often affects factual consistency. This effect can be quantified using the average rate of change (AROC), which estimates the change in factual consistency between the smallest model $\mathcal{M}_1$ and the largest model $\mathcal{M}_2$ of the family, based on their sizes m_1 and m_2. AROC is calculated within the range $[m_1, m_2]$ using equation 4.3.

$$\begin{aligned} \text{AROC} &= \frac{\text{absolute change in factual consistency}}{\text{absolute change in model size}} \\ &= \frac{f(\mathcal{M}_2) - f(\mathcal{M}_1)}{m_2 - m_1} \end{aligned} \tag{4.3}$$

LLM families differ significantly in model size scales; for instance, T5 sizes are in millions, while Llama2 sizes are in billions, impacting factual consistency. Comparing LLM families using AROC (equation 4.3) alone is therefore insufficient. Instead, we apply AROC on a logarithmic scale (equation 4.4), capturing relative changes by converting absolute differences into ratios (equation 4.5).

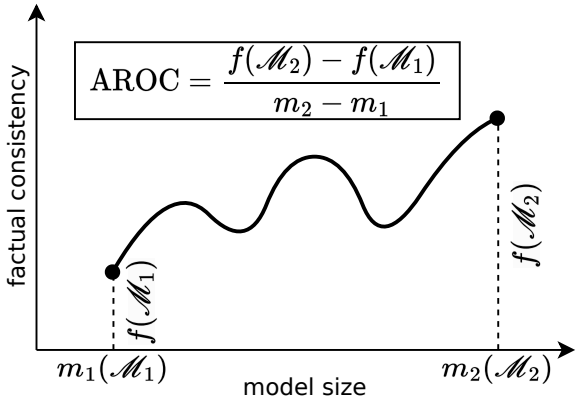


FIGURE 4.5: AROC of the function $f(\cdot)$ (representing factual consistency) for the model size interval $[m_1, m_2]$ between two LLMs, $\mathcal{M}_1$ and $\mathcal{M}_2$.

$$\text{AROC} = \frac{\log(f(\mathcal{M}_2)) - \log(f(\mathcal{M}_1))}{\log(m_2) - \log(m_1)} \tag{4.4}$$

$$= \frac{\log(f(\mathcal{M}_2)/f(\mathcal{M}_1))}{\log(m_2/m_1)} \tag{4.5}$$

A positive AROC value (equation 4.4 or equation 4.5) indicates a positive correlation between model size increment and factual consistency, while the magnitude of the AROC value reflects the strength of this correlation.

4.5 Evaluation of Factual Consistency: Results and Discussion

This section presents our extensive evaluation of factual consistency across five widely used LLM families and five prominent D2T datasets, employing four widely used metrics: SUMMAC-CONV, NEOVERLAP, ALIGNSCORE, and QAFACTEVAL.

4.5.1 Results of Factual Consistency Evaluation Using SummaC-Conv

NLI (natural language inference) models are commonly used to assess factual consistency due to their similarity to the task. However, a key limitation of NLI models is their focus on sentence-level granularity, which is insufficient for evaluating factual consistency that requires document-level analysis. SUMMAC-CONV [96] addresses this limitation by providing document-level granularity. It segments both the context and claim into sentences, computing NLI scores for each sentence pair. Additionally, SUMMAC-CONV

TABLE 4.2: Factual consistency evaluated through SUMMAC-CONV. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3).

family	model	model size (billion)	E2E	WikiTableText	ViGGo	DART	WebNLG
BART	BART-base	0.14	0.3805	0.3111	0.2330	0.3004	0.2894
	BART-large	0.41	**0.4110**	**0.3609**	**0.2498**	**0.3786**	**0.4173**
T5	T5-base	0.22	0.3899	0.3625	0.2435	0.3582	0.3687
	T5-large	0.75	**0.4135**	**0.4082**	**0.2684**	**0.4264**	**0.4593**
OPT	OPT-2.7B	2.70	**0.3637**	**0.4120**	0.2971	0.3926	0.4781
	OPT-6.7B	6.70	0.3597	0.4060	0.3139	0.3937	0.4799
	OPT-13B	13.00	0.3621	0.4092	**0.3291**	**0.4137**	**0.4970**
BLOOM	BLOOM-1.1B	1.10	0.3777	0.3871	0.2874	0.3850	0.3853
	BLOOM-3B	3.00	0.3842	0.4136	**0.3177**	0.4119	0.4610
	BLOOM-7B	7.00	**0.3872**	**0.4184**	0.3170	**0.4144**	**0.4780**
Llama 2	Llama2-7B	6.00	**0.4569**	0.4898	0.3622	0.5138	**0.5739**
	Llama2-13B	13.00	0.4540	**0.5009**	**0.3776**	**0.5201**	0.5693

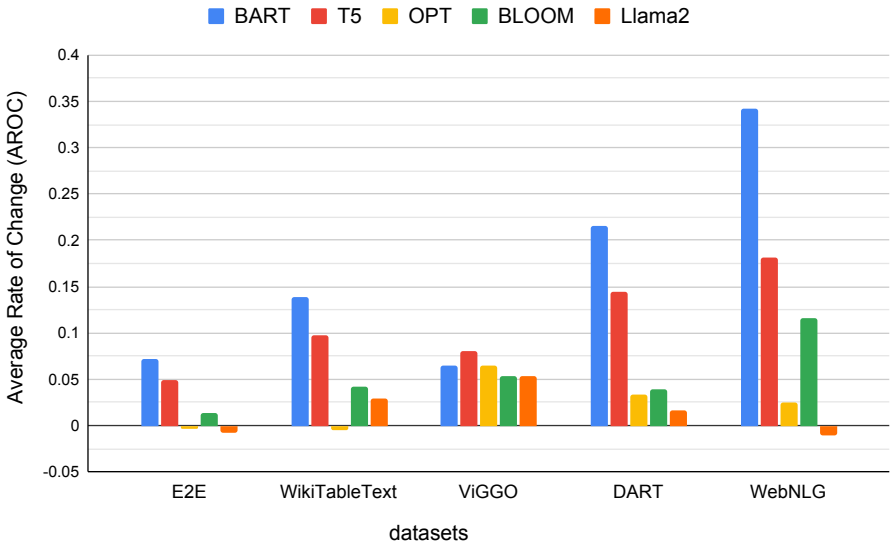


FIGURE 4.6: AROC values for all LLM families across the five D2T datasets based on SUMMAC-CONV scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency.

employs binning techniques and one-dimensional convolution to mitigate the impact of outliers, effectively capturing the factual consistency between the claim and context. SUMMAC-CONV is a commonly used metric for evaluating factual consistency.

Table 4.2 presents the factual consistency scores measured by SUMMAC-CONV across the five LLM families and five D2T datasets. A higher SUMMAC-CONV value indicates better factual consistency. Meanwhile, figure 4.6 displays the AROC values of SUMMAC-CONV, derived from table 4.2, for all LLM families and datasets.

TABLE 4.3: Factual consistency evaluated through NEOVERLAP. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3).

family	model	model size (billion)	E2E	WikiTableText	ViGGo	DART	WebNLG
BART	BART-base	0.14	0.8305	0.5183	0.7077	0.5643	0.3725
	BART-large	0.41	**0.8368**	**0.5700**	**0.7892**	**0.6204**	**0.4761**
T5	T5-base	0.22	0.8305	0.5556	0.7815	0.6164	0.4195
	T5-large	0.75	**0.8593**	**0.5982**	**0.8985**	**0.6813**	**0.5021**
OPT	OPT-2.7B	2.70	0.8323	**0.6267**	0.8908	0.6181	0.5385
	OPT-6.7B	6.70	0.8142	0.6058	0.9215	0.6358	0.5546
	OPT-13B	13.00	**0.8431**	0.6075	**0.9369**	**0.6499**	**0.5546**
BLOOM	BLOOM-1.1B	1.10	0.8133	0.5858	0.8938	0.5973	0.4373
	BLOOM-3B	3.00	0.8070	0.6050	0.9200	**0.6326**	0.4844
	BLOOM-7B	7.00	**0.8206**	**0.6208**	**0.9200**	0.6310	**0.5203**
Llama 2	Llama2-7B	6.00	0.8674	0.6350	0.9462	**0.7269**	0.6368
	Llama2-13B	13.00	**0.8683**	**0.6675**	**0.9523**	0.7158	**0.6532**

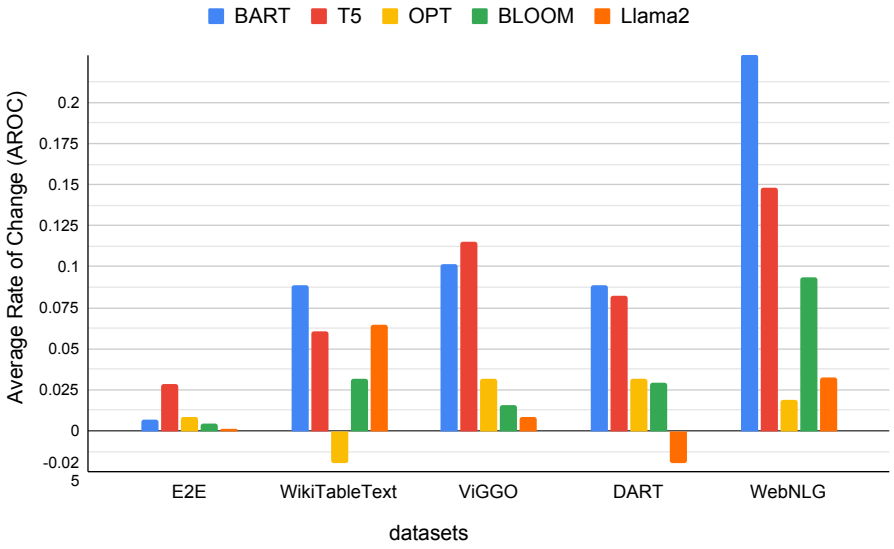


FIGURE 4.7: AROC values for all LLM families across the five D2T datasets based on NEOVERLAP scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency.

4.5.2 Results of Factual Consistency Evaluation Using NEOverlap

Named entities (NEs) are key indicators of factual knowledge in texts and are often used to assess factual consistencies. NEOVERLAP measures factual consistency by comparing the overlap of NEs between the claim and the context. We utilized the metric implementation from [96], focusing on specific NE types such as `PERSON`, `LOCATION`, and `ORGANIZATION`. This metric returns a value of 1 if the generated text is factually consistent with the reference in terms of NE overlap, and 0 if there is little or no factual consistency. In the D2T context, the reference text is treated as the context, and the generated text as the claim.

Table 4.3 displays the factual consistency scores evaluated using NEOverlap across the five LLM families and D2T datasets. Figure 4.7 presents the AROC values of NEOverlap, derived from table 4.3, for all LLM families and datasets.

Table 4.4: Factual consistency evaluated through AlignScore. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3).

family	model	model size (billion)	E2E	WikiTableText	ViGGo	DART	WebNLG
BART	BART-base	0.14	0.6825	0.3189	0.2579	0.4064	0.3869
	BART-large	0.41	**0.7663**	**0.4158**	**0.3538**	**0.6013**	**0.6551**
T5	T5-base	0.22	0.7049	0.4045	0.3238	0.5755	0.5457
	T5-large	0.75	**0.7646**	**0.4513**	**0.4459**	**0.7050**	**0.7207**
OPT	OPT-2.7B	2.70	0.6931	**0.4897**	0.5439	0.6326	0.7764
	OPT-6.7B	6.70	0.6792	0.4835	0.6036	0.6374	0.7876
	OPT-13B	13.00	**0.6999**	0.4810	**0.6356**	**0.6714**	**0.8145**
BLOOM	BLOOM-1.1B	1.10	0.7147	0.4887	0.4548	0.6175	0.6099
	BLOOM-3B	3.00	**0.7468**	**0.5326**	0.5653	0.6723	0.7507
	BLOOM-7B	7.00	0.7280	0.5296	**0.5884**	**0.6848**	**0.7797**
Llama 2	Llama2-7B	6.00	0.8106	0.5905	0.6798	**0.8192**	**0.8768**
	Llama2-13B	13.00	**0.8186**	**0.6144**	**0.7163**	0.8191	0.8749

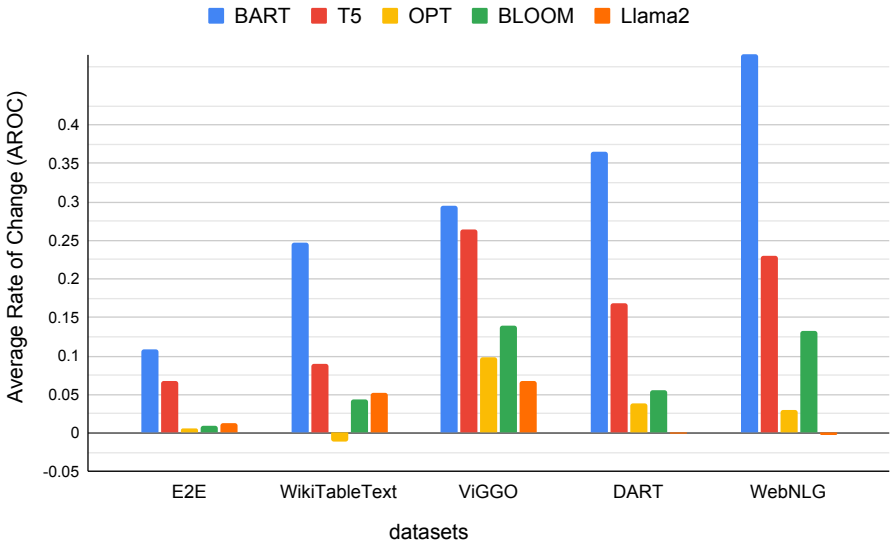


Figure 4.8: AROC values for all LLM families across the five D2T datasets based on AlignScore scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency.

4.5.3 Results of Factual Consistency Evaluation Using AlignScore

Unlike most factual consistency metrics, which typically focus on single tasks like semantic similarity or NLI (natural language inference), AlignScore [146] is trained on multiple tasks (fact verification, information retrieval, NLI, QA, paraphrasing, semantic similarity, and summarization) in a unified manner. AlignScore assesses factual consistency by analyzing the alignment of information between the claim and context,

capturing various qualities, including factual consistency. A key feature of AlignScore is its ability to handle longer contexts by splitting the context into chunks and aligning each chunk with the claim's sentences. A higher AlignScore reflects greater factual consistency. In our experiment, AlignScore refers specifically to the AlignScore-large variant, which utilizes the RoBERTa-large model for information alignment. Zha et al. [146] shows that AlignScore has been reported to achieve stronger correlations than several previously proposed measures, including LLM-based methods, in evaluating factual accuracy.

Table 4.4 presents the factual consistency scores evaluated using AlignScore across all LLM families and D2T datasets. Figure 4.8 shows the AROC values of AlignScore, obtained from table 4.4, for all LLM families and datasets..

Table 4.5: Factual consistency evaluated through QAFactEval. **Boldface** shows best result per LLM family and dataset. All results statistically significant (see 4.4.3).

family	model	model size (billion)	E2E	WikiTableText	ViGGo	DART	WebNLG
BART	BART-base	0.14	3.4109	1.8008	1.3894	2.2541	2.1009
	BART-large	0.41	**3.7393**	**2.3559**	**1.9500**	**3.1206**	**3.3009**
T5	T5-base	0.22	3.6086	2.1887	1.6307	2.7631	2.5079
	T5-large	0.75	**3.7393**	**2.6267**	**2.0971**	**3.4325**	**3.3865**
OPT	OPT-2.7B	2.70	3.2732	2.5878	2.4776	3.1245	3.5771
	OPT-6.7B	6.70	**3.2930**	**2.6805**	2.7332	3.1335	3.6794
	OPT-13B	13.00	3.2768	2.6085	**2.8302**	**3.2894**	**3.7640**
BLOOM	BLOOM-1.1B	1.10	3.5579	2.2128	2.0289	3.0540	2.9072
	BLOOM-3B	3.00	3.6179	2.4706	2.5181	3.2751	3.4617
	BLOOM-7B	7.00	**3.6245**	**2.5299**	**2.5265**	**3.3069**	**3.5181**
Llama 2	Llama2-7B	6.00	4.1619	3.0296	2.9628	4.0208	**4.1166**
	Llama2-13B	13.00	**4.1874**	**3.0933**	**3.0950**	**4.0411**	4.0892

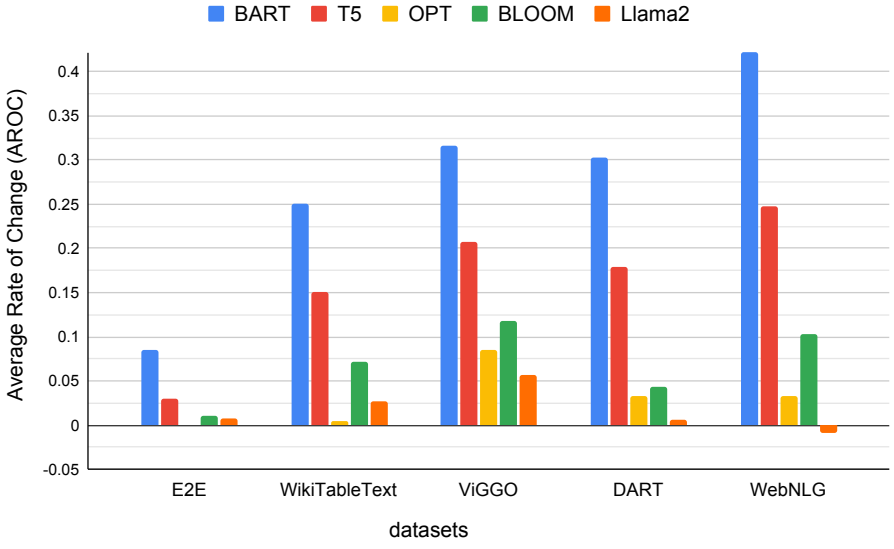


Figure 4.9: AROC values for all LLM families across the five D2T datasets based on QAFactEval scores. Higher positive AROC values reflect a stronger positive correlation between increases in model size and factual consistency.

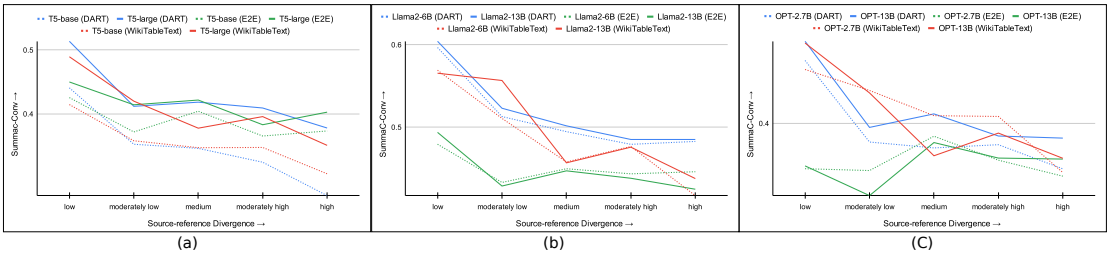


FIGURE 4.10: Impact of source-reference divergence on factual consistency (measured through SUMMAC-CONV) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines.

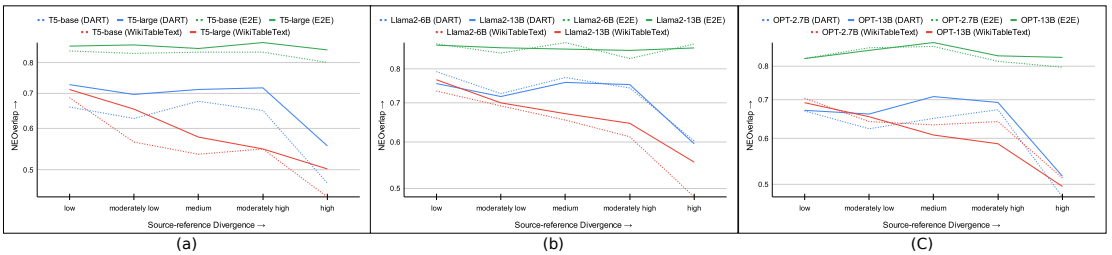


FIGURE 4.11: Impact of source-reference divergence on factual consistency (measured through NEOVERLAP) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines.

4.5.4 Results of Factual Consistency Evaluation Using QAFactEval

QAFACTEVAL [145] is a factual consistency metric that utilizes question generation and answering (QGA). It comprises four key components: 1) Answer Selection: Noun phrase chunks are extracted as answers from the generated text; 2) Question Generation: The BART-large model generates questions based on these answers; 3) Question Answering: The Electra-large model answers the questions using the source text; 4) Answer Overlap Determination: The overlap is assessed using the LERC metric, which provides a score from 1 to 5. QAFACTEVAL surpasses existing QGA-based metrics and complements NLI-based evaluations, with higher scores indicating greater factual consistency, reaching a maximum of 5.

Table 4.5 presents the factual consistency scores evaluated using QAFACTEVAL across all LLM families and D2T datasets. Figure 4.9 illustrates the AROC values of QAFACTEVAL, derived from table 4.5, for five LLM families across all D2T datasets.

4.5.5 Discussion of Factual Consistency Results

From the factual consistency results across the four metrics—SummaC-Conv, NEOverlap, AlignScore, and QAFactEval (Tables 4.2, 4.3, 4.4, and 4.5)—we observe several key findings. The Llama2 family consistently outperforms other models in factual consistency across all four metrics and nearly all D2T datasets. However, smaller LLMs like T5-large and BART-large remain competitive, often outperforming larger families such as BLOOM and OPT on datasets like E2E (where lexical richness, as measured by TTR in table 4.1, is low, and dataset size is large) and DART (which has a large number of instances, as shown in table 4.1). This indicates that smaller LLM families can excel depending on dataset characteristics. Larger LLM families like Llama2, BLOOM, and OPT tend to perform better on datasets such as WikiTableText and ViGGo, likely due to their higher lexical richness (see table 4.1). Similarly, for the WebNLG dataset, T5 and BART produce results similar to those of BLOOM and OPT, likely because WebNLG is a large dataset with moderate lexical richness. Hence, while larger LLMs often exhibit better factual consistency than smaller ones, the characteristics of D2T datasets, particularly their size and lexical richness, are essential in shaping this performance.

AROC values across all four metrics (figures 4.6, 4.7, 4.8, and 4.9) show that increasing model size generally improves the factual consistency of LLM families across all D2T datasets. The T5 and BLOOM families exhibit significant improvements in factual consistency with larger models. In contrast, the Llama2 family shows less noticeable gains, and in some cases, a decline in performance as model size increases. We hypothesize that this may be due to the log-scale applied to the AROC values, where smaller LLM families (like T5 and BART) show a higher relative improvement in factual consistency compared to larger families (like Llama2). Overall, the AROC values across all four metrics suggest that, for most LLM families, larger model sizes are correlated with improved factual consistency.

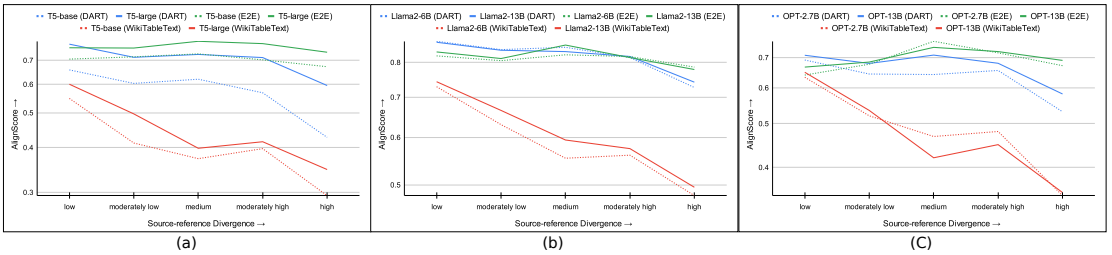


FIGURE 4.12: Impact of source-reference divergence on factual consistency (measured through ALIGNSCORE) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines.

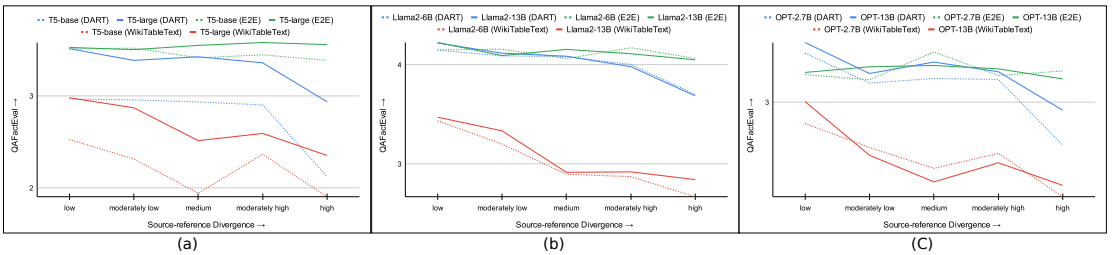


FIGURE 4.13: Impact of source-reference divergence on factual consistency (measured through QAFACTEVAL) at time of inference on three popular D2T datasets (DART, E2E and WikiTableText). The three sub-figures (a, b and c) represent results from three different families (T5, Llama2 and OPT), where the relatively smaller models (T5-base, Llama2-7B, and OPT-2.7B) are represented using dotted lines, while the relatively larger variants within each considered model family (T5-large, Llama2-13B, and OPT-13B) are represented using solid lines.

4.6 Impact of Source-reference Divergence on Factual Consistency

This section investigates how source-reference divergence impacts the factual consistency of LLMs in D2T tasks during the inference phase (i.e., at the time of text generation). For this analysis, we focus on three LLM families (T5, OPT, and Llama 2) with six specific models: T5-base, T5-large, OPT-2.7B, OPT-13B, Llama2-7B, and Llama2-13B. We evaluate the factual consistency of these LLMs using the four automatic metrics: SUMMAC-CONV, NEOVERLAP, ALIGNSCORE, and QAFACTEVAL. To cover the three main types of D2T tasks, we select one dataset for each: E2E for MR-to-Text, WikiTableText for Table-to-Text, and DART for Graph-to-Text. To better illustrate the effect of source-reference divergence, we divide the test/inference sets of each dataset into five groups based on their average source-reference divergence: `low`, `moderately low`, `medium`, `moderately high`, and `high`. The `low` group contains instances with minimal divergence, while the `high` group includes instances with significant divergence.

Given that the source data (s) in D2T tasks is often semi-structured and non-textual, traditional methods of calculating source-reference divergence, such as higher-order n-gram matching or contextual representation, may fail to capture critical information. Two common strategies for assessing source-reference divergence are unigram-based and longest common subsequence (LCS)-based approaches [39]. We choose the LCS-based approach over the unigram-based method because it often yields better performance by flexibly matching n-grams and capturing skip context. Source-reference divergence ($\text{div}(s, r)$) is then defined based on the differences between the source text (s) and the reference text (r) at the unigram level, as follows:

$$\text{div}(s, r) = 1 - \frac{1}{\max(|s|, |r|)} \text{LCS}(s, r)$$

Where $|x|$ denotes the cardinality, and $\text{LCS}(x, y)$ represents the length of the longest common subsequence between x and y (both x and y are segmented by whitespaces).

From the results shown in figures 4.10, 4.11, 4.12, and 4.13, we observe two important details regarding factual consistency in the presence of source-reference divergence. First, the factual consistency of all LLMs, regardless of model family or size, decreases with an increase in source-reference divergence. However, in the E2E dataset, the decrease in factual consistency with source-reference divergence is lower compared to the DART and WikiTableText datasets. We believe this is due to the higher lexical richness of the DART and WikiTableText datasets compared to the E2E dataset. Second, it is evident that within each LLM family, larger models consistently outperform smaller ones in terms of factual consistency. This suggests that increasing the model size of LLMs makes them more agnostic to source-reference divergence. This finding indicates that choosing larger model-sized LLMs is beneficial when source-reference divergence is present in D2T tasks.

4.7 Human Evaluation

Text generation tasks like D2T require human evaluation due to the complexity of natural language, and human assessment is often regarded as the gold standard for factual consistency [105, 106]. For our human evaluation of factual consistency, we selected three LLM families (T5, OPT, and Llama 2), each with two models (T5-base, T5-large, OPT-2.7B, OPT-13B, Llama2-7B, and Llama2-13B), and three D2T datasets: E2E, WikiTableText, and DART. In the annotator selection process, three graduate students from different academic disciplines were recruited to minimize potential bias. Participation

was voluntary, and all necessary arrangements were provided. Before beginning the annotation task, each annotator was given a set of five predefined input-output examples along with gold-standard labels (provided by the authors) to help them understand the objectives of the task. Annotators were instructed to assess factual consistency by identifying any incorrect or irrelevant information in the generated text with respect to the given source data. To ensure independent judgments, annotators were explicitly told not to discuss their evaluations with one another. For the main annotation phase, each annotator evaluated a set of 20 generated texts for every combination of LLM and dataset. Their responses were recorded as binary 'yes'/'no' answers. Finally, the results were aggregated using a majority voting scheme and reported as percentages. The inter-annotator agreement in our experiments, measured by Krippendorff's alpha (α), was 0.76, indicating moderate to high agreement. The detailed annotation guidelines used for the human evaluation are presented below.

Annotation Guidelines

Participation in this annotation task is entirely voluntary. By proceeding with the evaluation, you acknowledge that you have read and understood the instructions and agree to participate in the study.

Please carefully read the provided input data (source), reference text, and generated output before assigning a score. Your task is to evaluate the factual consistency of the generated text according to the following definition:

- **Factual consistency.** The degree to which the generated text is supported by the source data and reference text, without introducing unsupported or contradictory information.

Provide your score using the following five-point scale:

- **Score 1 (Very Poor):** The generated text contains substantial unsupported, contradictory, or misleading information and exhibits very low factual consistency.
- **Score 2 (Poor):** The generated text contains several factual inconsistencies or unsupported statements, resulting in low factual consistency.
- **Score 3 (Moderate):** The generated text is generally factually consistent but contains some unsupported, inaccurate, or ambiguous information.
- **Score 4 (Good):** The generated text is largely factually consistent, with only minor unsupported or inaccurate content.

- **Score 5 (Excellent):** The generated text is fully or nearly fully supported by the source data and reference text, with no significant factual inconsistencies.

Evaluate each sample independently and do not allow previous evaluations to influence subsequent judgments. Do not discuss the evaluation with other annotators. In addition to assigning a numerical score, provide a brief justification for your assessment, identifying any factual inconsistencies as (1) *missing fact*, (2) *fabricated fact*, and/or (3) *irrelevant fact*.

TABLE 4.6: Human evaluation of factual consistency. A lower value indicates higher factual consistency.

family	model	E2E	DART	WikiTableText
T5	T5-base	23%	32%	37%
	T5-large	15%	22%	32%
OPT	OPT-2.7B	19%	25%	18%
	OPT-13B	16%	22%	14%
Llama 2	Llama2-7B	12%	15%	21%
	Llama2-13B	12%	17%	17%

Table 4.6 shows the human evaluation results for factual consistency. We can see that, in human evaluation as well, the Llama 2 family performs better than OPT and T5. Similar to the automatic metric results, we find that T5 performs similarly to the OPT family on the DART and E2E datasets. However, OPT outperforms T5 on the WikiTableText dataset. The broad agreement between human and automatic evaluations provides additional evidence for the observed trends.

4.8 Conclusion

This chapter presents an extensive evaluation of the *factual consistency* of LLMs in D2T tasks. Through extensive experimentation involving five representative LLM families and five widely used D2T datasets, the study employs four state-of-the-art automatic factual consistency metrics—SUMMAC-CONV, QAFACTEVAL, ALIGNSCORE, and NEOVERLAP—in conjunction with human evaluations to provide a multi-faceted assessment of model performance. The findings of this study suggest three main observations. First, *Llama 2* generally attains the highest factual consistency scores across the evaluated datasets, while smaller models such as T5 and BART also exhibit competitive performance, particularly on larger and lexically less diverse datasets. Second, analysis using the *average rate of change (AROC)* metric indicates a generally positive

association between model size and factual consistency within the evaluated model families. Third, the results suggest that *source-reference divergence*—where the reference text deviates semantically from the source data—may be associated with lower factual consistency in LLM-generated outputs. Collectively, these findings provide empirical observations into the factors influencing factual consistency in D2T and provide empirical evidence that may inform future research aimed at developing approaches that may improve faithfulness and consistency in D2T generation systems.

Chapter 5

Enhancing Factual Consistency in LLM-Based D2T

Building on the insights gained from the previous chapter, which provided a broad evaluation of factual consistency in LLM-based D2T, this chapter[1] investigates a method for improving *faithfulness* by mitigating factual inconsistencies associated with *source–reference divergence*. Source–reference divergence, referring to semantic mismatches between the input data and its corresponding reference text, has been associated with unfaithful or hallucinated content in LLM-generated outputs. To address this issue, this chapter introduces a *divergence-aware selective generation* framework intended to improve factual consistency in LLM-based D2T systems. The proposed approach explicitly models source–reference divergence using three complementary metrics—Unigram-Level Divergence, Named Entity-Based Divergence, and Field-Aware Divergence—and integrates these divergence signals into the generation process to promote selective and faithful content realization. In this work, the goal is to enhance faithfulness in terms of factual consistency by enabling LLMs to generate text that remains aligned with the input data despite reference variations. Experimental results across the evaluated D2T datasets and LLM architectures indicate that the proposed framework generally improves factual consistency while largely maintaining fluency and informativeness. Overall, this chapter presents and discusses a divergence-aware strategy that improves factual consistency under the evaluated experimental conditions.

[1]The work presented in this chapter is based on our paper: Mahapatra, J. and Garain, U., 2025. Divergence-Aware Selective Data-to-Text Generation with LLMs for Factual Consistency. *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*. [69]

5.1 Background

The advent of LLMs such as T5 [55], OPT [99], and Qwen [33] has contributed to advances in D2T, leading to improvements in fluency, generalization, and informativeness on a variety of evaluated benchmarks. However, maintaining *factual consistency*—the degree to which generated text accurately reflects the information present in the input data—remains a persistent challenge in LLM-based D2T systems [35, 39]. The occurrence of *hallucinated content*, i.e., information that is unsupported or incorrect with respect to the source input, can critically undermine the reliability and trustworthiness of generated text, particularly in high-stakes domains such as healthcare and finance [50].

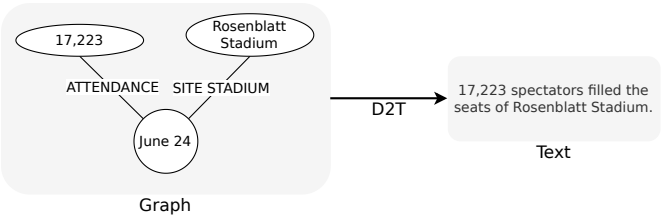


FIGURE 5.1: An illustration of graph-to-text generation, a representative form of Data-to-Text (D2T) generation.

Among the multiple factors contributing to factual inconsistency, one critical source of error is *source–reference divergence*—a semantic or structural mismatch between the input data and its corresponding reference text within training corpora. Such divergence commonly arises from loosely aligned annotations or inconsistencies in real-world data, leading models to internalize spurious correlations and generate factually incorrect outputs. Despite its potential impact on faithfulness, this phenomenon has not been systematically analyzed or quantified in prior D2T research.

This chapter addresses this gap by investigating the influence of source–reference divergence in LLM-based D2T and by introducing a principled approach to mitigate its effects. It presents a detailed analysis of divergence across three representative D2T datasets—WikiTableText (table-to-text), ViGGo (MR-to-text), and WebNLG (graph-to-text)—using three complementary measures: *unigram-level divergence*, *named entity-based divergence*, and *field-aware divergence*. Building upon these insights, the chapter proposes a *divergence-aware selective generation* framework, in which LLMs selectively abstain from generation when divergence exceeds a specified threshold, thereby prioritizing factual consistency. Experiments conducted with three widely used LLM families—Qwen2.5 [33], FLAN-T5 [55], and OPT [99]—depict consistent improvements in factual consistency, as measured by AlignScore, QAFactEval, SummaC-Conv, and

UniEval-Fact, particularly under moderate coverage thresholds. These findings highlight the significance of explicitly modeling source–reference divergence as a means to enhance the faithfulness and factual consistency of LLM-based D2T systems.

5.2 Quantifying Source–Reference Divergence in Data-to-Text Generation

A preliminary quantification of *source–reference divergence* in Data-to-Text (D2T) generation was discussed in Chapter 3 (subsection 3.3.3); however, in this chapter, the concept is examined in greater detail. Data-to-Text generation involves transforming structured or semi-structured data into coherent and contextually appropriate natural language text [35]. Despite the substantial advancements enabled by LLMs such as T5 [55], OPT [99], and Qwen [33], maintaining factual consistency between the generated output and the source data continues to pose a challenge [35, 39]. A major contributing factor to this challenge is *source–reference divergence*—a semantic or structural mismatch between the input data and its corresponding reference text used during training. Such divergence often arises from loosely aligned annotations or natural variability present in real-world datasets [7, 58], causing models to internalize spurious correlations and subsequently produce factually inconsistent or hallucinated outputs.

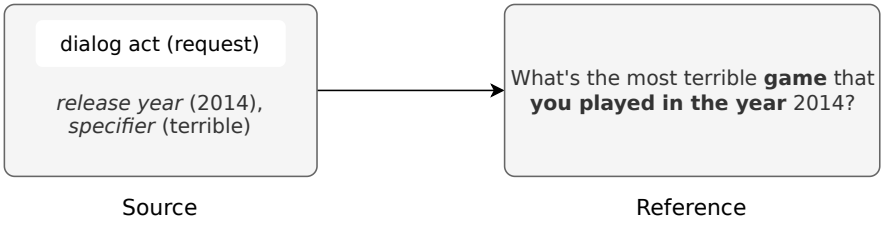


Figure 5.2: An illustration of source–reference divergence. The boldfaced segments in the reference text indicate content that diverges substantially from the source data.

To systematically analyze this phenomenon, we quantify source–reference divergence through three complementary measures—*unigram-level*, *named entity-based*, and *field-aware divergence*—collectively denoted as f_{div}. These measures capture distinct aspects of semantic and structural divergence, providing a multifaceted understanding of how closely the reference text aligns with its corresponding input.

Unigram-Level Divergence. This measure captures lexical divergence using unigram overlap between the source x and reference y, computed via Jaccard similarity. A

higher overlap indicates lower divergence:

$$\text{Unigram}_{\text{div}}(x, y) = 1 - \frac{\left|\text{unigram}(x) \cap \text{unigram}(y)\right|}{\left|\text{unigram}(x) \cup \text{unigram}(y)\right|}$$

where $\text{unigram}(t)$ denotes the set of unigrams extracted from text t.

Named Entity-Based Divergence. This measure focuses on factual elements by comparing named entities between the source and reference texts:

$$\text{NE}_{\text{div}}(x, y) = 1 - \frac{\left|\text{NE}(x) \cap \text{NE}(y)\right|}{\left|\text{NE}(y)\right|}$$

Here, $\text{NE}(x)$ and $\text{NE}(y)$ denote the sets of named entities extracted from the source and reference, respectively. Since named entities often encode factual information, this measure highlights discrepancies most relevant to factual consistency.

Field-Aware Divergence. Designed for structured inputs such as tables, graphs, or meaning representations (MRs), this measure assesses divergence at the field level, considering each field's contribution independently:

$$\text{Field}_{\text{div}}(x, y) = 1 - \frac{1}{|\mathcal{F}_x|} \sum_{f \in \mathcal{F}_x} w_f \cdot \frac{|\text{n-gram}(f) \cap \text{n-gram}(y)|}{|\text{n-gram}(y)|}$$

where $\mathcal{F}_x$ is the set of fields in the structured input x, $\text{n-gram}(f)$ denotes the set of n-grams extracted from the field value f, and w_f is a smoothed weight controlling the relative importance of each field.

These three formulations collectively provide a principled framework for quantifying source–reference divergence in D2T tasks. By capturing lexical, factual, and structural discrepancies, they enable an systematic analysis of how divergence affects faithfulness—particularly factual consistency—in LLM-based D2T.

5.3 Divergence-aware Selective Generation

Our approach—source-reference divergence-aware selective generation—is inspired by selective classification [147], where a model abstains under uncertainty. Analogously, here LLM refrains from generating when the input exhibits high source-reference divergence with the generated text, based on the hypothesis that such cases are more prone to factual inconsistency. This strategy enhances output reliability, especially in high-stakes (safety-critical) domains. Formally, the selective generation process is defined as follows:

$$\text{Selective generation} = \begin{cases} \widehat{y} & \text{if } \widehat{y} \sim \mathcal{G}_{\theta^*}(x) \text{ and } f_{\text{div}}(x, \widehat{y}) \leq \Phi \\ \text{abstain} & \text{otherwise} \end{cases} \tag{5.1}$$

Here, Φ is a divergence threshold used to decide whether to generate or abstain. It can be set empirically based on divergence distributions in training data. While selective generation may reduce coverage (i.e., fewer outputs), this trade-off is justified in safety-critical domains like healthcare, where improved factual consistency is paramount.

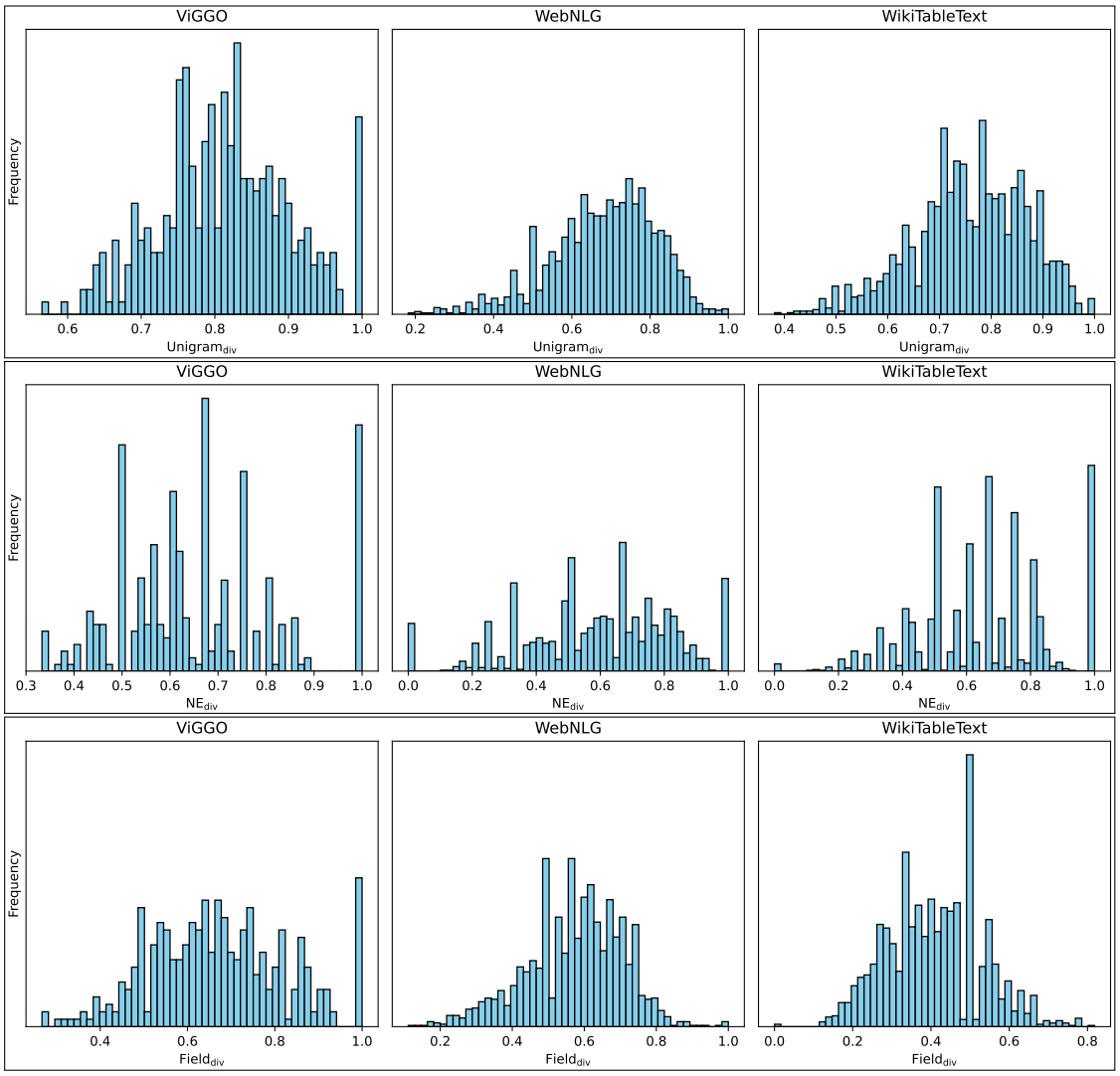


FIGURE 5.3: Distribution of source-reference divergence across the ViGGO, WikiTableText, and WebNLG datasets using our three proposed measures: unigram-level divergence, named entity-based divergence, and field-aware divergence.

TABLE 5.1: Factual consistency scores for FLAN-T5 under divergence-aware selective generation at varying coverage levels. Blue denotes favorable consistency gains; red indicates aberrant behavior.

		ViGGO				WebNLG				WikiTableText			
f_{div}	coverage	AlignScore	QAFactEval	SummaC-conv	UniEval-fact	AlignScore	QAFactEval	SummaC-conv	UniEval-fact	AlignScore	QAFactEval	SummaC-conv	UniEval-fact
NE_{div}	1.0	0.328	0.255	0.042	0.612	0.468	0.417	0.219	0.443	0.309	0.296	0.183	0.159
	0.9	0.586	0.333	0.090	0.441	0.516	0.650	0.452	0.492	0.351	0.527	0.410	0.201
	0.8	0.436	0.625	0.112	0.713	0.790	0.482	0.290	0.521	0.377	0.360	0.243	0.465
	0.7	0.708	0.652	0.360	0.724	0.803	0.727	0.540	0.769	0.617	0.603	0.482	0.462
	0.6	0.717	0.668	0.357	0.728	0.802	0.726	0.543	0.777	0.617	0.604	0.480	0.465
$Unigram_{div}$	1.0	0.328	0.255	0.272	0.612	0.468	0.647	0.219	0.443	0.309	0.296	0.183	0.159
	0.9	0.405	0.341	0.092	0.449	0.521	0.658	0.273	0.495	0.346	0.525	0.218	0.195
	0.8	0.449	0.405	0.120	0.480	0.549	0.733	0.539	0.526	0.366	0.597	0.478	0.212
	0.7	0.704	0.671	0.367	0.721	0.803	0.740	0.553	0.777	0.609	0.600	0.477	0.455
	0.6	0.711	0.678	0.368	0.723	0.816	0.744	0.563	0.788	0.611	0.605	0.477	0.457
$Field_{div}$	1.0	0.328	0.255	0.042	0.382	0.468	0.647	0.219	0.673	0.309	0.526	0.183	0.389
	0.9	0.412	0.535	0.094	0.449	0.525	0.473	0.271	0.497	0.534	0.337	0.220	0.193
	0.8	0.457	0.407	0.122	0.484	0.548	0.493	0.293	0.519	0.366	0.357	0.478	0.213
	0.7	0.717	0.672	0.368	0.736	0.800	0.737	0.541	0.766	0.607	0.597	0.483	0.451
	0.6	0.734	0.687	0.373	0.745	0.802	0.741	0.548	0.771	0.610	0.594	0.483	0.452

5.4 Experimental Details

We evaluate our source-reference divergence-aware selective generation framework across three LLM families—Qwen2.5 (1.5B), FLAN-T5 (248M), and OPT (790M)—fine-tuned using Huggingface `transformers` and QLoRA [85] with a learning rate of 2e−4 and 3 training epochs. Factual consistency is assessed using four complementary metrics: AlignScore, QAFactEval, SummaC-Conv, and UniEval-Fact, capturing diverse aspects of factual alignment. Evaluation spans three representative D2T benchmarks—WikiTableText (table-to-text) [2], ViGGO (MR-to-text) [62], and WebNLG (graph-to-text) [7]—ensuring coverage across key data formats in D2T. A possible limitation of the experimental setup is the potential risk of data contamination arising from the use of publicly available datasets in LLM evaluations. Nevertheless, considering the grounded nature of D2T tasks and the evaluation methodology adopted in this work, the influence of such contamination is expected to be minimal, as discussed in subsection 3.4.4. The effectiveness of LLM-based D2T systems can be influenced by the adopted input linearization and prompt structure. Therefore, the corresponding templates and prompting methodology used in this work are described in detail in subsection 6.5.6.

5.5 Results 1: Existence of Source-Reference Divergence in D2T

This result section depicts the presence of source-reference divergence in three widely used D2T datasets: ViGGO, WikiTableText, and WebNLG. We apply our proposed

divergence quantification methods—unigram-level divergence, named entity-based divergence, and field-aware divergence—on the training partitions of these datasets. Figure 5.3 illustrates the distribution of divergence scores for each dataset under the three approaches. The wide spread of these distributions suggests that source-reference divergence is prevalent and non-trivial across all datasets. These findings provide empirical support for the hypothesis that source-reference divergence is a meaningful phenomenon in the evaluated D2T benchmarks, suggesting that accounting for it may be beneficial during the training and evaluation of LLM-based generation systems.

TABLE 5.2: Factual consistency scores for OPT under divergence-aware selective generation at varying coverage levels. Blue denotes favorable consistency gains; red indicates aberrant behavior.

		ViGGO				WebNLG				WikiTableText			
$\int_{div}$	coverage	AlignScore	QAFactEval	SummaC-conv	UniEval-fact	AlignScore	QAFactEval	SummaC-conv	UniEval-fact	AlignScore	QAFactEval	SummaC-conv	UniEval-fact
NE_{div}	1.0	0.394	0.546	0.104	0.661	0.477	0.650	0.220	0.446	0.308	0.500	0.418	0.167
	0.9	0.462	0.384	0.155	0.678	0.521	0.649	0.260	0.678	0.351	0.503	0.418	0.209
	0.8	0.495	0.423	0.182	0.750	0.540	0.713	0.520	0.749	0.372	0.331	0.490	0.229
	0.7	0.741	0.677	0.424	0.745	0.782	0.712	0.523	0.751	0.614	0.571	0.489	0.468
	0.6	0.737	0.665	0.428	0.744	0.778	0.706	0.526	0.754	0.618	0.575	0.492	0.471
$Unigram_{div}$	1.0	0.394	0.316	0.334	0.661	0.477	0.650	0.450	0.446	0.538	0.500	0.188	0.167
	0.9	0.656	0.393	0.348	0.489	0.523	0.460	0.462	0.679	0.348	0.315	0.226	0.204
	0.8	0.499	0.441	0.427	0.755	0.788	0.723	0.540	0.516	0.372	0.582	0.487	0.225
	0.7	0.747	0.692	0.432	0.756	0.795	0.730	0.548	0.765	0.614	0.585	0.486	0.464
	0.6	0.759	0.707	0.438	0.761	0.803	0.734	0.554	0.772	0.626	0.590	0.495	0.473
$Field_{div}$	1.0	0.394	0.316	0.104	0.431	0.477	0.420	0.450	0.446	0.308	0.270	0.188	0.167
	0.9	0.656	0.581	0.155	0.677	0.525	0.465	0.269	0.485	0.353	0.320	0.230	0.207
	0.8	0.747	0.438	0.429	0.521	0.552	0.487	0.536	0.510	0.379	0.581	0.252	0.228
	0.7	0.766	0.702	0.438	0.773	0.794	0.726	0.542	0.752	0.623	0.581	0.494	0.468
	0.6	0.794	0.725	0.455	0.789	0.801	0.731	0.547	0.757	0.626	0.590	0.499	0.469

TABLE 5.3: Factual consistency scores for Qwen2.5 under divergence-aware selective generation at varying coverage levels. Blue denotes favorable consistency gains; red indicates aberrant behavior.

		ViGGO				WebNLG				WikiTableText			
$\int_{div}$	coverage	AlignScore	QAFactEval	SummaC-conv	UniEval-fact	AlignScore	QAFactEval	SummaC-conv	UniEval-fact	AlignScore	QAFactEval	SummaC-conv	UniEval-fact
NE_{div}	1.0	0.433	0.334	0.115	0.672	0.531	0.474	0.474	0.721	0.292	0.266	0.179	0.399
	0.9	0.500	0.411	0.166	0.505	0.572	0.513	0.286	0.535	0.518	0.303	0.214	0.207
	0.8	0.525	0.678	0.187	0.527	0.833	0.530	0.309	0.796	0.351	0.324	0.477	0.227
	0.7	0.782	0.692	0.434	0.776	0.833	0.770	0.553	0.796	0.588	0.557	0.475	0.462
	0.6	0.793	0.709	0.441	0.779	0.830	0.765	0.554	0.797	0.590	0.555	0.477	0.460
$Unigram_{div}$	1.0	0.433	0.334	0.115	0.442	0.531	0.474	0.474	0.491	0.292	0.496	0.409	0.169
	0.9	0.498	0.404	0.355	0.507	0.576	0.515	0.294	0.536	0.330	0.309	0.216	0.207
	0.8	0.526	0.451	0.188	0.765	0.592	0.531	0.319	0.797	0.351	0.329	0.472	0.465
	0.7	0.780	0.719	0.434	0.772	0.832	0.768	0.559	0.798	0.598	0.575	0.474	0.466
	0.6	0.790	0.732	0.439	0.777	0.835	0.765	0.572	0.808	0.601	0.579	0.475	0.469
$Field_{div}$	1.0	0.433	0.334	0.345	0.442	0.761	0.474	0.474	0.491	0.522	0.266	0.179	0.169
	0.9	0.496	0.594	0.355	0.504	0.571	0.517	0.289	0.532	0.333	0.501	0.216	0.207
	0.8	0.531	0.448	0.432	0.533	0.598	0.537	0.558	0.557	0.359	0.336	0.240	0.228
	0.7	0.795	0.713	0.442	0.788	0.836	0.778	0.556	0.796	0.598	0.578	0.477	0.464
	0.6	0.809	0.735	0.451	0.794	0.838	0.777	0.559	0.796	0.594	0.580	0.474	0.458

5.6 Results 2: Factual Consistency vs. Coverage in Divergence-aware Selective Generation

Table 5.1, Table 5.2, and Table 5.3 present the impact of divergence-aware selective generation on factual consistency across varying coverage thresholds for three LLMs: FLAN-T5, OPT, and Qwen2.5. Evaluation is conducted using four widely adopted factual

consistency metrics—AlignScore, QAFactEval, SummaC-Conv, and UniEval-Fact—to provide a multi-perspective assessment of factual consistency. As the coverage threshold is varied from 60% to 100%, the results suggest that selective generation is often associated with higher factual consistency across the evaluated datasets and models. In particular, the 70–80% coverage range (highlighted in blue) frequently exhibits comparatively high consistency scores, indicating that instances with lower levels of source–reference divergence tend to yield more factually consistent outputs. The divergence threshold Φ can be empirically estimated from the third quartile of the divergence distribution in the training data. Although slight drops in consistency are observed at higher coverage levels (90–100%, marked in red), such cases occur less frequently in the reported experiments. Overall, these results show that divergence-aware selective generation represents a promising approach within the evaluated benchmark settings. This also showing that

5.7 Conclusion

This chapter addresses an important challenge in LLM-based Data-to-Text generation—maintaining factual consistency in the presence of *source–reference divergence*. To characterize this phenomenon quantitatively, three complementary measures are introduced: *unigram-level divergence*, *named entity-based divergence*, and *field-aware divergence*, each designed to quantify the degree of misalignment between structured input data and their corresponding reference texts. Building upon these metrics, the chapter presents a *divergence-aware selective generation* framework that enables LLMs to selectively abstain from generating outputs when divergence levels are high, thereby prioritizing factual consistency. Extensive experiments conducted on three representative D2T benchmarks—ViGGO, WikiTableText, and WebNLG—using three LLM families—Qwen2.5, FLAN-T5, and OPT—suggest that the proposed framework is associated with higher factual consistency while maintaining relatively substantial coverage levels between 70% and 80% across the evaluated settings. The results suggest that the selective abstention mechanism is associated with reductions in factual inconsistencies in generated outputs and may provide a useful direction for future research on improving the reliability of D2T systems in domains where factual accuracy is important, such as healthcare and finance. Overall, the findings presented in this chapter underscore the potential value of explicitly modeling and managing source–reference divergence for improving factual consistency in LLM-based D2T systems.

Chapter 6

Exponential Scaling Factual Inconsistency of LLM in D2T

While the previous chapter focused on mitigating factual inconsistency through divergence-aware selective generation, this chapter[1] examines its scaling behavior with respect to key model and data parameters. Specifically, it investigates how the degree of factual inconsistency varies with respect to model size, computational capacity (FLOPs), and fine-tuning data size. Unlike prior studies that primarily examine generalization error using power law scaling, this chapter presents a detailed investigation into the scaling dynamics of factual inconsistency using both power law and exponential formulations. To enable a rigorous comparison, it introduces *VaCScal*, a three-stage statistical validation framework comprising predictive performance estimation, goodness-of-fit assessment, and comparative scaling analysis. Experiments conducted across six LLM families and five D2T datasets suggest that factual inconsistency in the evaluated D2T settings is often better characterized by exponential scaling than by power-law scaling with respect to model size, compute, and fine-tuning data size. Overall, this chapter reports findings that motivate reconsideration of whether power-law scaling is the most appropriate characterization of factual inconsistency in the evaluated D2T settings.

6.1 Introduction

D2T [17, 35] involves converting semi-structured data (e.g., tables, graphs) into coherent and contextually appropriate natural language text, with applications spanning dialogue

[1]The work presented in this chapter is based on our paper: Mahapatra, J., Roy, S. and Garain, U., 2025. Exponential Scaling of Factual Inconsistency in Data-to-Text Generation with Fine-Tuned LLMs. *Transactions on Machine Learning Research (TMLR)*. [120]

systems, automated journalism, and report generation. Large Language Models (LLMs) have achieved strong performance in D2T tasks, particularly in terms of coherence and informativeness [6]. However, *factual inconsistency*—where the generated text fails to reflect the information present in the source data—remains a persistent challenge [54, 93], often undermining the reliability and trustworthiness of LLM-based D2T systems (see Figure 6.1). To develop dependable models, it is essential to understand how factual inconsistency scales with key factors such as model size, computational capacity (FLOPs), and the amount of fine-tuning data. Despite the importance of this issue, existing scaling analyses in D2T have primarily concentrated on generalization errors or test perplexity—typically characterized under *power law scaling* [148, 149]—while the scaling behavior of factual inconsistency has received comparatively limited attention. Bridging this knowledge gap is important for improving our understanding of faithfulness and for developing more faithful and trustworthy D2T systems.

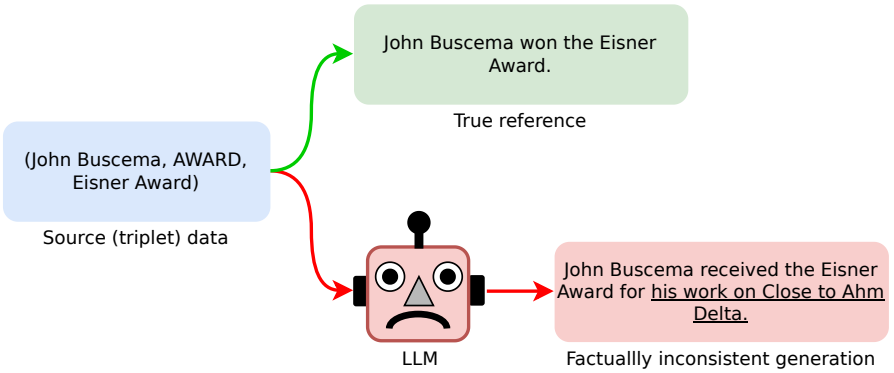


FIGURE 6.1: An example from the DART [1] dataset illustrating a factually inconsistent output generated by an LLM (Pythia-1.4B). The generated text contains unsupported content— *"his work on Close to Ahm Delta"*—which is not grounded in the source data.

This chapter addresses this research gap by systematically examining how factual inconsistency scales in LLM-based D2T. In contrast to prior work that focuses exclusively on power law scaling [148–150], the study also investigates *exponential scaling* using a statistical validation framework, *VaCScal*. The *VaCScal* framework consists of three stages—(1) predictive performance estimation, (2) goodness-of-fit assessment, and (3) comparative scaling analysis—allowing for a structured statistical comparison between power law and exponential scaling behaviors. The experimental setup spans three major LLM architectures (decoder-only, encoder–decoder, and state-space) and six LLM families—BLOOM, FLAN-T5, Mamba, OPT, Pythia, and Qwen2.5—evaluated across five D2T datasets: E2E, ViGGO, WebNLG, DART, and WikiTableText. Models are fine-tuned using QLoRA [85] and Prefix-Tuning [86], with additional experiments employing full fine-tuning to update all model parameters. For robustness, the study further reports results under both greedy and nucleus sampling decoding strategies. Factual inconsistency, defined as the inverse of factual consistency, is measured using four

widely used automatic metrics—AlignScore, QAFactEval, SummaC-Conv, and UniEval-Fact—supplemented by human evaluation. The findings, further examined using *VaCScal*, suggest that factual inconsistency is well characterized by an exponential scaling pattern under the evaluated settings. Furthermore, a mathematical rationale that offers a possible explanation for the emergence of exponential scaling behavior in factual inconsistency under the considered D2T conditions is presented, providing additional insights into the scaling dynamics of factual inconsistency in LLM-based D2T.

6.2 Related Work

D2T and factual inconsistency. D2T [35] involves generating natural language from structured inputs such as graphs [1, 7], tables [2], and meaning representations (slot-value pairs) [8, 62]. These three representation forms constitute the main categories of D2T: graph-to-text, table-to-text, and MR-to-text Recently, LLMs have become central to D2T due to their large-scale pretraining [99], scalability [72], and compatibility with efficient fine-tuning [85] and prompt-based learning [151]. These models often outperform traditional approaches in coherence and informativeness [116]. However, factual inconsistency—where generated text fails to reflect input data properly—remains a persistent challenge [145, 146], leading to hallucinations and reduced trustworthiness. Prior work attributes this to misalignment between source data and references [39, 68], LLM biases [100], and exposure bias [93]. While human evaluation is the gold standard, recent automatic metrics [145, 146] have shown strong correlation with human judgments and offer scalable alternatives.

Scaling in LLM. In LLMs, *scaling* describes how LLM performance varies with scaling factors such as model size, compute, data, etc [55, 150, 152, 153]. Hestness et al. [154] showed that deep language models follow power law scaling, which was extended by Kaplan et al. [148] to include model size, dataset size, and compute. Subsequent works have explored scaling across tasks [148, 155], paradigms such as sparse modeling [156], and parameter-efficient fine-tuning [152]. More recently, joint scaling—such as additive and multiplicative formulations—has garnered increasing attention [149, 152]. Scaling plays a critical role in hyperparameter tuning [43], training cost estimation [157], and model performance prediction [149]. A set of recent studies [158–160] has further reinforced the theoretical foundations of scaling.

6.3 Desiderata: Scaling Models

Moving beyond prior work on power law scaling [148, 150, 161], we examine two scaling models for factual inconsistency in D2T with respect to LLM size: power law scaling (modeled by a power law function) and exponential scaling (modeled by an exponential function). The two scaling models are defined as follows:

$$\text{Power Law Scaling Model} : f(x) = \begin{cases} Ax^{\alpha} + B & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

$$\text{Exponential Scaling Model} : f(x) = \begin{cases} Ce^{\beta x} + D & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

Here, A and C are case-specific parameters, α is the power law exponent, β is the exponential scaling rate, x represents the scaling factor (e.g., LLM size, compute (FLOPs), etc.), and $f(x)$ denotes the scaling objective (e.g., factual inconsistency). B and D represent the irreducible error or entropy of an LLM family [150]; in other studies, they are interpreted as the performance threshold or capacity of a given LLM family [154]. Following the significant work by Zhang et al. [152], we estimate the parameters of both models via maximum likelihood estimation on the factual inconsistency data $\mathcal{D}$, optimizing with the standard Huber loss ($\delta = 1$) due to its robustness to outliers.

6.4 VaCScal: Statistical Validation Framework

Prior empirical studies on LLM scaling often assess predictive performance using held-out loss alone. However, we find this insufficient for rigorously evaluating or comparing multiple scaling models. To address this, we use *VaCScal* (Validation and Comparison of Scaling law)—a structured three-stage framework comprising predictive performance estimation, goodness-of-fit assessment, and comparative analysis. Each stage is detailed below.

- **Stage I: Predictive Performance Estimation.** The first stage evaluates how well the scaling models generalize in terms of their predictive ability on unseen data. To achieve this, we assess the scaling models on held-out data using the Huber loss ($\delta = 1$). The Huber loss is known for its robustness to outliers, making it well-suited for reliable estimation. Given the limited data availability, we employ five-fold cross-validation to assess predictive performance.

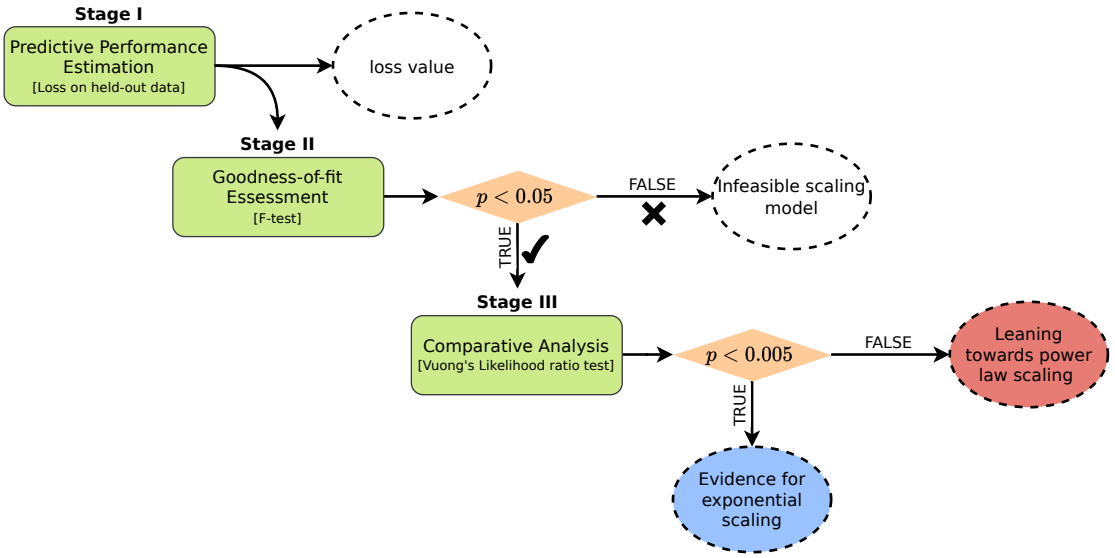


FIGURE 6.2: All three stages of *VaCScal*, the statistical validation framework for scaling models. Dashed ovals indicate outputs of each stage.

- **Stage II: Goodness-of-fit Assessment.** Predictive performance alone is not sufficient to validate a scaling model; assessing its goodness-of-fit is also important for evaluating its suitability. Therefore, in this stage, we assess the goodness-of-fit of the scaling models using an F-test for regression [162, 163]. The test statistic for the F-test is computed as follows:

$$F_{\text{stat}} = \frac{SSR_{\mathfrak{R}} - SSR_{\mathfrak{E}}}{df_{\mathfrak{R}} - df_{\mathfrak{E}}} \Big/ \frac{SSR_{\mathfrak{E}}}{df_{\mathfrak{E}}}$$
$$F_{\text{stat}} \sim \text{F-distribution}(x)$$

  Here, $SSR_{\mathfrak{R}}$ and $SSR_{\mathfrak{E}}$ represent the sum of squared residuals for the reduced and exact models, respectively. Similarly, $df_{\mathfrak{R}}$ and $df_{\mathfrak{E}}$ denote the degrees of freedom for the reduced and exact models, respectively. We consider both of our scaling models as exact models, while the reduced model is represented by a simple mean-response model. Since the F-test applies only to linear regression models, we use a log transformation to convert our scaling models into their linear forms. We perform the F-test with a significance level of $p < 0.05$, which is commonly considered a moderate threshold. If a scaling model fails to achieve a p-value below 0.05 in the goodness-of-fit assessment, we do not consider it feasible for modeling scaling behavior.

- **Stage III: Comparative analysis.** In this final stage of validation, we compare the two scaling models through hypothesis testing to determine which better explains the data. Since power law and exponential scaling models are not nested

hypotheses, the standard likelihood-ratio test is not applicable. Instead, we employ Vuong's likelihood-ratio test [164] for comparison. The test statistic for Vuong's likelihood-ratio test is computed as follows:

$$V_{\text{stat}} = \frac{\sqrt{n} \cdot \text{mean}(d)}{\sqrt{\text{Var}(d)}} \tag{6.3}$$

$$V_{\text{stat}} \sim \text{normal}(0, 1) \tag{6.4}$$

Where n represents the sample size, and d denotes the n-sized sample of the log-likelihood differences between the two scaling models. We conduct Vuong's likelihood ratio test at a stringent significance level of $p < 0.005$ to provide highly compelling evidence for our conclusion.

Normality Assumptions Verification. In Stage II and III of our validation framework, we incorporate the F-test and Vuong's likelihood-ratio test. Both tests rely, directly or indirectly, on the assumption that the residuals of our scaling models follow a normal distribution. Therefore, validating this assumption is essential. To assess the normality of residuals for both scaling models, we employ the Shapiro–Wilk test [165] (with $p < 0.05$) throughout all of our experiments.

6.5 Experimental Setup

This section presents the overall experimental setup, including details of the LLM families, D2T datasets, and the evaluation metrics used to quantify factual inconsistency. It also describes the fine-tuning and decoding strategies, training libraries, input linearisation methods, prompt structures, and other essential implementation details that ensure the consistency and reproducibility of our experiments.

6.5.1 D2T Datasets

The experimental setup incorporates five widely used D2T datasets, covering three major D2T task types: *graph-to-text*, *table-to-text*, and *meaning representation (MR)-to-text*. Specifically, DART [1] and WebNLG [7] represent graph-to-text generation, WikiTableText [2] represents table-to-text generation, and E2E [58] and ViGGO [65] correspond to MR-to-text generation. Among these datasets, DART (constructed from diverse knowledge triples), WebNLG (derived from DBpedia), and WikiTableText (based on Wikipedia tables) are *open-domain* corpora that capture broad and heterogeneous

knowledge sources. In contrast, E2E (restaurant domain) and ViGGO (video game domain) are *closed-domain* datasets characterized by consistent, domain-specific semantic structures. E2E [8, 58] is an MR-to-text dataset from the restaurant domain, consisting of 36.8K instances with an average output length of 21 words. ViGGO [65] is another MR-to-text dataset comprising approximately 6.9K instances spanning nine dialogue acts in the video game domain, with an average output length of 14 words. WikiTableText [2] is an open-domain table-to-text dataset containing around 13.3K pairs derived from Wikipedia tables, while DART [1] consists of roughly 70.5K knowledge graph triples with an average output length of 34 words. WebNLG [7], an RDF-to-text dataset, includes approximately 38.8K samples with an average output length of 30 words. Together, DART and WebNLG serve as representative open-domain benchmarks for evaluating graph-to-text generation. All datasets used in this study are sourced from [138] and [137]. A detailed summary of dataset characteristics—including domain type, D2T type, and token-level statistics for both source and reference texts—is presented in Table 6.1.

TABLE 6.1: Summary of key statistics, D2T types, and domains for the five datasets used in this study. The table reports the average text length (in tokens), number of unique tokens, and total tokens for both source data (linearized to text) and reference outputs.

Dataset	D2T Type	Domain Type	Dataset Size	Source Data			Reference Text		
				Avg. Length	Unique Tokens	Total Tokens	Avg. Length	Unique Tokens	Total Tokens
E2E	MR-to-text	Closed	36,856	27.3	125	1M	20.8	4.5K	885K
ViGGO	MR-to-text	Closed	6,900	29.9	618	206K	21.5	4.4K	148K
WikiTableText	Table-to-text	Open	13,318	35.2	29K	469K	13.9	24K	185K
DART	Graph-to-text	Open	70,524	34.8	44K	2.5M	19.3	45K	1.5M
WebNLG	Graph-to-text	Open	38,872	30.4	7K	1.2M	20.1	19K	905K

Although the publicly available datasets used in this work may introduce a potential possibility of data contamination in LLM evaluations, the impact of such effects is expected to be limited due to the grounding-oriented nature of D2T tasks and the comparative evaluation framework adopted in this study, as discussed in subsection 3.4.4.

6.5.2 LLM Families

We evaluate six prominent families of Large Language Models (LLMs), representing three distinct architectural paradigms: *decoder-only*, *encoder–decoder*, and *state-space models*. These include BLOOM [72], FLAN-T5 [55], OPT [99], Pythia [166], Mamba [56], and Qwen2.5 [33]. All models and their variants are sourced from the Hugging Face hub [137]. A summary of these LLM families, along with their architectural paradigms and model scales, is provided in Table 6.2.

TABLE 6.2: Overview of the six LLM families, their architectural paradigms, number of models, and corresponding parameter sizes (M = million, B = billion).

Family	Model Paradigm	Model Count	Parameter Sizes
BLOOM	Decoder-only	5	560M, 1.1B, 1.7B, 3B, 7B
FLAN-T5	Encoder–decoder	5	77M, 248M, 783M, 2.85B, 11.3B
Mamba	State-space model	5	130M, 370M, 790M, 1.4B, 2.8B
OPT	Decoder-only	6	125M, 350M, 1.3B, 2.7B, 6.7B, 13B
Pythia	Decoder-only	8	70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, 12B
Qwen2.5	Decoder-only	7	0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B

Decoder-only Models. BLOOM [72] is a multilingual decoder-only language model trained on the ROOT dataset, designed to support open-access research across multiple languages. Five BLOOM variants are included in our experiments, ranging from 560M to 7B parameters. OPT [99] is another decoder-only model family, trained on a diverse corpus combining Reddit, the Pile [167], and RoBERTa [168] pretraining datasets, following the design of Brown et al. [76]. We evaluate six OPT models, spanning from 125M to 13B parameters. Pythia [166] comprises eight GPT-style autoregressive models (70M–12B parameters) trained on the Pile dataset using a consistent data ordering scheme and optimized with flash attention for efficient training and inference. Qwen2.5 [33] is a widely regarded decoder-only Transformer [30] model family. We include seven Qwen2.5 variants ranging from 0.5B to 72B parameters. Pretrained on over 18 trillion tokens and post-trained with more than one million supervised examples, Qwen2.5 exhibits strong reasoning, mathematical, and commonsense capabilities. Its exposure to structured data formats such as tables and JSON suggests strong potential for D2T tasks, making it particularly valuable for analyzing scaling behavior in factual consistency.

Encoder–Decoder Models. FLAN-T5 [55] represents the encoder–decoder paradigm, extending the original T5 architecture with instruction tuning and large-scale task generalization. Unlike decoder-only architectures, FLAN-T5 processes prompts through an encoder to enhance contextual understanding before decoding. We evaluate five models from this family, ranging from 77M to 11.3B parameters.

State-Space Models. Mamba [56] represents a recent paradigm shift in sequence modeling, employing state-space mechanisms as an efficient alternative to attention-based architectures. State-space models capture long-range dependencies while maintaining linear-time inference complexity. We evaluate five Mamba variants ranging from 130M to 2.8B parameters, providing insights into the scalability and consistency of non-transformer architectures in D2T.

6.5.3 Fine-tuning and Decoding Strategies

All Large Language Models (LLMs) are fine-tuned independently on each of the five D2T datasets using two parameter-efficient fine-tuning techniques: QLoRA [85] (Quantized Low-Rank Adapter) and Prefix-Tuning [86]. In the QLoRA configuration, a reduced rank of $r = 16$ is applied primarily to the attention and feedforward modules of the LLMs to balance computational efficiency and adaptation performance. For Prefix-Tuning, a virtual prefix length of 32 tokens is used, consistent with prior findings that highlight its effectiveness as a soft prompt-tuning approach. Prefix-Tuning is applied to all evaluated LLM families except Mamba, where the state-space architecture exhibits unstable initialization for virtual token embeddings. For both fine-tuning methods, the learning rate for trainable parameters is fixed at 1.0×10^{-4}. To further investigate scaling effects, full fine-tuning—where all trainable parameters of the model are updated—is also performed, using a learning rate of 5.0×10^{-5}. This setup enables the analysis of both efficient adaptation and complete model optimization under comparable training conditions. Given the pivotal role of decoding strategies in D2T, two decoding paradigms are explored: deterministic (*greedy decoding*) and stochastic (*nucleus sampling*). For nucleus sampling, the nucleus size is set to 0.95, ensuring a balance between diversity and factual accuracy. Unless otherwise specified, most results presented in this chapter are based on outputs generated through greedy decoding, which facilitates consistent and reproducible evaluation across all models and datasets.

6.5.4 Quantification of Factual Inconsistency

Factual inconsistency is defined as the inverse of factual consistency and is computed as $(1 - z)$, where z denotes the factual consistency score ranging from 0 to 1. To quantify factual inconsistency in LLM-based D2T, this study employs four widely used automatic evaluation metrics, each suggesting strong correlation with human judgments. AlignScore measures factual alignment between the generated text and the source data through information alignment techniques [146]. QAFactEval assesses factual consistency by generating and answering questions derived from the source and generated text [145]. SummaC-Conv evaluates factual consistency using natural language inference (NLI)-based methods [96], while UniEval-Fact applies a unified, multi-task evaluation framework trained across various factual consistency tasks [169]. In addition to these automatic metrics, human evaluation is conducted to provide qualitative validation, where factual consistency is manually annotated by human experts to complement and verify the automatic assessments.

6.5.5 Training Settings and Libraries

All experiments are conducted using the `SciPy` library [170] to train both scaling models—power law and exponential—based on the observed factual inconsistency results. For each fitted scaling model, the margin of error (MoE) is computed on the residual deviation with a 95% confidence interval, providing an estimate of the uncertainty associated with model fitting. The same `SciPy` library is also utilized for all hypothesis testing procedures, including the implementation of Vuong's likelihood-ratio test in Python. For fine-tuning and model quantization across all LLM families, the experiments employ the `transformers` library [137] from Hugging Face. Fine-tuning and evaluation are performed separately on each D2T dataset using their standard training and testing splits to ensure fair and consistent comparisons. The experimental setup includes one NVIDIA A6000 GPU (48 GB) and one NVIDIA A100 GPU (80 GB). The cumulative computational time across both devices is approximately 22.5 days. Throughout all training and evaluation phases, no catastrophic failures or system instabilities were encountered, suggesting the robustness of the experimental pipeline.

6.5.6 Input Linearisation and Prompt Structure

Input linearisation and prompt structure are essential components of D2T, influencing both language model fine-tuning and inference.

Input Linearisation. Linearisation refers to the process of transforming semi-structured input data into a sequential format suitable for LLM-based generation. As discussed earlier, we consider three types of D2T tasks: MR-to-text (E2E and ViGGO), table-to-text (WikiTableText), and graph-to-text (DART and WebNLG). For MR-to-text and table-to-text generation, inputs are linearised as sequences of slot–value pairs in the following format:

```
[(SlotName1, Value1), (SlotName2, Value2), ...,
(SlotNameN, ValueN)]
```

For graph-to-text generation, inputs are linearised as lists of relational triplets in the following format:

```
[(EntityA1, Relation1, EntityB1), (EntityA2,
Relation2, EntityB2), ..., (EntityAN, RelationN,
EntityBN)]
```

Prompt Structure. Prompt design plays an important role in both supervised fine-tuning (and other post-training methods) and inference for D2T tasks. For our experiments, we use the following unified prompt template for both supervised fine-tuning [33, 99] and inference:

```
Instruction: Please generate a textual description
('Output') from the provided structured data
('Input').
Input: [linearised input data]
Output: [reference text]
```

6.6 Results

This section presents the plots and *VaCScal* results for factual inconsistency scaling (measure using AlignScore and QAFactEval) in D2T with respect to two scaling factor model size and compute (FLOPs), based on two scaling models—power law and exponential. All results in this study are based on text generated using greedy decoding.

6.6.1 Scaling Behavior Based on Model Size of Fine-Tuned (QLoRA) LLM Families

Here, we present the scaling behavior of factual inconsistency in D2T with respect to model size for LLM families fine-tuned using QLoRA. Here, we present results for four LLM families—FLAN-T5, Mamba, Pythia, and Qwen2.5. Results for the other two families, BLOOM and OPT, are provided in paragraph* A.2. For results using SummaC-conv and UniEval-fact as factual inconsistency metrics, please refer to Appendix A.3.

6.6.1.1 Factual Inconsistency Measured via AlignScore

Figure 6.3 shows the fitted curves for the power law and exponential scaling models, along with 95% confidence intervals derived from residual error. Anomalies are particularly observed in the E2E dataset for the Mamba family, where factual inconsistency unexpectedly increases with model size. A possible hypothesis for this aberrant behavior is discussed in Appendix A.8. In sporadic exceptional cases involving the E2E and ViGGO datasets, both scaling models exhibit higher margins of error (MoE), although the exponential model consistently shows lower MoE than the power law. Outside these

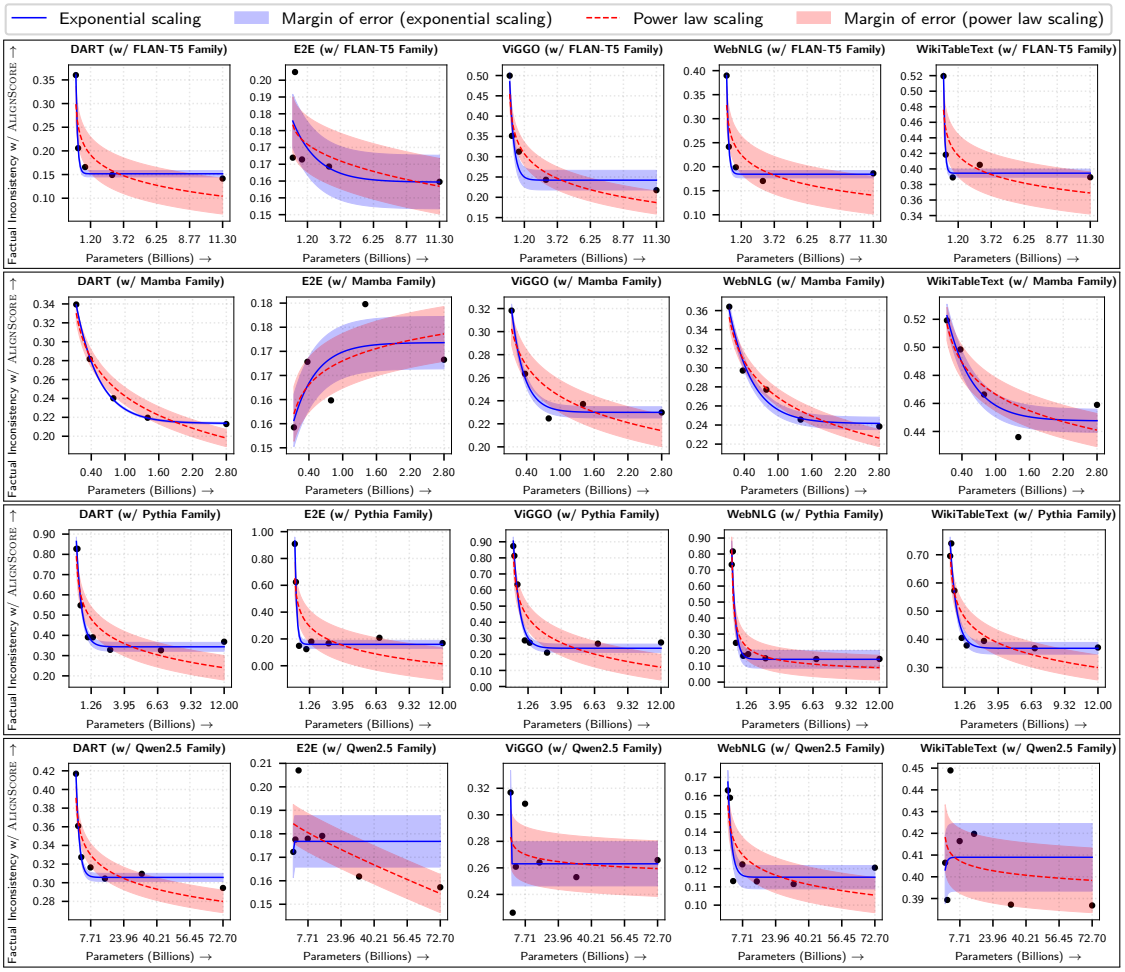


FIGURE 6.3: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by ALIGNSCORE, plotted against model size across different LLM families.

exceptions, exponential scaling generally yields much narrower MoEs, indicating more stable and reliable fits in terms of uncertainty. Table 6.3 summarizes the outcomes of the *VaCScal* framework. Exponential scaling consistently outperforms the power law across all three validation stages: in Stage I, it achieves lower held-out loss; in Stage II, it passes the F-test more frequently (✓); and in Stage III, Vuong's test favors exponential scaling in most cases (highlighted in blue). In several scenarios, power law scaling is found to be infeasible. Although in multiple cases for Qwen2.5 both scaling models fail to pass the Stage II feasibility test of *VaCScal* (i.e., meeting the minimum p-value for goodness-of-fit), exponential scaling still achieves lower or comparable held-out loss (Stage I) in most scenarios. Despite a few exceptions (e.g., E2E), *VaCScal* results indicate that exponential scaling more reliably models factual inconsistency—measured via ALIGNSCORE—in D2T.

Takeaway. *Exponential scaling models better capture factual inconsistency (measured via* AlignScore*) with respect to LLM model size in D2T, exhibiting lower error and stronger statistical fit compared to power law scaling, despite a few outliers.*

Table 6.3: Results from *VaCScal* corresponding to Figure 6.3. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
FLAN-T5	Exponential	4.15E-04/✓	8.60E-04/✗	1.14E-03/✓	2.16E-03/✓	2.56E-03/✓
	Power Law	1.47E-04/✓	3.11E-03/✗	7.80E-01/✗	6.16E-04/✓	1.45E-04/✓
Mamba	Exponential	2.14E-05/✓	1.43E-04/✗	4.40E+01/✓	1.05E-04/✓	3.09E-04/✗
	Power Law	3.77E+05/✗	1.99E+03/✗	1.08E+167/✗	3.50E-04/✗	1.42E+13/✗
Pythia	Exponential	2.03E-03/✓	1.71E+01/✓	2.48E-03/✓	2.30E-02/✓	3.31E-03/✓
	Power Law	1.68E+11/✗	1.24E-02/✓	1.92E-02/✗	8.71E+11/✗	1.96E+04/✗
Qwen2.5	Exponential	4.93E-05/✓	2.30E-04/✗	3.98E-02/✗	7.73E+01/✓	3.41E-04/✗
	Power Law	4.79E-01/✓	2.00E-03/✗	6.81E-04/✗	2.71E-04/✗	2.59E-03/✗

6.6.1.2 Factual Inconsistency Measured via QAFactEval

Figure 6.4 shows the fitted power law and exponential scaling curves, along with 95% MoE, using QAFactEval as the inconsistency metric. Similar to AlignScore, notable anomalies are observed—especially for the E2E dataset, where all LLM families (except Pythia) exhibit increasing inconsistency with model size. Aside from a few exceptions in E2E and ViGGO, where both scaling models show higher MoE, exponential scaling generally yields narrower MoE than power law scaling, indicating more stable fits with lower uncertainty. Table 6.4 summarizes the corresponding results under the *VaCScal* framework. In most Stage I results, exponential models consistently achieve lower or comparable held-out loss. For Stage II, they pass the F-test (✓) more reliably than power law models, which often fail (✗). Overall, Stage III also favors exponential scaling, with numerous significant Vuong test results (highlighted in blue). In two rare cases—DART and WebNLG with the FLAN-T5 family—power law scaling outperforms exponential models. We hypothesize that FLAN-T5, having been pre-trained on graph-to-text formats, may exhibit divergence from exponential scaling in these cases due to prior alignment with power law behavior. In Qwen2.5, apart from ViGGO, all other datasets show significantly lower held-out loss (Stage I) for exponential scaling. Overall, aside from these few rare exceptions, exponential scaling emerges as a more consistent and statistically supported characterization of factual inconsistency in D2T under the evaluated settings, when measured with QAFactEval.

Takeaway. *Compared to power law scaling, exponential scaling more effectively captures the relationship between factual inconsistency (measured via* QAFactEval*) and*

model size in D2T across most LLM families, despite some anomalies (e.g., E2E, ViGGO).

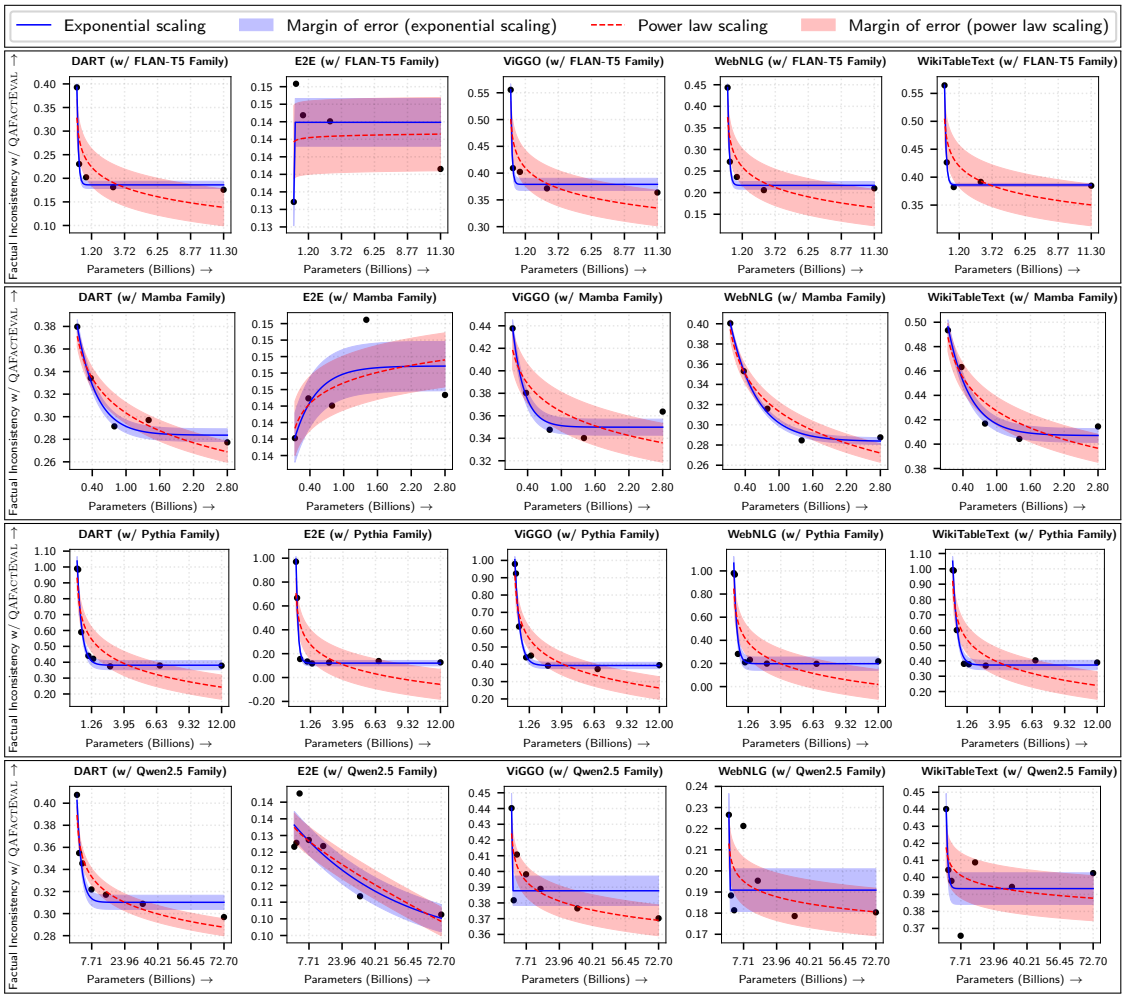


FIGURE 6.4: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFACTEVAL, plotted against model size across different LLM families.

6.6.2 Scaling Behaviour Based on FLOPs of Fine-Tuned (QLoRA) LLM Families

Estimating compute budget is central to LLM scaling research [55, 148], with several early works linking generalization error to FLOPs (floating-point operations). This section examines the scaling behavior of LLM factual inconsistency in D2T with respect to FLOPs. We evaluate LLM families across five D2T datasets using two automatic metrics—ALIGNSCORE and QAFACTEVAL. Here, we present results for four LLM

TABLE 6.4: Results from *VaCScal* corresponding to Figure 6.4. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
FLAN-T5	Exponential	3.37E-03/✓	6.59E-05/✗	4.64E-03/✓	2.46E-03/✓	3.22E+00/✓
	Power Law	1.22E+02/✓	7.08E-04/✗	1.30E-03/✓	2.57E-03/✓	4.57E-04/✓
Mamba	Exponential	3.74E-04/✓	3.75E-05/✗	1.52E+01/✓	4.38E-05/✓	5.68E-05/✓
	Power Law	4.01E-05/✗	2.17E-05/✗	4.43E-04/✗	2.63E+05/✗	9.08E+20/✗
Pythia	Exponential	3.97E-03/✓	7.47E-02/✓	3.65E-03/✓	4.54E-02/✓	3.24E-01/✓
	Power Law	1.04E+06/✗	2.86E-02/✓	2.09E+06/✓	9.44E+14/✗	1.73E-02/✗
Qwen2.5	Exponential	3.26E-04/✓	7.24E-05/✓	4.64E-04/✗	2.06E-04/✗	2.88E-04/✗
	Power Law	1.06E-01/✓	2.14E-03/✗	3.63E-02/✓	2.91E-04/✗	6.27E+16/✗

families—FLAN-T5, Mamba, Pythia, and Qwen2.5. Results for the other two families, BLOOM and OPT, are provided in Appendix A.2. FLOPs are computed using `calflops`[2], and all models are fine-tuned using QLoRA.

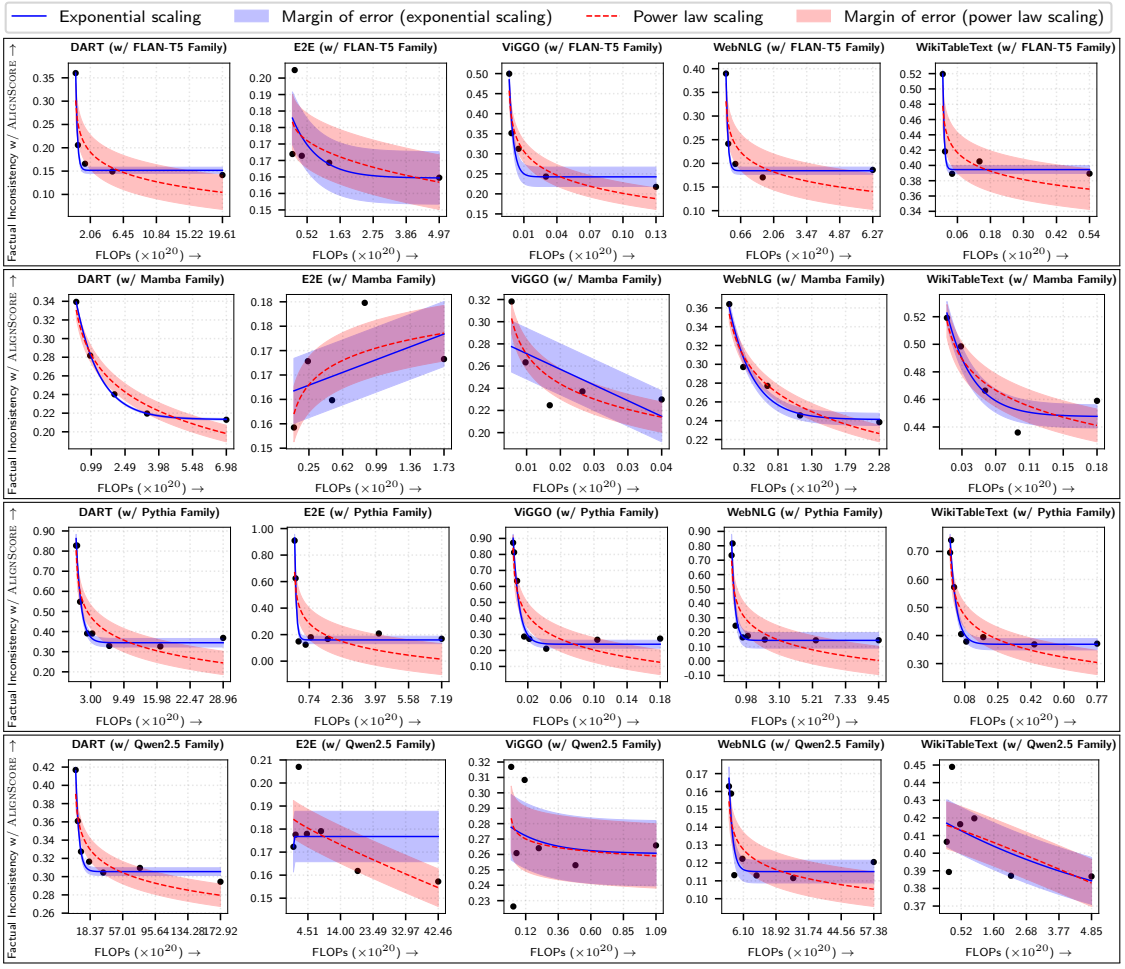


FIGURE 6.5: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by AlignScore, plotted against FLOPs across different LLM families.

[2] https://github.com/MrYxJ/calculate-flops.pytorch

TABLE 6.5: Results from *VaCScal* corresponding to Figure 6.5. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
FLAN-T5	Exponential	1.52E-03/✓	4.62E-04/✗	3.28E-03/✗	4.48E-04/✓	3.88E-03/✓
	Power Law	6.14E-05/✓	6.15E-03/✗	3.47E-01/✗	2.36E-03/✓	1.33E-04/✓
Mamba	Exponential	1.62E-05/✓	2.36E-04/✗	8.04E+01/✓	3.92E-04/✓	4.56E-04/✗
	Power Law	4.51E+05/✗	1.47E-03/✗	5.07E+176/✗	1.13E+00/✗	4.44E+13/✗
Pythia	Exponential	3.82E-03/✓	3.22E+01/✓	1.24E-03/✓	4.92E-02/✓	1.03E-03/✓
	Power Law	1.95E+11/✗	2.56E-02/✓	2.48E+03/✗	1.77E+11/✗	1.37E+03/✗
Qwen2.5	Exponential	9.97E-05/✓	1.76E-04/✗	3.71E-04/✗	1.03E+02/✓	5.53E-04/✗
	Power Law	1.55E+03/✓	1.49E-04/✗	4.51E-03/✗	2.18E-03/✗	4.13E-04/✗

6.6.2.1 Factual Inconsistency Measured via AlignScore

Figure 6.5 illustrates how factual inconsistency—measured via ALIGNSCORE—scales with FLOPs across LLM families under both power law and exponential models, along with 95% confidence margins. As with the model size results, similar anomalies persist in FLOP-based scaling—particularly in the E2E dataset (notably with the Mamba family), where inconsistency increases with FLOPs. Similar to the model size-based findings, Appendix A.8 provides a plausible explanation for this aberrant behavior. In Figure 6.5, E2E also exhibits higher MoE across both scaling models and most LLM families. Outside these exceptions, exponential scaling consistently yields narrower MoEs, indicating more stable fits with lower residual uncertainty. Table 6.5 summarizes the *VaCScal* evaluation for this FLOP-based analysis. In Stage I, exponential models consistently achieve lower held-out loss. While Stage II includes a few infeasible cases (e.g., E2E and ViGGO with Qwen2.5), exponential scaling is statistically preferred in most model–dataset pairs, as indicated by the frequent blue marks showing significant Vuong test outcomes in Stage III. Here, Qwen2.5 shows conclusions similar to its model size–based scaling behavior for factual inconsistency measured via ALIGNSCORE. Overall, the *VaCScal* results indicate that exponential scaling provides a more robust and reliable model for capturing how factual inconsistency in D2T decreases with increasing compute—offering actionable insights for compute-budget planning with fine-tuned LLMs.

Takeaway. *Exponential scaling model better captures how factual inconsistency (via* ALIGNSCORE*) decreases with FLOPs across LLM families, offering a more stable and practical scaling model than power law scaling.*

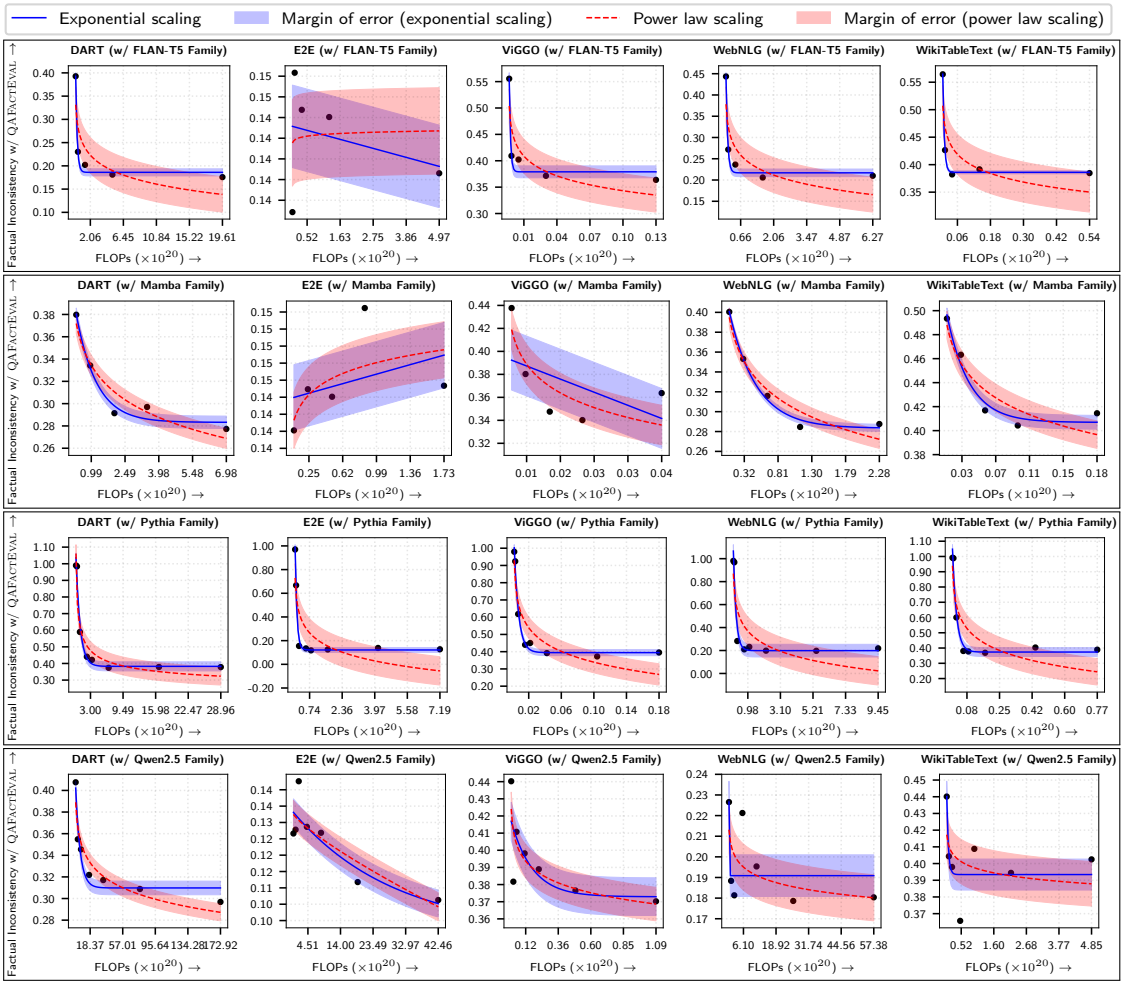


FIGURE 6.6: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFACTEVAL, plotted against FLOPs across different LLM families.

6.6.2.2 Factual Inconsistency Measured via QAFactEval

Figure 6.6 displays fitted power law and exponential scaling curves with 95% confidence intervals (MoE), showing how factual inconsistency—measured using QAFACTEVAL—varies with FLOPs. Again, certain anomalies persist; for instance, factual inconsistency increases with compute for the E2E dataset under the FLAN-T5 and Mamba families. The E2E dataset (within FLAN-T5 and Mamba families) often exhibits large MoEs, indicating high uncertainty. Nevertheless, exponential scaling consistently delivers narrower confidence bounds than power law fits, suggesting more reliable and stable modeling. The associated *VaCScal* outcomes, summarized in Table 6.6, reinforce this conclusion: most model–dataset pairs show a clear preference for exponential scaling, as indicated by the frequent blue-highlighted Vuong test results, with only a few exceptions (in DART and WebNLG with FLAN-T5). While a few configurations—such as E2E—are deemed infeasible in Stage II, the broader pattern of *VaCScal* results remains

consistent with exponential scaling as a more suitable framework for characterizing how factual inconsistency declines under the evaluated settings.

TABLE 6.6: Results from *VaCScal* corresponding to Figure 6.6. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
FLAN-T5	Exponential	2.37E-03/✓	6.81E-05/✗	4.00E-03/✓	2.34E-04/✓	1.64E-04/✓
	Power Law	2.24E+02/✓	5.52E-06/✗	3.73E-03/✓	6.87E-03/✓	3.64E-04/✓
Mamba	Exponential	8.12E-05/✓	1.76E-06/✗	1.96E-04/✓	3.72E-05/✓	1.10E-03/✓
	Power Law	8.31E+08/✗	2.60E-04/✗	7.56E-04/✗	7.46E+05/✗	2.66E+21/✗
Pythia	Exponential	2.81E-03/✓	2.32E-02/✓	9.32E-04/✓	3.67E-01/✓	7.35E+01/✓
	Power Law	5.64E+06/✗	5.29E-02/✓	9.25E+03/✓	1.01E+15/✗	4.59E-02/✗
Qwen2.5	Exponential	3.80E-04/✓	7.23E-05/✓	3.41E-04/✗	1.91E-04/✗	5.13E-04/✗
	Power Law	6.40E-01/✓	1.58E-02/✗	1.33E-01/✓	1.76E-04/✗	4.90E-04/✗

Takeaway. *Compared to power law scaling, the exponential model more effectively captures the relationship between FLOPs and factual inconsistency (measured via* QAFACTEVAL*), exhibiting greater stability across LLM families.*

6.7 Discussion

The findings presented in this chapter consistently suggest that *exponential scaling* provides a statistically grounded explanation of *factual inconsistency* in D2T compared to power law scaling. This conclusion holds robustly across multiple evaluation metrics, scaling factors, and experimental configurations. In the main results (section 6.6), exponential scaling outperforms power law scaling in modeling factual inconsistency, as measured by ALIGNSCORE and QAFACTEVAL, with respect to both model size and computational budget (FLOPs). This trend remains consistent when alternative factual consistency metrics—such as SUMMAC-CONV and UNIEVAL-FACT (Appendix A.3)—are employed, and when the results are validated against human annotations (Appendix A.1), further supporting the consistency of the observed scaling behavior. Moreover, exponential scaling continues to dominate when additional scaling factors are examined, including the size of the fine-tuning dataset (Appendix A.6).

To verify that these findings are not contingent upon a specific fine-tuning or decoding approach, supplementary experiments are conducted using Prefix-Tuning (Appendix A.5) and nucleus sampling (Appendix A.4). Furthermore, to provide a thorough understanding of scaling behavior in factual inconsistency, full fine-tuning experiments—where all model parameters are updated—are also performed (Appendix A.7).

Across all experimental conditions, exponential scaling consistently suggests strong performance over power law scaling. While a few exceptions are observed, particularly in the E2E and ViGGO datasets, these anomalies are limited in scope and do not undermine the broader empirical trend. It is hypothesized that the atypical behavior of these datasets arises from their low lexical diversity and closed-domain nature, which increase susceptibility to overfitting in larger LLMs. In contrast, open-domain datasets such as DART, WebNLG, and WikiTableText exhibit greater linguistic and semantic diversity, mitigating such effects. A detailed analysis of these exceptions is provided in Appendix A.8.

A key enabler of this analysis is the proposed *VaCScal* framework, which enables statistically rigorous validation of scaling models beyond traditional indicators such as margin of error (MoE) or held-out loss. While MoE provides an intuitive measure of uncertainty, it does not assess a model's statistical plausibility. In several instances, models exhibiting low MoE were found to fail Stage II of *VaCScal* due to high p-values, suggesting that visual fit alone is insufficient for establishing statistical validity. Stage III of *VaCScal* offers a principled method for comparing competing scaling models, frequently providing evidence favoring exponential scaling when both models satisfy Stage II feasibility criteria. Additionally, Stage I analysis reveals that exponential scaling often achieves lower held-out loss even in cases where neither model passes feasibility checks, highlighting its robustness in practical applications.

In summary, the broad evaluation presented in this chapter—spanning multiple metrics, datasets, scaling factors, and validation strategies—provides empirical evidence that exponential scaling offers a better fit than power-law scaling for modeling factual inconsistency within the evaluated fine-tuned LLM-based D2T systems.

6.8 A Conjecture on the Exponential Scaling of Factual Inconsistency in D2T

Building on the empirical evidence observed in our experiments and the presence of source–reference divergence in D2T datasets [34, 39]—that is, discrepancies between the structured input data and the corresponding reference text in the training corpus—we formulate the following conjecture on the scaling behavior of factual inconsistency with model size. We then provide a mathematical rationale for this conjecture under a set of explicit assumptions.

Conjecture 6.1. *Consider a D2T dataset exhibiting source–reference divergence. Under the assumptions described in the following mathematical rationale, we conjecture that*

factual inconsistency in LLM-based D2T decreases exponentially with model size N, such that

$$\textit{Factual Inconsistency} \propto e^{-\xi(N)}, \tag{6.5}$$

where $\xi(N)$ is a model-dependent function that is linear or sub-linear in the model (LLM) size N.

Mathematical Rationale for Conjecture 6.1. To provide a mathematical rationale for the conjectured exponential scaling of factual inconsistency with LLM size, we consider the role of *source–reference divergence*, defined as the discrepancy between the structured input data and the corresponding reference texts used during training or fine-tuning. Such divergence may arise from heuristic data collection procedures, annotation noise, or the incorporation of information that is not explicitly grounded in the source data. Under this setting, minimizing the conventional autoregressive training objective may encourage the model to assign probability mass to outputs that deviate from the factual content of the input. Consequently, standard perplexity, which is computed with respect to the potentially divergent reference text, may not accurately reflect the model's factual consistency with the source.

To facilitate the analysis, we therefore introduce a relative perplexity measure $\mathcal{R}$ that compares the likelihood assigned to a reference sequence against that assigned to the model's most likely output. Unlike conventional perplexity, which primarily measures the likelihood of a given sequence, this relative measure captures the model's preference between a reference sequence and its most likely alternative. In particular, when the reference sequence is constructed from information directly grounded in the input, $\mathcal{R}$ provides a proxy for the extent to which the model prefers a potentially inconsistent output over a source-grounded alternative.

Formally, given an input x, a reference sequence $y^{\text{ref}} = y_1^{\text{ref}} y_2^{\text{ref}} \dots y_L^{\text{ref}}$, and the model's most likely output sequence $y^{\max} = y_1^{\max} y_2^{\max} \dots y_L^{\max}$, we define $\mathcal{R}$ as the ratio between their corresponding perplexities (Equation 6.6):

$$\mathcal{R}(x, y^{\text{ref}}) = \frac{\text{perplexity}(y^{\text{ref}})}{\text{perplexity}(y^{\max})}\bigg|_{\text{input}=x} = \left(\frac{\prod_{i=1}^{L} \Pr(y_i^{\text{ref}})}{\prod_{i=1}^{L} \Pr(y_i^{\max})} \right)^{-\frac{1}{L}} \Bigg|_{\text{input}=x} \tag{6.6}$$

By construction, the relative perplexity measure $\mathcal{R}$ takes values in the range $[1, \infty)$, with $\mathcal{R} = 1$ corresponding to the limiting case in which the model assigns maximal likelihood to the reference relative to its generated alternative. Under the proposed interpretation of $\mathcal{R}$ as a proxy for factual inconsistency, values close to unity indicate a strong alignment between the model's likelihood assignments and the reference, whereas larger values of $\mathcal{R}$ suggest an increasing preference for outputs that deviate from the reference. Consequently, $\mathcal{R}$ can be viewed as a relative measure of the discrepancy between factual and model-preferred outputs, rather than an exact measure of factual inconsistency itself. In practical settings, the effective range of $\mathcal{R}$ may be moderated by calibration effects and the bounded variability of model logits.

To further analyze the behavior of the relative perplexity measure in Equation 6.6, we consider its dependence on the underlying token probabilities produced by the model. Since these probabilities are obtained by applying the softmax function to the corresponding logits at inference time, the behavior of $\mathcal{R}$ can be examined through the properties of the associated logit values. Let z^{ref} and $z^{\max}$ denote the logits assigned to the tokens in the reference sequence y^{ref} and the model's most likely output sequence $y^{\max}$, respectively. Expressing $\mathcal{R}$ in terms of these logits provides a convenient framework for analyzing how deviations between factual and model-preferred outputs influence the proposed proxy measure of factual inconsistency.

$$\begin{aligned}\mathcal{R}(x, y^{\text{ref}}) = \left(\frac{\prod_{i=1}^{L} \frac{e^{z_i^{\text{ref}}}}{\sum_j e^{z_{i,j}}}}{\prod_{i=1}^{L} \frac{e^{z_i^{\max}}}{\sum_j e^{z_{i,j}}}} \right)^{-\frac{1}{L}} \Bigg|_{\text{input}=x} &\underset{(1)}{=} \left(\frac{\prod_{i=1}^{L} e^{z_i^{\text{ref}}}}{\prod_{i=1}^{L} e^{z_i^{\max}}} \right)^{-\frac{1}{L}} \Bigg|_{\text{input}=x} \\ &\underset{(2)}{=} \left(\prod_{i=1}^{L} e^{z_i^{\text{ref}} - z_i^{\max}} \right)^{-\frac{1}{L}} \Bigg|_{\text{input}=x}\end{aligned} \tag{6.7}$$

The simplification in step (1) of Equation 6.7 follows from the fact that the softmax normalization terms appearing in both the numerator and denominator are equal and therefore cancel. It is important to note that z_i^{ref} and $z_i^{\max}$ do not represent the i-th vocabulary logits; rather, they denote the logits assigned to the i-th tokens of the sequences y^{ref} and $y^{\max}$, respectively. Under this formulation, the relative perplexity measure $\mathcal{R}$ depends on the aggregate differences between the logits associated with the reference and those associated with the model's most likely output. Consequently, the behavior of $\mathcal{R}$ can be analyzed through the statistical properties of these logit differences. In settings where source-reference divergence is present, discrepancies between the logits

assigned to reference tokens and model-preferred tokens may arise, motivating a closer examination of their magnitude and scaling behavior. We therefore proceed to analyze the properties of these logit deviations and their implications for the proposed proxy measure of factual inconsistency.

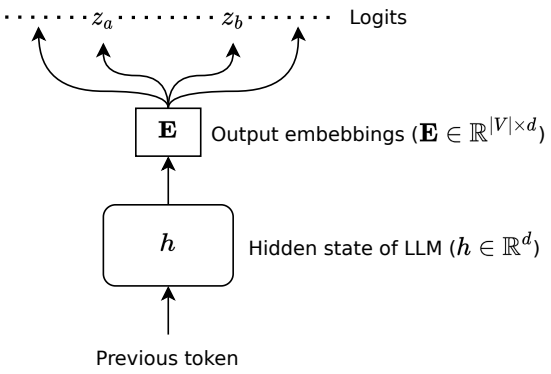


FIGURE 6.7: A simplified systematic view of the logits output in a LLM.

For analytical convenience, let $z_a = z_i^{\text{ref}}$ and $z_b = z_i^{\max}$, where a and b denote the vocabulary indices corresponding to the i-th tokens in the reference and model-generated sequences, respectively. To study the statistical behavior of these logits, we adopt a simplified probabilistic model of logit generation. Specifically, at a given decoding step, the logit vector is represented as $z = \mathbf{E}h$, where $\mathbf{E} \in \mathbb{R}^{|V| \times d}$ denotes the output embedding matrix and $h \in \mathbb{R}^d$ denotes the hidden representation of the model. Let d be the model dimension. In accordance with standard mean-field assumptions in high-dimensional neural network theory, we consider an idealized normalized regime in which the entries of $\mathbf{E}$ and h are independent random variables with zero mean and variances σ_E^2 and σ_h^2, respectively; that is, $\mathbf{E}_{i,j} \sim \mathcal{N}(0, \sigma_E^2)$ and $h_j \sim \mathcal{N}(0, \sigma_h^2)$. While these assumptions do not aim to characterize the exact behavior of trained LLMs, they provide a tractable framework for examining how logit deviations may scale with model size. Under this probabilistic model, the logit associated with the k-th vocabulary token is given by Equation 6.8.

$$z_k = \sum_{j=1}^{d} \mathbf{E}_{k,j} h_j \tag{6.8}$$

Under the probabilistic assumptions introduced above, the variance of an individual logit can be derived directly from the representation of z_k in Equation 6.8 as follows:

$$\text{var}[z_k] = d \cdot \sigma_h^2 \cdot \sigma_E^2 \tag{6.9}$$

$$= d \cdot \sigma^2 \qquad \text{Assuming } \sigma^2 = \sigma_h^2 \cdot \sigma_E^2 \tag{6.10}$$

Prior work on neural scaling laws suggests that the model dimension d grows with model size N according to a function ϕ, typically exhibiting linear or sub-linear behavior [159, 173]. Motivated by this relationship, we examine how the magnitude of the logit deviation $|z_a - z_b|$ may evolve as a function of model size. To characterize the statistical behavior of this deviation under the proposed probabilistic model, we employ Chebyshev's inequality to bound the probability of large departures from its mean. This provides a tractable mechanism for relating the distribution of logit differences to model scaling and, consequently, to the behavior of the relative perplexity measure introduced as a proxy for factual inconsistency.

$$\Pr\left(|z_a - z_b| \geq \xi(N)\right) \leq \frac{\text{Var}(z_a - z_b)}{\xi(N)^2} \tag{6.11}$$

In Equation 6.11, $\xi(N)$ denotes a model-dependent function of the model size N that characterizes the aggregate scale of logit deviations under the proposed probabilistic framework and is assumed to exhibit linear or sub-linear growth. To evaluate the resulting probability bound, it is necessary to characterize the variance of the logit difference $|z_a - z_b|$. We therefore consider the variance of the difference between the two random variables and employ the standard variance decomposition formula, which for independent variables yields:

$$\text{Var}(z_a - z_b) = \text{Var}(z_a) + \text{Var}(z_b) - 2\,\text{Cov}(z_a, z_b) \tag{6.12}$$

Using the logit representation introduced above, the logits corresponding to the reference and model-preferred tokens can be written as $z_a = \mathbf{E}_{a,:}^\top \mathbf{h}$ and $z_b = \mathbf{E}_{b,:}^\top \mathbf{h}$, where $\mathbf{E}_{a,:}$ and $\mathbf{E}_{b,:}$ denote the rows of the output embedding matrix associated with vocabulary entries a and b, respectively. This representation makes it possible to analyze the statistical relationship between the two logits through the properties of the embedding matrix and hidden representation. In particular, the covariance between z_a and z_b can be expressed in terms of the covariance structure of $\mathbf{h}$ and the geometric relationship between the corresponding embedding vectors:

$$\mathrm{Cov}(z_a, z_b) = \mathbf{E}_{a,:}^{\top} \mathrm{Cov}(\mathbf{h}) \mathbf{E}_{b,:} \tag{6.13}$$

$$= \sigma_h^2 \cdot \mathbf{E}_{a,:}^{\top} \mathbf{E}_{b,:} \tag{6.14}$$

An important observation is that the term $\mathbf{E}_{a,:}^{\top}\mathbf{E}_{b,:}$ represents the inner product between the embedding vectors associated with vocabulary entries a and b, and therefore serves as a measure of their similarity. When the embeddings are normalized, this quantity is bounded within the interval $[-1, 1]$. Substituting this relationship into Equation 6.14 yields $\mathrm{Var}(z_a - z_b) = \mathcal{O}(d \cdot \sigma^2)$, where $\sigma^2 = \sigma_E^2 \sigma_h^2$ and the model dimension satisfies $d \propto \phi(N)$ for a linear or sub-linear function ϕ of the model size N. Under the proposed probabilistic framework, reduced similarity between the embedding vectors contributes to larger variance in the corresponding logit difference. Such situations may arise when the tokens associated with the reference and the model-preferred output occupy distinct regions of the representation space, as may occur in the presence of source-reference divergence. Applying Chebyshev's inequality from Equation 6.11 then yields the upper bound $\mathrm{Var}(z_a - z_b)/\xi(N)^2$, where $\xi(N)$ denotes a model-dependent function of N. Since $\mathrm{Var}(z_a - z_b)$ scales with the model dimension, the resulting bound need not diminish rapidly with increasing model size, implying that non-trivial deviations of the form $|z_a - z_b| \geq \xi(N)$ remain theoretically plausible under the stated assumptions. Consequently, the existence of meaningful logit discrepancies between reference tokens and model-preferred outputs is consistent with the proposed probabilistic model and provides a theoretical basis for analyzing the scaling behavior of the relative perplexity measure introduced as a proxy for factual inconsistency.

Recall that $z_a = z_i^{\mathrm{ref}}$ and $z_b = z_i^{\max}$. Under the probabilistic framework developed above, the preceding analysis suggests that non-trivial deviations between the logits assigned to reference tokens and model-preferred tokens may arise as a consequence of the variance structure induced by the model. Motivated by this observation, we assume that the typical magnitude of such deviations can be characterized by a model-dependent function $\xi(N)$, where N denotes the model size. Approximating the aggregate logit deviation in Equation 6.7 by this characteristic quantity allows the relative perplexity expression to be written in terms of $\xi(N)$, yielding:

$$\mathcal{R}(x, y^{\mathrm{ref}}) = \left(\prod_{i=1}^{L} e^{\xi(N)} \right)^{-\frac{1}{L}} \Bigg|_{\mathrm{input}=x} = e^{-\xi(N)} \Big|_{\mathrm{input}=x} \tag{6.15}$$

Taken together, the preceding analysis suggests that the relative perplexity measure $\mathcal{R}$, which is introduced as a proxy rather than an exact measure of factual inconsistency, may exhibit an exponential dependence on the model size N under the assumptions of the proposed probabilistic framework. In particular, if the aggregate logit deviations governing $\mathcal{R}$ can be characterized by a model-dependent function $\xi(N)$, then the resulting expression implies an exponential relationship between the proxy measure and model scale. Motivated by this observation, we generalize the form obtained in Equation 6.15 and arrive at the following scaling relationship:

$$\text{Factual Inconsistency} \propto e^{-\xi(N)}\big|_{\text{input}=x} \tag{6.16}$$

where $\xi(N)$ denotes a model-dependent function of the model size N, assumed to exhibit linear or sub-linear growth. Thus, under the stated assumptions, this mathematical rationale provides a probabilistic framework for the conjectured exponential scaling of factual inconsistency with model size in LLM-based D2T, as formulated in section 6.8.

$\Diamond$

6.9 Conclusion

This chapter presents an empirical analysis of the scaling behavior of *factual inconsistency* in D2T with respect to model size, computational capacity (FLOPs), and fine-tuning data size. While much of the existing scaling literature has focused on *power-law scaling*, the experimental results obtained across multiple LLM families, datasets, and evaluation metrics suggest that factual inconsistency may exhibit an *exponential scaling* pattern under the evaluated settings. To facilitate a systematic comparison of these scaling models, a three-stage statistical framework, *VaCScal*, is introduced, comprising predictive performance estimation, goodness-of-fit evaluation, and model comparison. Across the evaluated datasets and model families, the results generally indicate that *exponential scaling* provides a closer fit to the observed data than *power-law scaling* according to the considered evaluation criteria. In addition, a mathematical rationale is presented to provide further insight into these empirical observations, suggesting that under certain D2T conditions, particularly in the presence of source–reference divergence, exponential scaling may arise as a plausible characterization of factual inconsistency with increasing model size. Overall, the findings suggest that exponential scaling may offer a useful perspective for understanding the scaling behavior of factual inconsistency in fine-tuned LLM-based D2T systems within the scope of the evaluated settings.

Chapter 7

Conclusion and Future Directions

This thesis has undertaken a detailed investigation into the problem of *faithfulness* in language model-based D2T, with a particular emphasis on LLMs. Faithfulness, as conceptualized throughout this research, refers to the extent to which the generated text accurately reflects the underlying source and reference information, while carefully accounting for potential source–reference divergence. In this thesis, faithfulness is characterized through two central dimensions: *informativeness* and *factual consistency*. The work systematically explores both the evaluation and enhancement of faithfulness, bridging pre-LLM encoder–decoder models and the latest LLM-based systems, and contributes an integrated empirical perspective on several mechanisms, challenges, and scaling behaviors associated with faithfulness in D2T generation.

The Chapter 2 introduced a *structural encoder* for pre-LLM, encoder–decoder-based D2T models designed to effectively capture record–field structured data. By integrating attention gates, the model selectively extracted and utilized relevant content from encoded representations. Extensive experiments on the WikiBio and E2E datasets demonstrated that the proposed encoder robustly handled both sequential and non-sequential data, improving *informativeness*-related aspects of faithfulness in the evaluated settings. These findings suggest that explicit structural representations can play an important role in supporting accurate and contextually coherent generation.

The Chapter 3 shifted focus toward the capabilities of LLMs in D2T generation, providing a broad evaluations to date across twelve models from five prominent LLM families—BART, T5, BLOOM, OPT, and Llama 2—using multiple datasets and evaluation metrics. The study combined automatic and human evaluations with robustness analyses and introduced a *source–reference divergence-weighted training* strategy to improve model grounding. Findings revealed that larger models generally improve readability and informativeness, with Llama 2 often achieving strong performance across several

evaluated metrics and datasets. However, increases in model size did not consistently enhance faithfulness and, in some cases, resulted in degradation. Human evaluations showed that while larger models tend to produce more expressive and coherent text, minor factual inconsistencies are often overlooked by human raters but over-penalized by automatic metrics. Robustness analyses suggested that these observations hold across fine-tuning (QLoRA vs. Prefix-Tuning) and decoding (Beam Search vs. Nucleus Sampling) setups. The results further indicate that the divergence-weighted training method was associated with improvements in faithfulness, informativeness, and readability across the evaluated settings, suggesting its potential utility as a regularization mechanism for promoting factual grounding.

The Chapter 4 advanced the analysis by providing an extensive evaluation of *factual consistency* in LLM-based D2T generation. Five representative LLM families—T5, BART, OPT, BLOOM, and Llama 2—were evaluated on five widely used D2T datasets (E2E, ViGGo, WikiTableText, DART, and WebNLG) using four widely used factual consistency metrics: SummaC-Conv, QAFactEval, AlignScore, and NEOverlap. Human evaluations further complemented the automatic metrics. Three key observations emerged: (1) Llama 2 consistently exhibited the highest factual consistency; (2) smaller models such as T5 and BART achieved strong performance on large, lexically less diverse datasets; and (3) factual consistency generally declined as source–reference divergence increased across the evaluated datasets. By introducing the *Average Rate of Change (AROC)* metric, this study examined how factual consistency varied with model size across the evaluated model families. These findings deepened understanding of how LLM capacity, data characteristics, and divergence interact to influence faithfulness in D2T generation, providing empirical evidence that may inform future improvement strategies.

Chapter 5 investigated source–reference divergence, a phenomenon in D2T in which the reference text semantically deviates from the source data. To characterize this phenomenon quantitatively, three complementary measures were introduced: *unigram-level divergence*, *named entity-based divergence*, and *field-aware divergence*. Building upon these measures, the chapter presented a *divergence-aware selective generation* framework designed to enable LLMs to abstain from generating outputs for instances exhibiting high levels of divergence. Experiments conducted on the ViGGo, WikiTableText, and WebNLG datasets using three LLM families—Qwen2.5, FLAN-T5, and OPT—suggested that the proposed framework was often associated with higher factual consistency under coverage settings of 70–80% in the evaluated settings while retaining a large proportion of outputs. The findings further indicated that selective abstention was associated with reductions in factual inconsistencies and ungrounded content. Overall, the results suggest that explicitly modeling source–reference divergence may provide a useful approach

for studying and mitigating factual inconsistencies in LLM-based D2T systems and may inform future research on improving faithfulness in settings where factual accuracy is important.

Finally, Chapter 6 extended the investigation from empirical evaluation to the analysis of scaling behavior, examining how *factual inconsistency*—the inverse of factual consistency—varies with model size, computational capacity (FLOPs), and fine-tuning data size. While much of the existing scaling literature has focused on *power-law scaling*, the empirical results obtained in this chapter suggest that factual inconsistency may exhibit an *exponential scaling* pattern under the evaluated settings. To facilitate a systematic comparison between alternative scaling models, the chapter introduced a three-stage statistical framework, *VaCScal*, comprising predictive performance estimation, goodness-of-fit evaluation, and comparative model analysis. Experiments conducted across six LLM families and five D2T datasets generally indicated that exponential scaling provided a closer fit to the observed data than power-law scaling according to the considered evaluation criteria and under a range of fine-tuning and decoding configurations. In addition, a mathematical rationale was presented to provide further insight into these empirical observations, suggesting that under certain D2T conditions involving source–reference divergence, exponential scaling may arise as a plausible characterization of factual inconsistency with increasing model size. Collectively, the findings suggest that exponential scaling may offer a useful perspective for understanding the relationship between model capacity and factual inconsistency in LLM-based D2T systems within the scope of the evaluated settings.

7.1 Synthesis and Key Actionable Insights

The overall findings of this thesis suggest that faithfulness in language model-based D2T is influenced by multiple interacting factors and cannot be explained solely by model scale. The studies spanning both pre-LLM and LLM-based paradigms indicate that improvements in fluency or increases in model size do not necessarily correspond to improvements in factual consistency. Taken together, the findings provide the following actionable insights for the design and evaluation of D2T systems:

- **Preserve structured grounding:** Explicit structural representations and content selection mechanisms may help maintain informativeness and support alignment between generated text and source data.
- **Move beyond scale-centric design:** While larger models were generally associated with improved readability and informativeness in the evaluated settings,

improvements in faithfulness may require additional faithfulness-oriented mechanisms.

- **Identify and model divergence:** Source–reference divergence was frequently associated with lower faithfulness in the evaluated datasets, suggesting that it may be useful to quantify and account for divergence during training and generation.
- **Incorporate faithfulness-aware generation strategies:** Divergence-sensitive training and selective generation mechanisms were associated with improvements in factual consistency while maintaining reasonable coverage and generation quality in the evaluated settings.
- **Examine scaling behavior of factual inconsistency:** The *VaCScal* framework suggests that factual inconsistency may exhibit exponential scaling behavior under the evaluated conditions, providing an additional perspective for analyzing the relationship between model scale and factual consistency in D2T systems.

Collectively, these actionable insights provide a basis for studying and improving faithfulness in D2T systems through the analysis of grounding, source–reference divergence, and faithfulness-aware training and generation strategies. While the findings are derived from D2T tasks, some of the underlying principles may also be relevant to other forms of conditional text generation.

7.2 Future Directions

Building on the findings and frameworks presented in this thesis, several promising directions for future research emerge. These directions aim to deepen the understanding of faithfulness in D2T generation, enhance model grounding, and extend the frameworks developed herein to broader contexts of faithfulness and trustworthiness in Large Language Models (LLMs).

- **Divergence-aware reinforcement learning:** Future research can extend the proposed divergence-aware training paradigm into *reinforcement learning from human feedback* (RLHF). Integrating divergence-awareness within RLHF would allow models to dynamically reward factually grounded generations while penalizing unfaithful or hallucinated outputs. By coupling human preference signals with divergence-sensitive rewards, such frameworks could foster adaptive learning that balances fluency, informativeness, and factual accuracy, particularly in safety-critical applications such as medical or financial text generation.

- **Extending scaling analyses:** The *VaCScal* framework, introduced in this thesis for analyzing the scaling behavior of factual inconsistency, can be generalized to other generative properties, such as coherence, diversity, controllability, and stylistic variation. Investigating how these dimensions scale with model parameters, compute, and data would contribute to a broader understanding of LLM scaling behavior, offering deeper insights into the trade-offs between expressiveness and reliability in large-scale generative models.

- **Exploring other aspects of factual consistency in D2T generation:** While this thesis focused primarily on *source–reference divergence* as a key contributor to factual inconsistency, future research should investigate additional factors affecting factual consistency in D2T, such as input data sparsity, representational noise, contextual under-specification, and biases in pre-training corpora. A deeper understanding of these factors may help disentangle different sources of unfaithfulness and guide the development of more robust model architectures and training strategies. Furthermore, stronger integration among the concepts of faithfulness, trustworthiness, and factual consistency remains an important future direction in D2T. In particular, a closer synthesis of the frameworks and insights developed in Chapter 3 and Chapter 4 could provide a more unified foundation for building reliable and trustworthy language model-based D2T systems.

- **Exploring Source–Reference Divergence in Greater Detail:** While this thesis primarily focused on source–reference divergence in D2T using lexical and semantic perspectives, future research can investigate richer and more comprehensive formulations of divergence. These may include contextual, structural, discourse-level, or knowledge-based divergence measures that capture subtler forms of mismatch between source data and generated text. Furthermore, the concept of source–reference divergence can potentially be extended beyond D2T to broader conditional and controllable text generation tasks, such as summarization, dialogue generation, machine translation, and controllable generation settings where outputs are expected to satisfy specific attributes or constraints. Another promising direction is to predict source–reference divergence in real-world scenarios where reference text is unavailable, particularly in open-ended generation settings. Developing methods to estimate divergence directly from input and generated outputs could enable proactive detection of factual inconsistency and improve the reliability of language generation systems.

- **Factual consistency in different applications:** Investigating factual consistency across diverse real-world applications of D2T—such as clinical report

generation, automated journalism, sports commentary, and dialogue summarization—will help assess the generalizability of current findings. Each application domain has unique fidelity requirements and tolerance levels for factual deviations, and understanding these domain-specific constraints is vital for developing specialized LLM-based generation systems with improved factual consistency.

- **Evaluating factuality in D2T:** As discussed in section 1.4, an important direction for future research is the evaluation of factuality in D2T, where the truthfulness of generated text is assessed with respect to broader real-world knowledge rather than being confined solely to the input source data. This notion of factuality is particularly relevant to open-domain D2T tasks, where generated content may draw upon world knowledge that is not explicitly represented in the structured input. Developing models and evaluation metrics that align D2T outputs with verified real-world knowledge may further improve the factual grounding and reliability of D2T-generated text.

- **Factual consistency under prompting:** This thesis primarily investigates factual consistency in D2T under supervised fine-tuning (SFT) settings with LLMs. An important direction for future research is to examine how factual consistency and factuality in D2T is influenced by prompting strategies, such as in-context learning [76, 171] and chain-of-thought prompting [172]. Understanding the relationship between prompting techniques and factual consistency may help identify prompting configurations that better preserve source-grounded information. Such analyses could also provide insights into the mechanisms through which prompting affects factual errors and hallucinations in D2T systems.

7.3 Concluding Remarks

In conclusion, this thesis provides a systematic exploration of faithfulness in Data-to-Text generation, from structured encoder architectures in pre-LLM models to advanced LLM-based frameworks for evaluation, enhancement, and analytical insights. It offers practical solutions—such as divergence-weighted training and divergence-aware selective generation—and introduces the *VaCScal* framework, which suggests that exponential scaling may provide a useful characterization of factual inconsistency in D2T under the evaluated settings. The collective contributions of this work contribute to both empirical research and systematic understanding, may provide a basis for future research toward building LLM-based D2T that are not only fluent and informative but also factually grounded, and trustworthy in real-world applications.

Appendix A

Appendix: Exponential Scaling Factual Inconsistency of LLM in D2T

A.1 Scaling Behaviour Based on Model Size of Fine-Tuned (QLoRA) LLM Families With Human Evaluation

Although automatic metrics for evaluating factual consistency have improved in recent years, they still fall short compared to human-based evaluations. To further assess the suitability of exponential scaling for modeling factual inconsistency in D2T, we conducted a human evaluation involving five university student annotators from diverse academic backgrounds. This voluntary task was performed on outputs generated by four LLM families—FLAN-T5, Mamba, OPT, and Qwen2.5. Importantly, this evaluation was limited to QLoRA-fine-tuned versions of these LLM families. Prior to the annotation process, each annotator participated in a brief training session that included five example instances, along with corresponding source data and reference texts. Factual inconsistency was assessed on a Likert scale ranging from 5 (completely inconsistent) to 0 (fully consistent). In addition to assigning scores, annotators were required to justify inconsistencies using one or more of the following categories: (1) *missing fact*, (2) *fabricated fact*, and (3) *irrelevant fact*. The detailed annotation guidelines used for the human evaluation are similar to those described in section 4.7, except that the scoring scale was inverted to measure factual inconsistency as the inverse of factual consistency. A gold-labeled reference set was provided to guide the evaluation process (Table A.1). Each annotator independently evaluated 30 generated outputs across five D2T datasets, without consulting others or using external resources. Inter-annotator agreement was

high, with a Fleiss' kappa coefficient of 0.79, indicating substantial agreement among annotators.

TABLE A.1: The five short sample given to each annotators for training before annotation task.

source	reference	generated text	label (Likert Scale 0-5, with reasons)
('MotorSport Vision', 'city', 'Fawkham')	MotorSport Vision is located in Fawkham .	Fawkham is in the UK.	5 (inconsistent: irrelevant fact)
(('subjtitle', 'kf tirana'), ('subjsubtitle', 'presidents history'), ('name', 'lutfi nuri'), ('from–to', '1998–99'))	lutfi nuri was president of kf tirana during 1998–99 .	lutfi nuri president of kf tirana who took office 1998–99 .	0 (consistent)
(('subjtitle', 'aachen'), ('subjsubtitle', 'demographics'), ('year', '2015'), ('population', '245,885'))	the population of aachen 245,885 in 2015 .	245,885 people in aachen lived in 2015 .	0 (consistent)
(('subjtitle', 'fc rubin kazan'), ('subjsubtitle', 'average attendance'), ('year', '2011'), ('reg . season', '16,380'))	the average attendance of fc rubin kazanin 2011 were 16,380 .	fc rubin kazan got 16,380 attendences in 2011 .	1 (almost consistent: missing 'average')
(('name', 'Clowns'), ('eatType', 'pub'), ('near', 'The Sorrento'))	Close to The Sorrento you can find Clowns pub.	Clowns pub serves food for under £20 and is located near The Sorrento.	3 (inconsistent: fabricated fact)

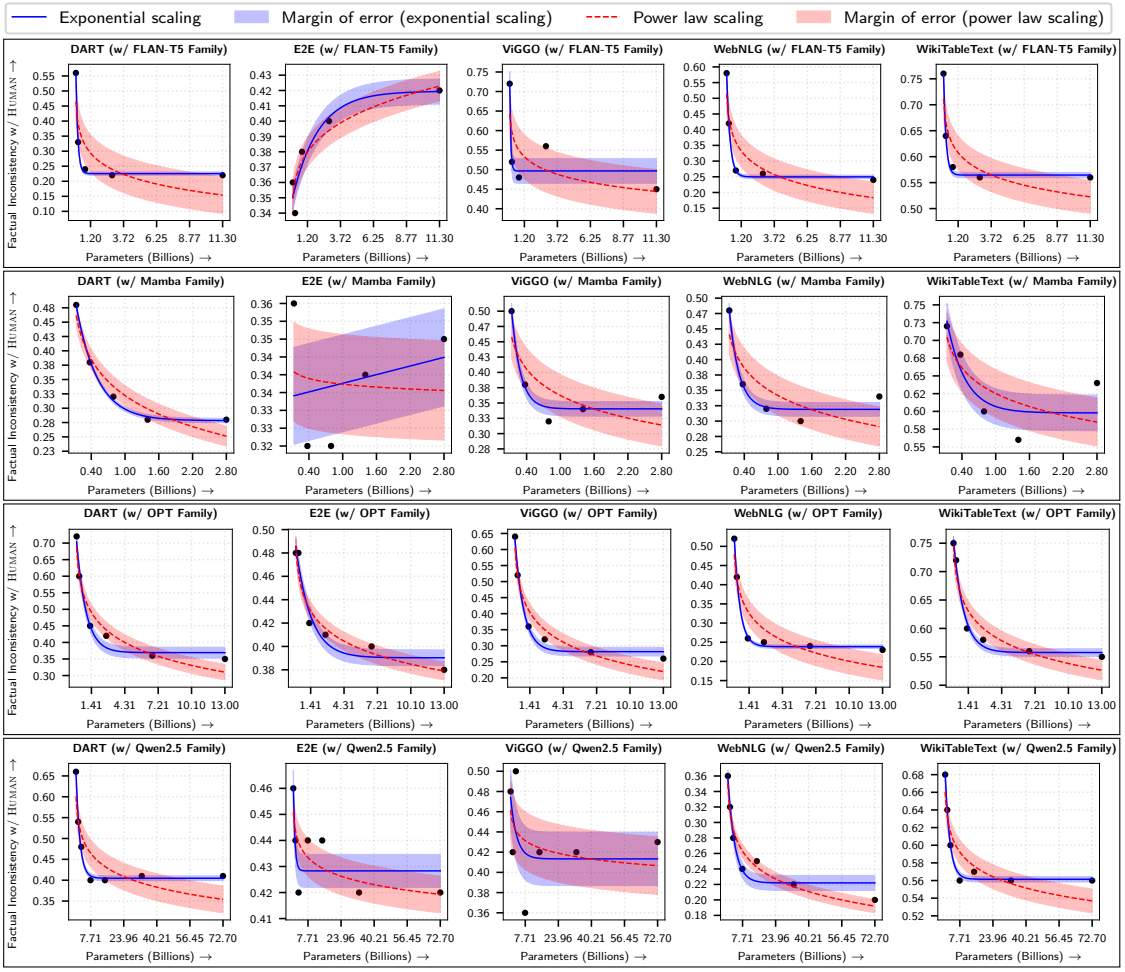


FIGURE A.1: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, based on HUMAN evaluation, plotted against model size for QLoRA fine-tuned FLAN-T5, Mamba, OPT, and Qwen2.5 LLM families.

Figure A.1 presents the fitted power law and exponential scaling curves, along with 95% confidence intervals (MoE), based on human-annotated assessments of factual inconsistency across varying model sizes (in terms of parameter count). Across all D2T datasets,

exponential scaling consistently exhibits lower MoEs than power law scaling, indicating lower uncertainty under human evaluation. The corresponding *VaCScal* results are shown in Table A.2, where, in almost all datasets (except E2E), the LLM families favor exponential scaling. This is evidenced by the blue-highlighted cells denoting strong statistical preference for exponential scaling in Stage III of the Vuong test. While some exceptions are observed in the E2E dataset, even in these cases, exponential scaling achieves lower held-out error in Stage I compared to power law scaling—indicating better predictive performance under the evaluated settings. Although in two rare cases, where power law scaling outperforms exponential scaling for the DART and WebNLG datasets with FLAN-T5 models, exponential scaling remains competitive in Stages I and II of the *VaCScal* framework. Except for these rare cases, all other consistent *VaCScal* outcomes further support the suitability of exponential scaling for modeling the relationship between model size and factual inconsistency in D2T.

Takeaway. *In human evaluations across the FLAN-T5, Mamba, OPT, and Qwen2.5 LLM families, exponential scaling receives greater support as a model for the decrease of factual inconsistency with increasing model size, offering a more stable characterization than power-law scaling under the evaluated settings.*

TABLE A.2: Results from *VaCScal* corresponding to Figure A.1. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
FLAN-T5	Exponential	1.14E-04/✓	4.31E-05/✗	1.54E-03/✓	2.48E-03/✓	4.35E+00/✓
	Power Law	1.54E-04/✓	7.06E-04/✗	8.62E-05/✓	5.05E-04/✓	6.02E-04/✓
Mamba	Exponential	3.28E-06/✓	8.17E-05/✗	1.69E-04/✓	2.84E-04/✓	6.68E-04/✗
	Power Law	4.51E+05/✗	1.47E-03/✗	2.53E+176/✗	5.66E-01/✗	8.87E+13/✗
OPT	Exponential	1.25E-03/✓	4.29E-04/✓	4.13E-03/✓	2.54E-04/✓	2.84E-03/✓
	Power Law	7.00E+01/✗	1.26E+00/✗	6.51E+02/✗	1.58E+10/✗	2.03E-03/✗
Qwen2.5	Exponential	4.51E-04/✓	1.26E-04/✗	4.05E+00/✓	6.08E-05/✓	7.98E-05/✓
	Power Law	3.88E+00/✓	6.44E-04/✗	1.03E-04/✗	9.36E-05/✗	4.05E-01/✗

A.2 Result: Scaling Behavior Based on Model Size and FLOPs for Fine-Tuned (QLoRA) BLOOM and OPT Families

Here, we present the scaling behavior of fine-tuned (QLoRA) BLOOM and OPT families with respect to factual inconsistency, measured using AlignScore and QAFactEval. Figure A.2 and Figure A.3 illustrate the fitted power law and exponential scaling models for both families, based on model size and FLOPs, respectively. We find that scaling trends based on model size and FLOPs are largely consistent with each other. Similar to the other four LLM families, exponential scaling generally provides a better fit than power law scaling, particularly for the OPT family, as indicated by narrower margins of error (MoE). Although BLOOM exhibits relatively higher MoEs, exponential scaling still consistently achieves lower MoEs compared to power law scaling. Despite the aberrant scaling behavior observed for the E2E and ViGGO datasets (Appendix A.8), the corresponding *VaCScal* results in Table A.3 and Table A.4 provide evidence favoring exponential scaling over power-law scaling across the three stages of *VaCScal*.

Table A.3: Results from *VaCScal* corresponding to Figure A.2. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Metric	LLM family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
AlignScore	BLOOM	Exponential	1.84E-04/✗	6.29E+03/✗	5.82E+03/✗	2.51E-04/✓	1.43E+03/✗
		Power Law	6.92E-04/✗	6.65E-04/✗	3.30E-04/✗	2.43E-02/✗	1.55E-03/✗
	OPT	Exponential	1.64E-04/✓	1.15E-01/✗	1.37E-03/✓	3.66E-04/✓	9.04E-04/✓
		Power Law	2.96E-03/✗	1.16E+82/✗	3.72E+81/✗	1.94E-03/✗	1.10E+03/✓
QAFactEval	BLOOM	Exponential	2.07E-04/✓	7.81E-05/✗	7.41E-04/✗	4.81E-04/✓	3.76E-04/✗
		Power Law	1.00E-01/✗	2.45E-03/✗	8.86E-04/✗	3.10E-02/✗	4.43E+06/✗
	OPT	Exponential	1.13E-04/✓	5.28E-05/✗	1.28E-03/✗	3.63E-03/✓	1.35E-04/✓
		Power Law	2.00E+01/✗	9.51E-02/✗	1.63E-02/✗	5.69E+02/✓	4.42E+00/✗

Table A.4: Results from *VaCScal* corresponding to Figure A.3. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Metrics	LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
AlignScore	BLOOM	Exponential	4.77E-03/✗	1.35E+03/✗	1.13E+03/✗	3.60E-04/✓	2.22E+02/✗
		Power Law	1.64E-01/✗	1.39E-02/✗	2.83E+198/✗	2.64E-04/✗	1.35E-03/✗
	OPT	Exponential	3.40E-04/✓	1.83E-04/✗	1.19E-03/✓	1.16E-04/✓	4.44E+02/✓
		Power Law	9.22E+01/✗	1.71E-04/✗	5.96E-04/✗	1.69E+08/✗	6.29E+02/✓
QAFactEval	BLOOM	Exponential	4.31E-04/✓	1.33E-04/✗	6.10E-04/✗	6.60E-04/✓	2.27E-04/✗
		Power Law	4.75E-02/✗	6.22E-05/✗	4.54E-04/✗	2.88E-02/✗	1.54E-02/✗
	OPT	Exponential	1.85E-04/✓	2.00E-05/✗	1.03E-03/✗	2.97E-03/✓	1.23E-04/✓
		Power Law	2.59E+01/✗	1.71E-04/✗	6.37E+04/✗	1.30E+07/✓	1.52E+00/✗

Takeaway: *In D2T, exponential scaling demontrate the factual inconsistency (measured by* AlignScore *and* QAFactEval*) with increasing model size and FLOPs more*

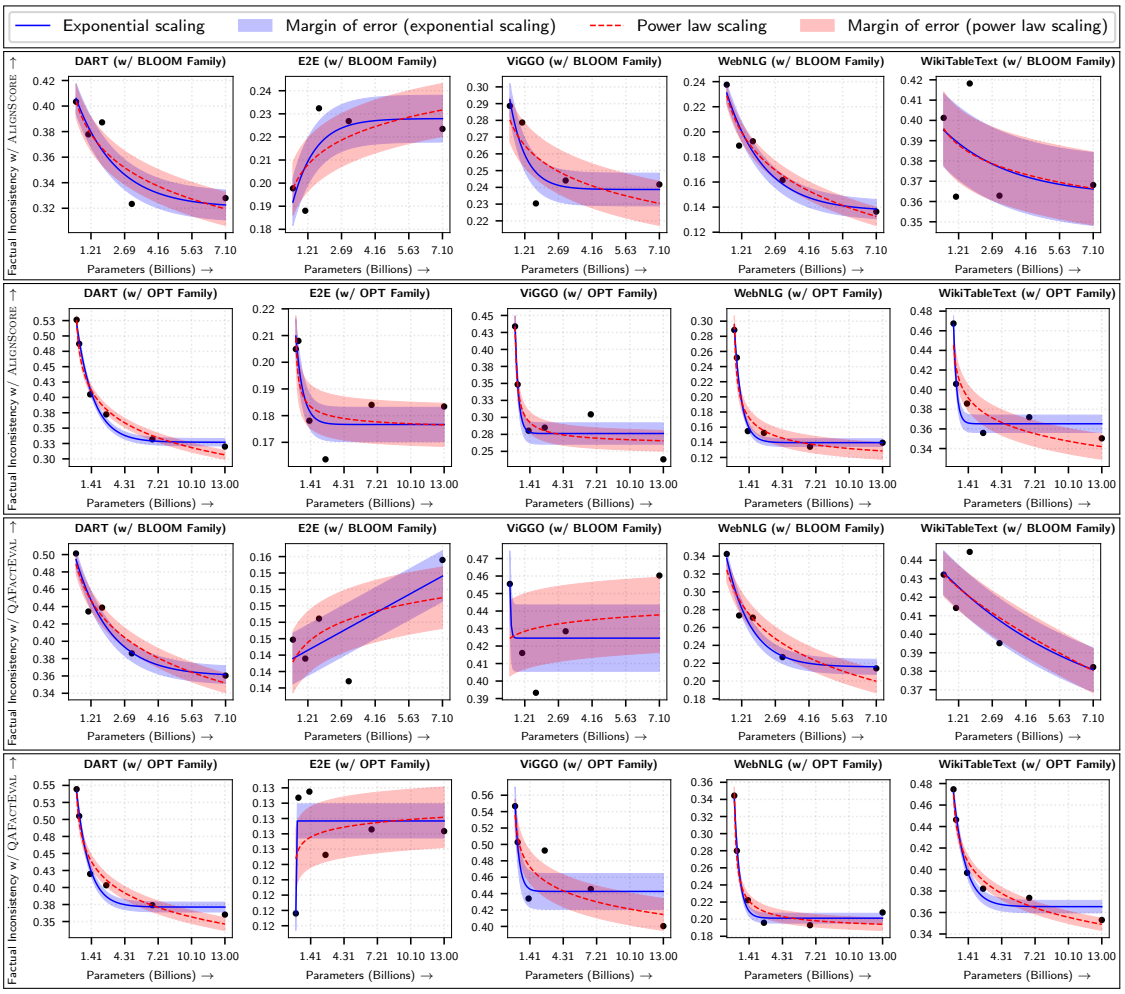


FIGURE A.2: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFACTEVAL, plotted against model size across different LLM families.

reliably than power law scaling across BLOOM and OPT families, except a few minor exception.

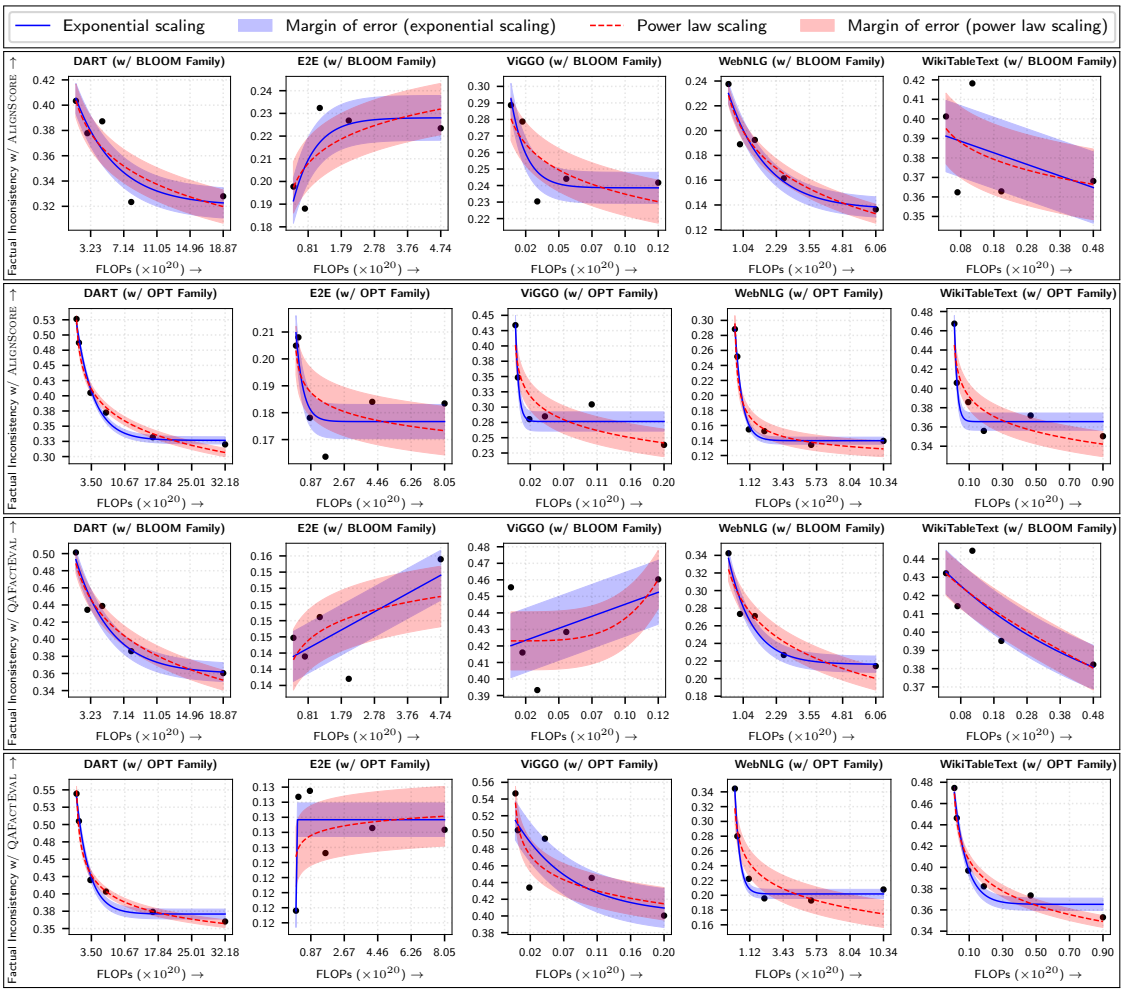


FIGURE A.3: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by ALIGNSCORE, plotted against FLOPs across different LLM families.

A.3 Result: Scaling Behavior Based on Model Size of Fine-Tuned (QLoRA) LLM Families

A.3.1 Factual Inconsistency Measured via SummaC-conv

Figure A.4 presents fitted power law and exponential scaling curves (with 95% confidence intervals) for factual inconsistency measured via SUMMAC-CONV, plotted against model size across all six LLM families. Although some anomalies appear—particularly in the E2E dataset with BLOOM—exponential scaling consistently yields lower MoE and lower held-out loss. Table A.5 shows corresponding *VaCScal* results, where exponential models pass statistical tests more frequently and are preferred by Vuong's test in most settings. These *VaCScal* results provide additional evidence for the consistency of the observed exponential scaling trend in modeling factual inconsistency—measured via SUMMAC-CONV—with respect to LLM model size across families.

Takeaway: *Exponential scaling captures better the decline of factual inconsistency (measured by* SummaC-conv*) with increasing model size more reliably than power law scaling across LLM families, supported by tighter confidence margins and stronger statistical validation from VaCScal.*

Table A.5: Results from *VaCScal* corresponding to Figure A.4. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
BLOOM	Exponential	1.79E-04/✗	2.39E+02/✗	2.45E-04/✗	5.40E-04/✗	2.20E-05/✓
	Power Law	2.72E-02/✗	2.98E-02/✗	2.11E-04/✗	7.86E-03/✗	9.19E-03/✗
FLAN-T5	Exponential	6.46E-04/✓	9.07E-06/✗	2.39E-04/✓	2.83E-04/✓	3.13E+00/✓
	Power Law	2.12E-03/✓	1.35E-05/✗	2.43E+01/✗	2.96E-04/✓	4.74E-05/✓
Mamba	Exponential	7.88E-05/✓	1.76E-06/✗	1.88E+01/✗	1.52E-04/✓	3.04E-04/✓
	Power Law	8.26E-06/✗	1.00E-04/✗	1.08E+168/✗	1.14E-04/✗	5.13E+11/✗
OPT	Exponential	1.92E-04/✓	2.85E-01/✗	6.21E-04/✗	3.48E-05/✓	4.75E-04/✓
	Power Law	3.94E-01/✗	8.17E+80/✗	1.42E-04/✗	6.67E+10/✗	5.98E+04/✓
Pythia	Exponential	1.38E-03/✓	1.05E-03/✓	8.16E-04/✓	5.51E-03/✓	4.07E-04/✓
	Power Law	5.21E-03/✗	1.25E-02/✓	3.14E-03/✗	1.42E+09/✗	7.34E-03/✗
Qwen2.5	Exponential	4.96E-04/✓	1.15E-04/✓	1.21E-03/✗	2.12E-04/✓	2.52E-04/✗
	Power Law	1.38E-03/✓	3.70E-03/✗	3.08E-04/✗	1.57E-02/✗	7.73E-03/✗

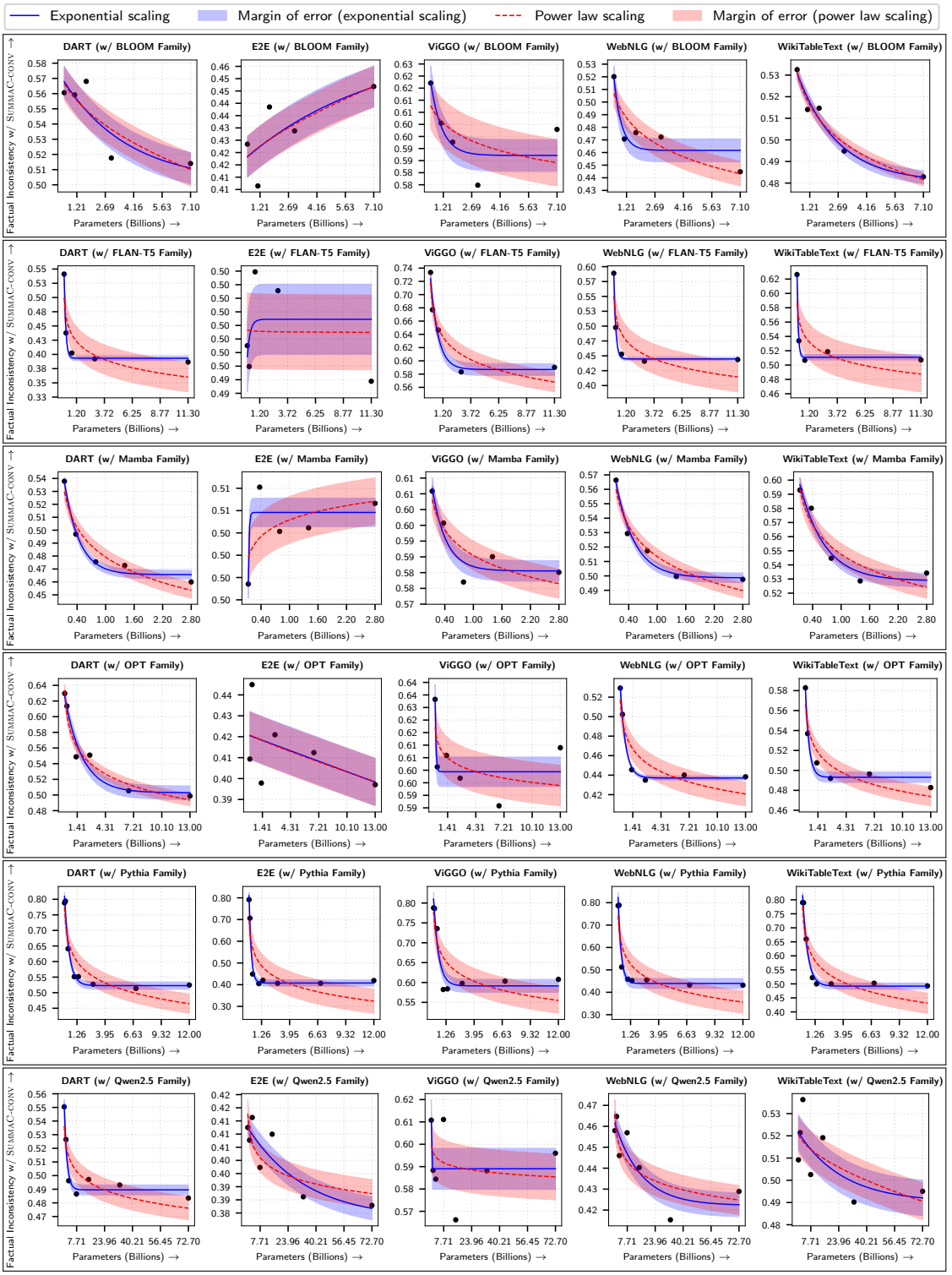


FIGURE A.4: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by SUMMAC-CONV, plotted against model size across different LLM families.

A.3.2 Factual Inconsistency Measured via UniEval-fact

Figure A.5 shows the fitted power law and exponential scaling curves with 95% confidence intervals (MoE) for factual inconsistency measured by UniEval-fact. As with other metrics (AlignScore, QAFactEval, SummaC-conv), anomalies persist—most notably in the E2E dataset and BLOOM family, which exhibit increasing inconsistency or large MoEs. Despite these outliers, exponential scaling consistently yields substantially lower MoEs across all cases compared to power law scaling, indicating reduced uncertainty in estimation. The corresponding *VaCScal* results (Table A.6) further support this observation: exponential scaling achieves significantly lower held-out loss in Stage I. In Stages II and III, exponential models suggest greater statistical feasibility (marked with ✓) and receive more support from Vuong's test (highlighted in blue) than their power law counterparts. Overall, the *VaCScal* results are consistent with the observation that exponential scaling provides a better fit than power-law scaling for characterizing how factual inconsistency varies with model size under the evaluated D2T settings using UniEval-fact.

Takeaway. *Exponential scaling more reliably models the relationship between model size of LLMs and factual inconsistency in D2T tasks, as measured by* UniEval-fact*, outperforming power law scaling despite a few dataset-specific anomalies.*

Table A.6: Results from *VaCScal* corresponding to Figure A.5. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
BLOOM	Exponential	3.07E-04/✓	1.39E-04/✗	6.31E-04/✗	4.78E-06/✓	3.45E-04/✓
	Power Law	3.25E-02/✗	8.95E-03/✗	1.49E-04/✗	1.06E-02/✗	1.36E-04/✗
FLAN-T5	Exponential	2.15E-04/✓	5.01E-05/✗	8.94E-04/✓	1.26E-03/✓	7.04E-01/✓
	Power Law	2.64E-04/✓	4.26E+00/✗	1.74E-03/✗	1.74E-03/✓	8.80E-05/✓
Mamba	Exponential	4.89E-05/✓	1.22E-04/✗	1.98E+02/✓	4.13E-05/✓	2.92E-04/✓
	Power Law	3.56E+06/✗	1.18E-04/✗	4.22E+170/✗	3.91E-04/✗	1.09E-03/✗
OPT	Exponential	1.13E-03/✓	8.96E-05/✗	2.81E-03/✓	5.28E-01/✓	1.23E-04/✓
	Power Law	3.08E+02/✗	5.89E-02/✗	6.76E-04/✓	1.20E+87/✗	3.69E+03/✓
Pythia	Exponential	1.06E-02/✓	1.06E-02/✓	7.36E-04/✓	6.82E-03/✓	4.14E-02/✓
	Power Law	1.36E+05/✗	2.06E+13/✓	2.11E+03/✗	4.45E+10/✗	4.23E+00/✗
Qwen2.5	Exponential	4.51E-04/✓	1.26E-04/✗	4.05E+00/✓	6.08E-05/✓	7.98E-05/✓
	Power Law	3.88E+00/✓	6.44E-04/✗	1.03E-04/✗	9.36E-05/✗	4.05E-01/✗

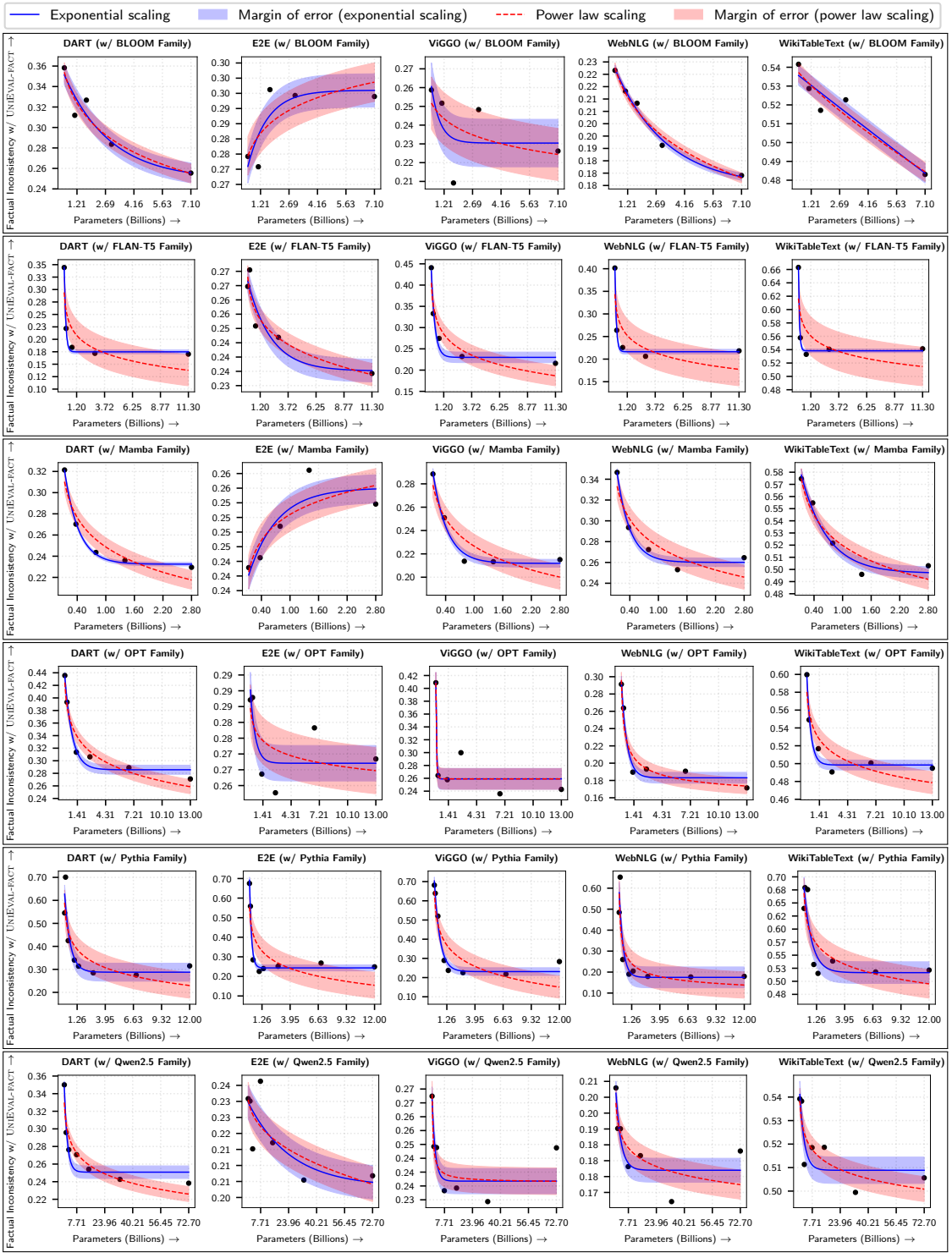


FIGURE A.5: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by UNIEVAL-FACT, plotted against model size across different LLM families.

A.4 Results: Scaling Behavior Based on Model Size of Fine-Tuned (QLoRA) LLM Families with Nucleus Sampling

As discussed in section 6.7, all prior results were based on greedy decoding. However, decoding strategy significantly influences the factual inconsistency of LLM outputs. To examine whether the observed preference for exponential scaling remains consistent under alternative decoding settings over power law scaling, we conduct additional experiments using nucleus sampling (with a nucleus probability of 0.95)—a widely used stochastic decoding method. For this setup, five LLM families (BLOOM, FLAN-T5, Mamba, OPT, and Pythia) are fine-tuned on the five D2T datasets using QLoRA, and factual inconsistency is assessed using AlignScore and QAFactEval.

Table A.7: Results from *VaCScal* corresponding to Figure A.6. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
BLOOM	Exponential	2.99E-04/✓	1.43E-03/✗	1.62E-03/✓	2.60E-04/✓	2.03E-03/✗
	Power Law	2.63E+02/✗	2.82E-04/✗	3.84E-01/✗	2.68E-03/✗	5.65E-01/✗
FLAN-T5	Exponential	2.59E-04/✓	5.44E-04/✗	7.12E-03/✓	2.59E-04/✓	9.13E-05/✓
	Power Law	2.26E-03/✓	5.46E-06/✗	2.55E-01/✗	4.34E+08/✗	5.38E+14/✓
Mamba	Exponential	1.01E-03/✓	5.39E-05/✗	1.39E-03/✓	4.22E-04/✓	1.97E-04/✓
	Power Law	3.88E+03/✗	1.05E-03/✗	6.78E+05/✗	2.51E+02/✗	5.08E+03/✗
OPT	Exponential	9.81E-04/✓	8.72E-01/✗	3.64E-04/✓	2.68E-04/✓	3.99E-03/✓
	Power Law	9.50E+03/✗	1.86E+79/✗	2.37E+03/✗	7.56E-03/✗	1.51E-03/✗
Pythia	Exponential	5.34E-03/✓	3.06E-01/✓	3.18E-03/✓	5.82E-03/✓	4.26E-03/✓
	Power Law	1.56E-02/✗	1.97E-02/✓	2.17E+03/✗	8.13E+04/✗	1.24E+04/✗

Figure A.6 and Figure A.7 show fitted power law and exponential scaling curves (with 95% confidence intervals) for factual inconsistency—measured by AlignScore and QAFactEval—under nucleus sampling ($p = 0.95$). The trends largely mirror those from greedy decoding: exponential scaling yields consistently lower MoEs, indicating greater stability. Exceptions such as the E2E dataset and the BLOOM family persist, with large uncertainties and irregular patterns. *VaCScal* results in Table A.7 and Table A.8 reaffirm exponential scaling as the better-supported model across most dataset-model combinations. These findings suggest that the observed conclusions remain largely consistent under stochastic decoding, consistent with the results obtained using greedy decoding across all our experiments.

Takeaway: *Exponential scaling continues to receive greater support from the evaluated statistical criteria for characterizing factual inconsistency in D2T—even under*

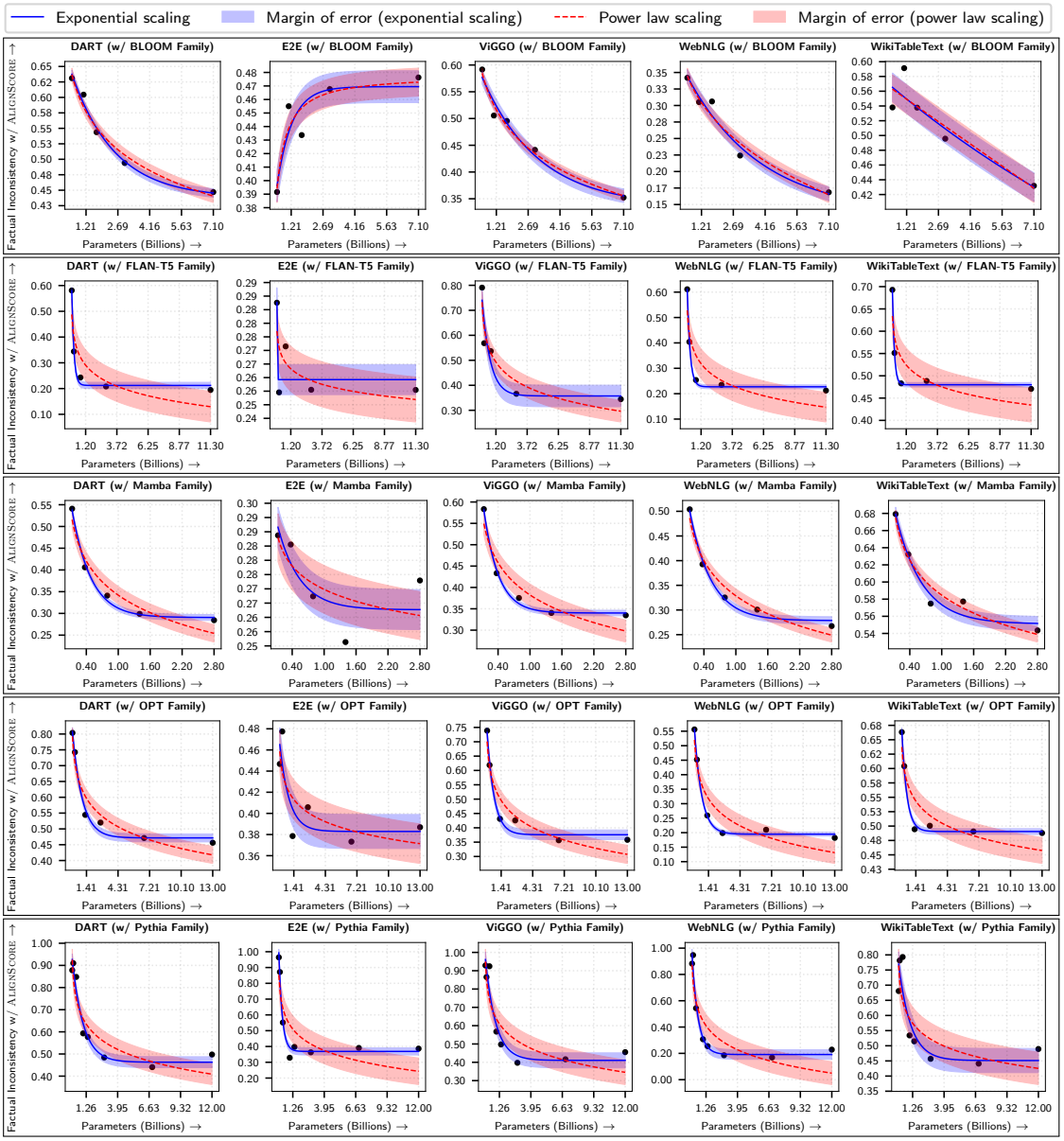


FIGURE A.6: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by ALIGNSCORE, plotted against model size across different LLM families. Here we have considered nucleus sampling for text generation.

stochastic decoding with nucleus sampling—providing additional evidence that the observed trends remain consistent under the evaluated decoding settings across decoding strategies.

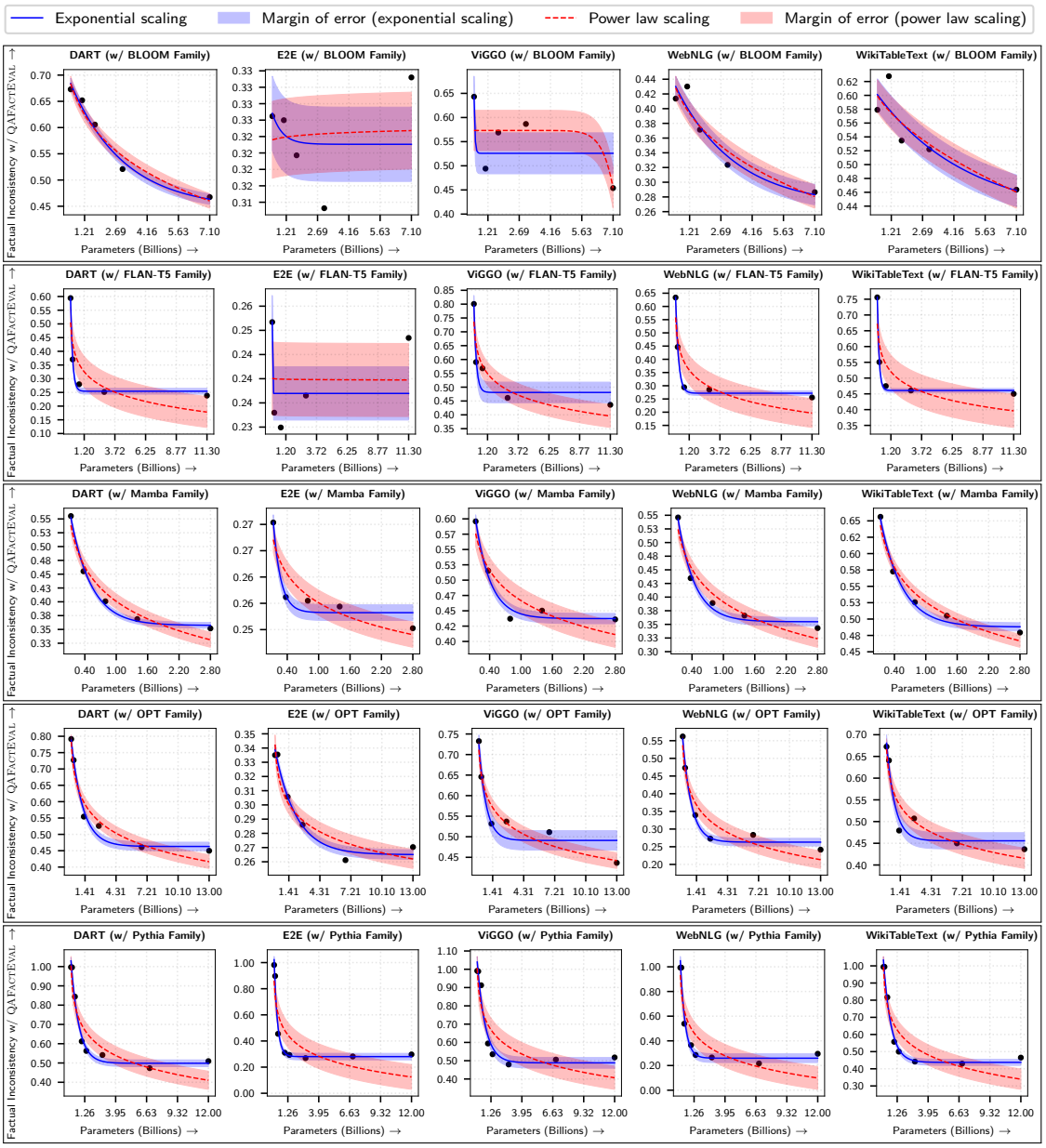


FIGURE A.7: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFACTEVAL, plotted against model size across different LLM families. Here we have considered nucleus sampling for text generation.

TABLE A.8: Results from *VaCScal* corresponding to Figure A.7. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
BLOOM	Exponential	5.01E-03/✓	7.46E+01/✗	4.92E-03/✗	2.76E-03/✓	1.49E-03/✗
	Power Law	8.35E+00/✗	5.92E-04/✗	9.12E-03/✗	2.51E-03/✗	2.84E+05/✗
FLAN-T5	Exponential	2.80E-04/✓	6.21E-05/✗	8.37E-03/✗	3.01E-04/✓	3.11E-03/✓
	Power Law	5.65E+05/✓	1.12E-03/✗	8.98E-02/✗	5.46E-03/✗	1.81E+07/✓
Mamba	Exponential	3.32E-04/✓	5.82E-05/✓	9.18E+01/✓	1.85E-04/✓	3.01E-04/✓
	Power Law	2.23E+03/✗	2.81E-06/✓	7.28E-04/✗	1.74E+01/✗	1.37E-04/✗
OPT	Exponential	2.83E-03/✓	1.58E-05/✓	2.14E-03/✓	1.13E-03/✓	2.64E-02/✓
	Power Law	2.31E+01/✗	6.34E-01/✗	1.02E+08/✗	3.28E+05/✗	1.24E+08/✗
Pythia	Exponential	1.30E-03/✓	2.67E-03/✓	3.60E-03/✓	7.31E-03/✓	5.69E-04/✓
	Power Law	4.16E+08/✗	4.99E+07/✓	2.04E+02/✗	2.90E-02/✗	4.00E+10/✗

A.5 Result: Scaling Behavior Based on Model Size of Fine-Tuned (Prefix-Tuning) LLM Families

In this section, we report results from experiments where LLM families (BLOOM, FLAN-T5, OPT, and Pythia) were fine-tuned using Prefix-Tuning, to examine whether the observed conclusions remain consistent across different fine-tuning strategies. We compare these results against the previously used QLoRA setup. Factual inconsistency is evaluated using AlignScore and QAFactEval, as these metrics were sufficient to assess whether the trends observed under QLoRA remained consistent. Note that Prefix-Tuning was not applied to the Mamba family due to stability issues encountered during fine-tuning.

A.5.1 Factual Inconsistency Measured via AlignScore and QAFactEval.

Here we present experimental results using *Prefix-Tuning* as an alternative fine-tuning strategy to the previously employed QLoRA. This comparison aims to assess the robustness of our findings across different fine-tuning approaches. We evaluate factual inconsistency using two representative automatic metrics—AlignScore and QAFactEval—which have consistently captured key trends and are sufficient to support our conclusions. Note that the Mamba family was excluded from this setup due to instability encountered during Prefix-Tuning.

Table A.9: Results from *VaCScal* corresponding to Figure A.8. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
BLOOM	Exponential	2.14E-04/✓	1.16E-02/✗	4.73E+03/✗	2.78E-05/✓	3.00E+03/✓
	Power Law	1.31E+05/✗	6.16E-03/✗	3.25E-03/✗	1.08E-03/✗	1.17E+203/✓
FLAN-T5	Exponential	1.09E-02/✗	3.01E-02/✗	5.14E-03/✗	3.88E-02/✗	3.08E-03/✗
	Power Law	1.49E-01/✗	8.41E-02/✗	6.03E-04/✗	8.07E-03/✗	8.31E-04/✗
OPT	Exponential	7.88E-04/✓	4.38E-03/✓	5.62E-03/✗	3.40E-03/✓	6.92E-04/✓
	Power Law	1.75E+09/✗	4.50E+56/✓	1.92E+01/✗	1.82E+02/✓	1.37E+03/✗
Pythia	Exponential	2.49E-02/✗	1.82E-02/✗	6.60E-02/✗	8.25E-03/✗	2.65E-03/✓
	Power Law	4.80E-02/✗	2.07E-02/✗	1.84E-02/✗	2.04E-02/✗	9.44E-02/✓

Figure A.8 and Figure A.9 present the fitted power law and exponential scaling curves with 95% confidence intervals (MoE) for factual inconsistency, measured using AlignScore and QAFactEval, plotted against model size (in terms of parameter count). For AlignScore, both the FLAN-T5 and Pythia families exhibit unusually large MoEs under both scaling models; similarly, for QAFactEval, the Pythia family shows high

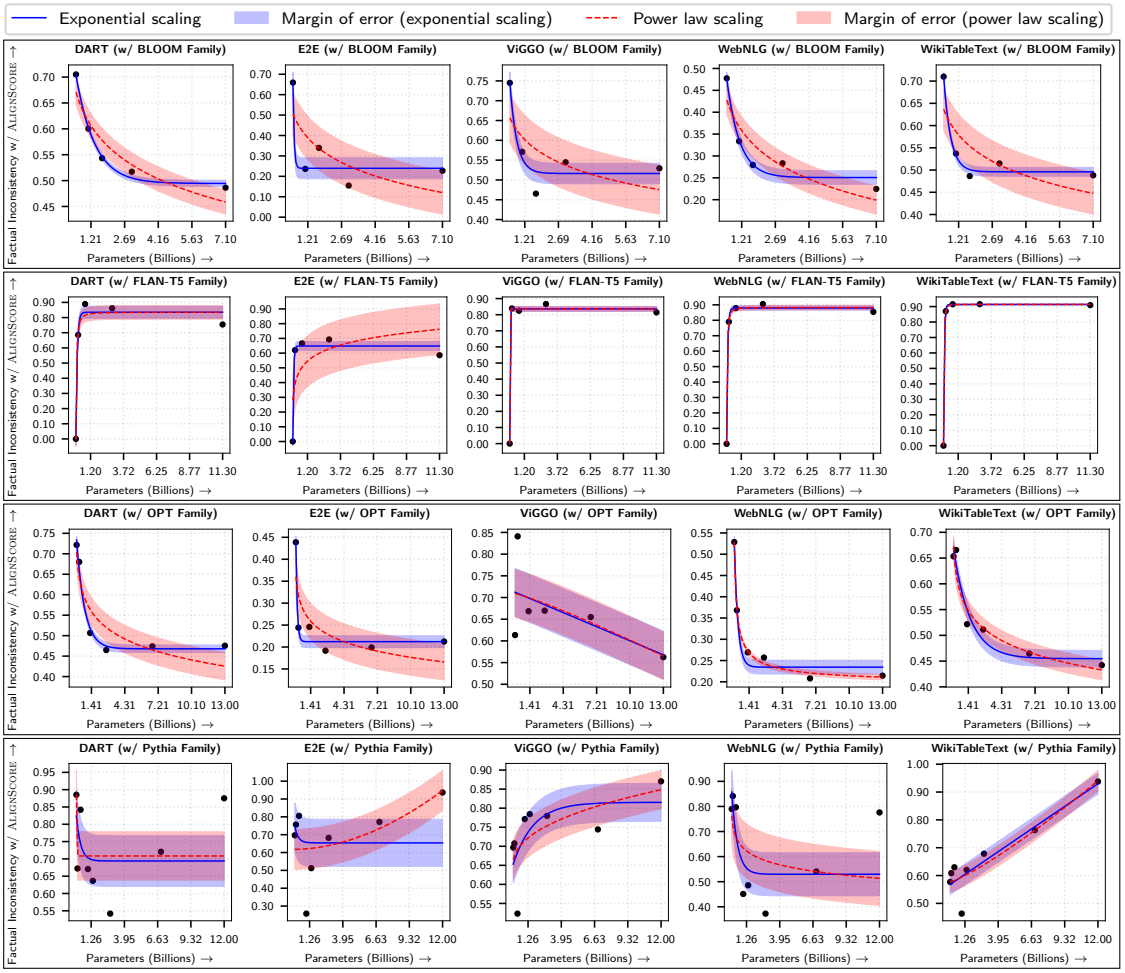


FIGURE A.8: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by ALIGNSCORE, plotted against model size across different LLM families. Here we consider Prefix-Tuning for fine-tuning all LLM families.

uncertainty. We hypothesize that this arises from the uniform addition of 32 virtual tokens during Prefix-Tuning across all models within a family. This fixed augmentation likely benefits smaller models disproportionately, while offering diminishing returns for larger ones, thereby limiting their improvement in factual consistency. Despite these anomalies, exponential scaling generally results in tighter MoEs than power law scaling, indicating more stable and confident fits. The corresponding *VaCScal* results, shown in Table A.9 and Table A.10, reveal that although many model–dataset combinations fail feasibility in Stage II, exponential scaling consistently achieves lower held-out loss (Stage I) compared to power law scaling. Moreover, in the feasible cases examined, exponential scaling generally provides a better fit than power-law scaling in capturing the relationship between factual inconsistency and model size.

Takeaway. *Exponential scaling continues to receive greater support from the evaluated statistical criteria than power-law scaling in modeling factual inconsistency under Prefix-Tuning, despite some anomalies caused by uniformly applied virtual tokens—especially in smaller models like FLAN-T5 and Pythia. Even when feasibility is limited, exponential scaling consistently yields lower held-out loss and better fits in all valid cases.*

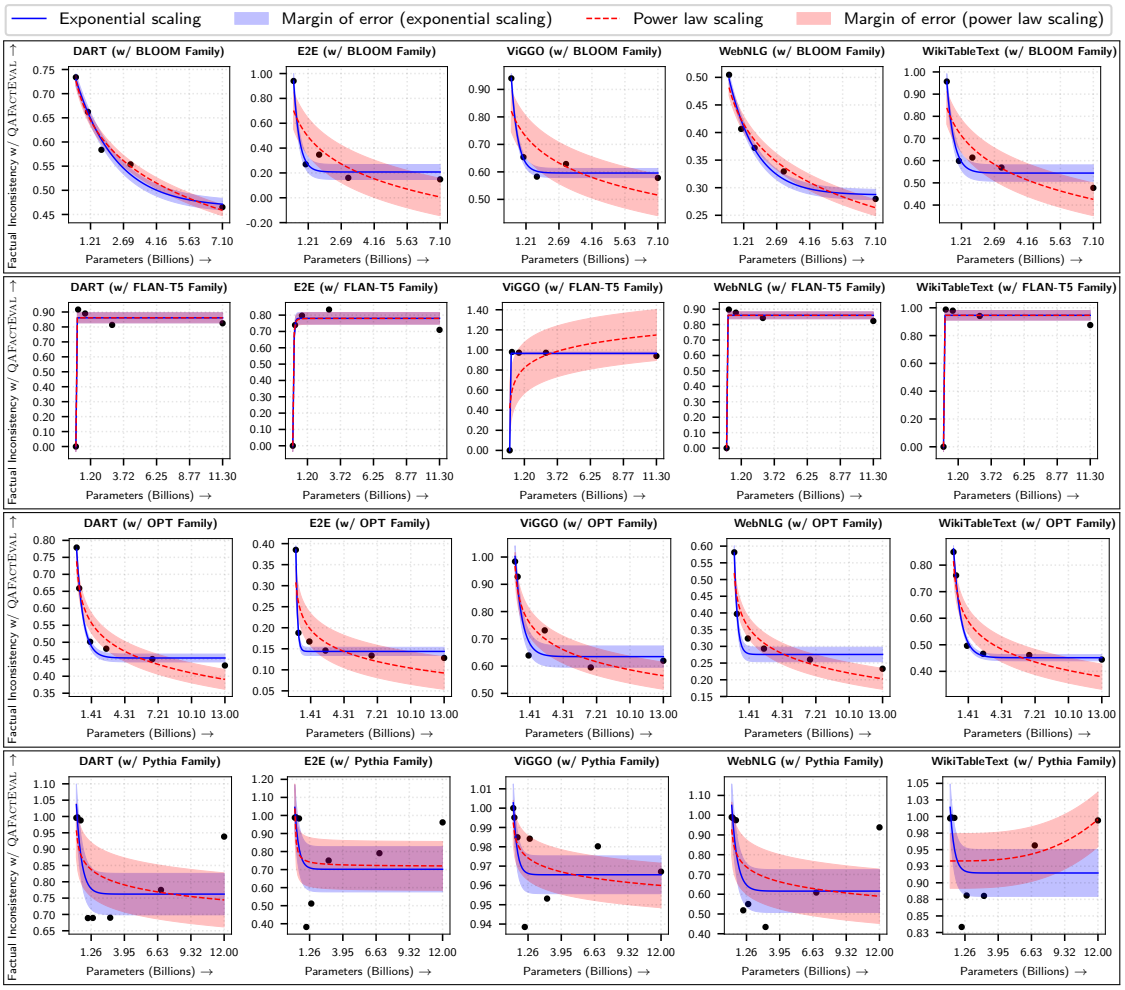


FIGURE A.9: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFACTEVAL, plotted against model size across different LLM families. Here we consider Prefix-Tuning for fine-tuning all LLM families.

TABLE A.10: Results from *VaCScal* corresponding to Figure A.9. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

LLM Family	Scaling	DART	E2E	ViGGO	WebNLG	WikiTableText
BLOOM	Exponential	2.99E-04/✓	1.17E-02/✗	1.01E-03/✓	9.22E-04/✓	3.07E-03/✗
	Power Law	6.86E-01/✗	8.30E-02/✗	1.05E-03/✓	2.33E-01/✗	7.28E-02/✓
FLAN-T5	Exponential	1.13E+00/✗	1.77E-02/✗	2.36E-04/✗	1.42E-05/✓	7.38E-04/✓
	Power Law	9.30E+95/✗	1.52E-01/✗	1.11E-02/✗	1.55E+02/✗	7.74E-04/✗
OPT	Exponential	7.88E-04/✓	4.38E-03/✓	5.62E-03/✗	3.40E-03/✓	6.92E-04/✓
	Power Law	1.75E+09/✗	4.50E+56/✓	1.92E+01/✗	1.82E+02/✓	1.37E+03/✗
Pythia	Exponential	1.51E-02/✗	2.06E-02/✗	9.79E-05/✗	2.04E-02/✗	2.93E-03/✗
	Power Law	1.46E+51/✗	3.66E-02/✗	4.51E+07/✗	2.80E-02/✗	1.09E+50/✗

A.6 Result: Scaling Behavior Based on Fine-tune Data Size

Data size is a fundamental factor in understanding scaling behavior. To investigate its effect on factual inconsistency in D2T tasks with LLMs, we analyze how varying the amount of fine-tuning data influences factual inconsistency outcomes. Here, "fine-tuning data" refers to different proportions of the original training set—specifically 10%, 20%, 40%, 60%, and 100%—used during fine-tuning with the QLoRA technique. This experiment is conducted on three representative D2T datasets: DART (graph-to-text), E2E (MR-to-text), and WikiTableText (table-to-text), with factual inconsistency measured using AlignScore and QAFactEval. We use the largest variants from each of the five incorporated LLM families (BLOOM, FLAN-T5, Mamba, OPT, and Pythia). Figure A.10 and Figure A.11 show the fitted power law and exponential scaling curves along with 95% confidence intervals (MoE), derived from residual-based estimation. Across both metrics and all model families, the E2E dataset consistently exhibits aberrant behavior—characterized by wide confidence bands and unstable scaling patterns—likely due to its inherent characteristics, which we discuss in more detail in Appendix A.8. DART also shows elevated MoEs, especially under AlignScore, whereas WikiTableText suggests relatively stable scaling behavior, with low MoEs under both models. The corresponding *VaCScal* results, presented in Table A.11 and Table A.12, support these observations. For the E2E dataset, neither scaling model yields feasible fits, indicating unreliable scaling behavior. However, DART (particularly under QAFactEval) and WikiTableText show stronger support for exponential scaling. A few exceptions persist—for instance, the FLAN-T5 model on WikiTableText favors power law scaling, as indicated by a red-marked cell. In summary, while no single scaling law appears to provide consistent behavior across all evaluated settings with respect to fine-tuning data size, the results broadly support exponential scaling as the model that more consistently fits the observed relationship between factual inconsistency and data size across LLMs.

Takeaway. *Exponential scaling generally provides a better fit than power-law scaling according to the evaluated criteria for characterizing how factual inconsistency in D2T tasks changes with increasing fine-tuning data size, despite some datasets (e.g., E2E) exhibiting aberrant or unstable behavior.*

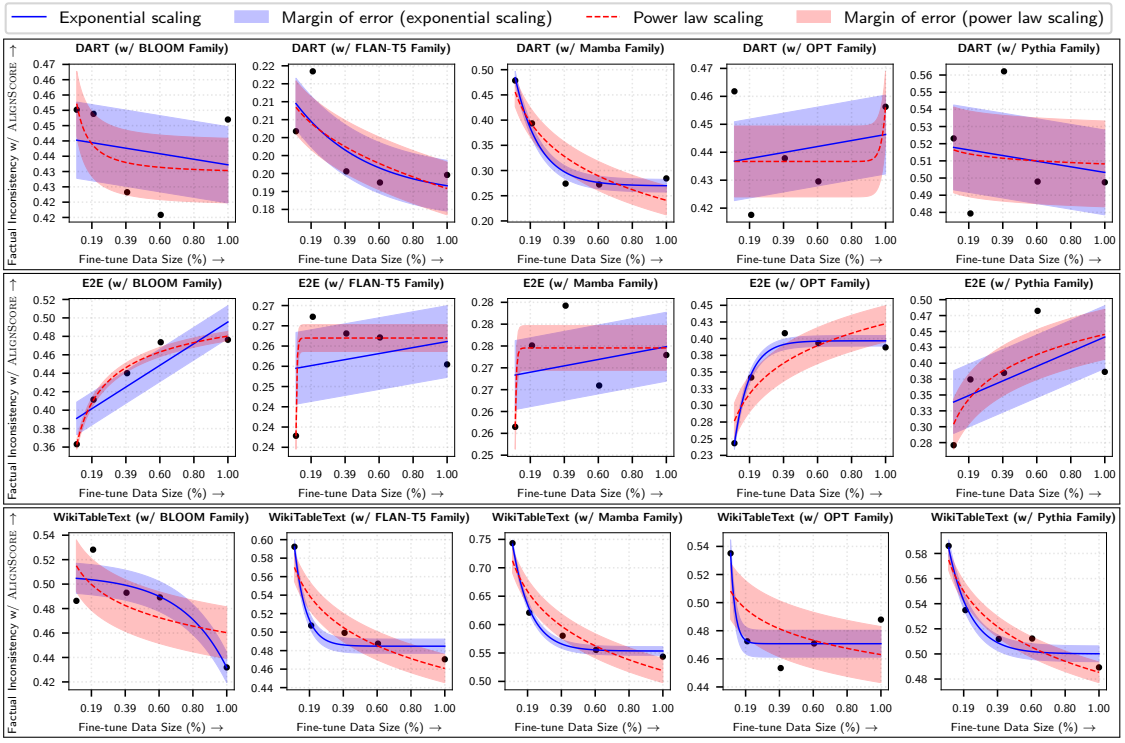


FIGURE A.10: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by ALIGNSCORE, plotted against fine-tune data size across three datasets—DART, E2E and WikiTableText.

TABLE A.11: Results from *VaCScal* corresponding to Figure A.10. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	BLOOM	FLAN-T5	Mamba	OPT	Pythia
DART	Exponential	4.17E+00/✗	2.00E-04/✗	1.20E+02/✓	2.95E-04/✗	1.39E-03/✗
	Power Law	7.72E+144/✗	1.24E-04/✗	2.90E-03/✗	3.16E-02/✗	4.60E-03/✗
E2E	Exponential	1.20E-04/✓	5.14E-06/✗	4.61E-05/✗	6.41E+01/✓	1.56E-03/✗
	Power Law	1.48E+00/✓	2.49E-05/✗	6.86E-04/✗	4.24E-02/✓	2.25E-03/✗
WikiTableText	Exponential	7.02E-04/✗	1.35E-03/✓	1.64E-03/✓	1.32E-04/✗	1.80E-04/✓
	Power Law	1.89E-01/✗	8.59E-05/✓	6.68E+00/✗	4.95E-03/✗	3.10E-02/✗

TABLE A.12: Results from *VaCScal* corresponding to Figure A.11. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	BLOOM	FLAN-T5	Mamba	OPT	Pythia
DART	Exponential	5.83E-04/✗	1.53E-05/✓	7.41E+01/✓	6.25E+00/✓	1.52E+01/✗
	Power Law	1.74E+147/✗	2.87E-03/✗	6.66E+163/✗	5.70E-04/✗	1.23E-02/✗
E2E	Exponential	3.40E-05/✗	3.38E+00/✗	9.86E+00/✗	1.54E+01/✗	5.85E-04/✗
	Power Law	2.00E+04/✗	3.47E-05/✗	4.59E+05/✗	4.32E-04/✗	6.65E-03/✗
WikiTableText	Exponential	3.72E-04/✗	4.82E-04/✓	2.22E-03/✓	4.03E-04/✗	6.34E-04/✓
	Power Law	8.04E-04/✗	3.13E-04/✓	2.13E-04/✗	6.41E-02/✗	1.41E-02/✗

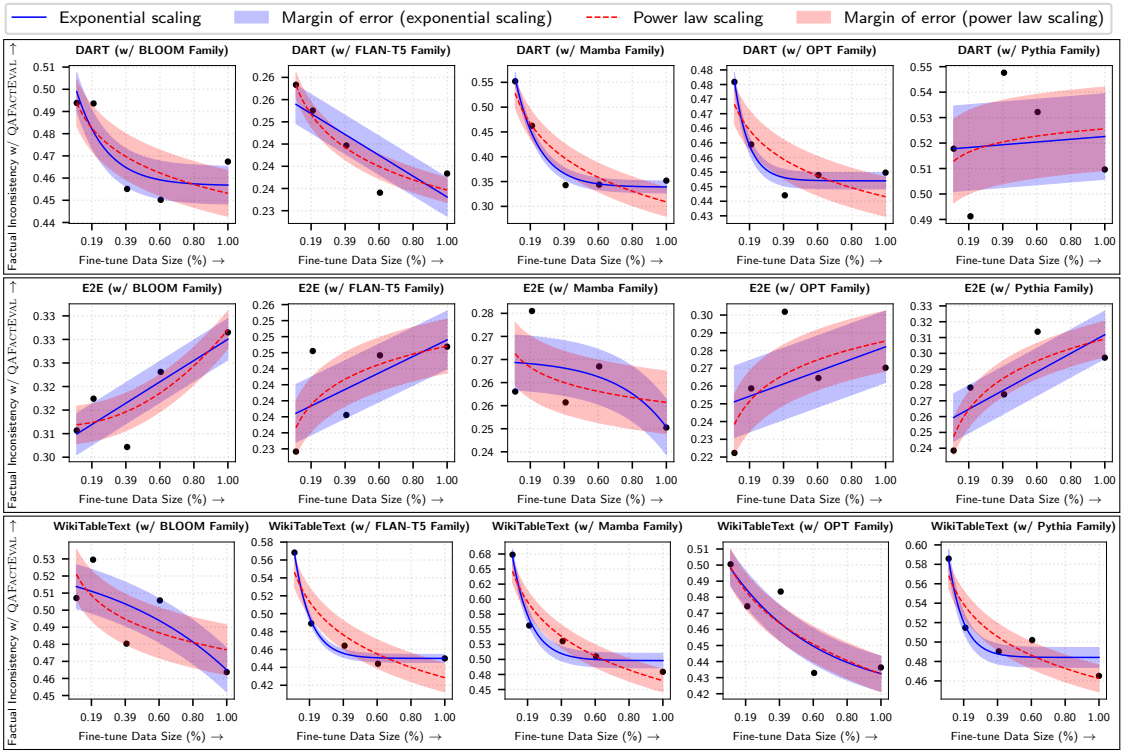


FIGURE A.11: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency, measured by QAFACTEVAL, plotted against fine-tune data size across three datasets—DART, E2E and WikiTableText.

A.7 Scaling Experiments with Full Fine-tuning

Unlike parameter-efficient fine-tuning methods such as QLoRA and prefix-tuning, full fine-tuning updates all model parameters, offering greater flexibility at the cost of substantially higher computational demands. Due to resource constraints, we perform full fine-tuning experiments on four LLM families—FLAN-T5, Mamba, Pythia, and Qwen2.5—across three D2T datasets: DART, E2E, and WikiTableText. Within the Qwen2.5 family, we consider only five models ranging from 0.5B to 14B parameters, as full fine-tuning of `Qwen2.5-32B` and `Qwen2.5-72B` exceeds our computational resource limits. We investigate scaling behavior with respect to both model size and computational budget (measured in FLOPs). For all outputs generated in these full fine-tuning experiments, we employ greedy decoding.

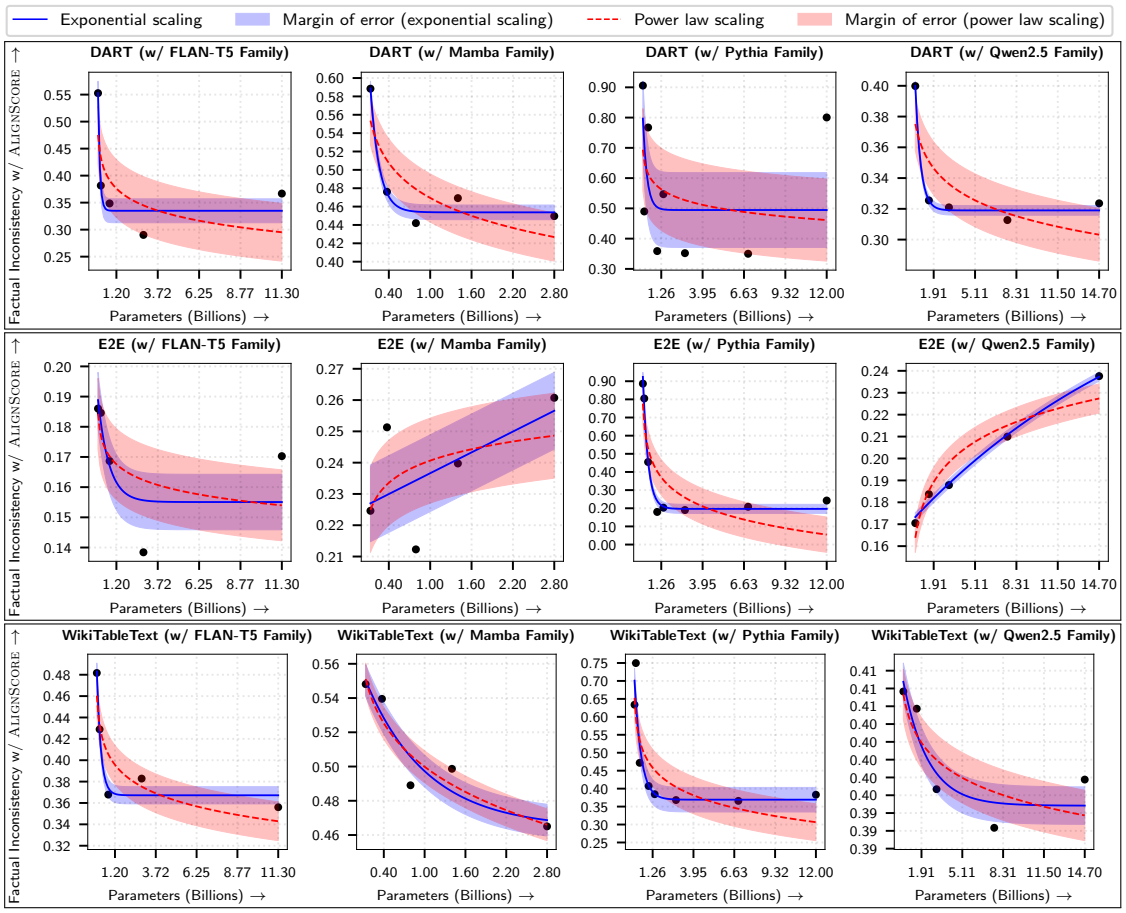


FIGURE A.12: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by AlignScore, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families.

A.7.1 Scaling Behavior Based on Model Size of Full Fine-Tuned LLM Families

The scaling results based on all four automatic metrics—AlignScore, QAFactEval, SummaC-conv, and UniEval-fact—across four LLM families (FLAN-T5, Mamba, Pythia, and Qwen2.5) are shown in Figure A.12, Figure A.13, Figure A.14, and Figure A.15. These results (especially in DART and WikiTableText) consistently suggest that exponential scaling outperforms power law scaling when model size is used as the scaling factor, as evidenced by its lower margin of error (MoE), indicating more stable and reliable fits to factual inconsistency trends in D2T tasks. This observation is further supported by the outputs of the *VaCScal* framework (Table A.13, Table A.14, Table A.15, and Table A.16), which consistently favor exponential scaling over power law scaling across all metrics. Similar to our findings with QLoRA, exponential scaling under full fine-tuning also exhibits strong support in Stage III of *VaCScal* (highlighted in blue), suggesting better empirical consistency. Compared to prefix-tuning, full fine-tuning shows greater alignment with exponential scaling. Although full fine-tuning yields fewer accepted fits in Stages II and III of *VaCScal* compared to QLoRA—exponential scaling still achieves significantly lower predictive error (Stage I) than power law scaling, further supporting its lower predictive error under the evaluated settings.

Table A.13: Results from *VaCScal* corresponding to Figure A.12, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	FLAN-T5	Mamba	Pythia	Qwen2.5
DART	Exponential	3.63E-03/✗	4.03E+01/✓	2.28E-02/✗	4.44E-04/✓
	Power Law	3.61E+11/✗	2.25E-04/✓	3.08E+02/✗	4.46E+17/✓
E2E	Exponential	6.22E-04/✗	2.28E+01/✗	6.29E-04/✓	1.46E-05/✓
	Power Law	1.26E-03/✗	2.94E-04/✗	3.64E+05/✗	1.30E-05/✓
WikiTableText	Exponential	1.22E+00/✓	2.61E-04/✗	6.12E-03/✓	9.80E-06/✗
	Power Law	1.21E-02/✗	5.18E+03/✗	8.45E+09/✗	2.43E+22/✗

Takeaway. *In the full fine-tuning setting, across all four widely used automatic metrics and LLM families (FLAN-T5, Mamba, Pythia, and Qwen2.5), exponential scaling suggests better agreement with how factual inconsistency decreases with model size, offering a more stable and practically reliable scaling model than power law scaling.*

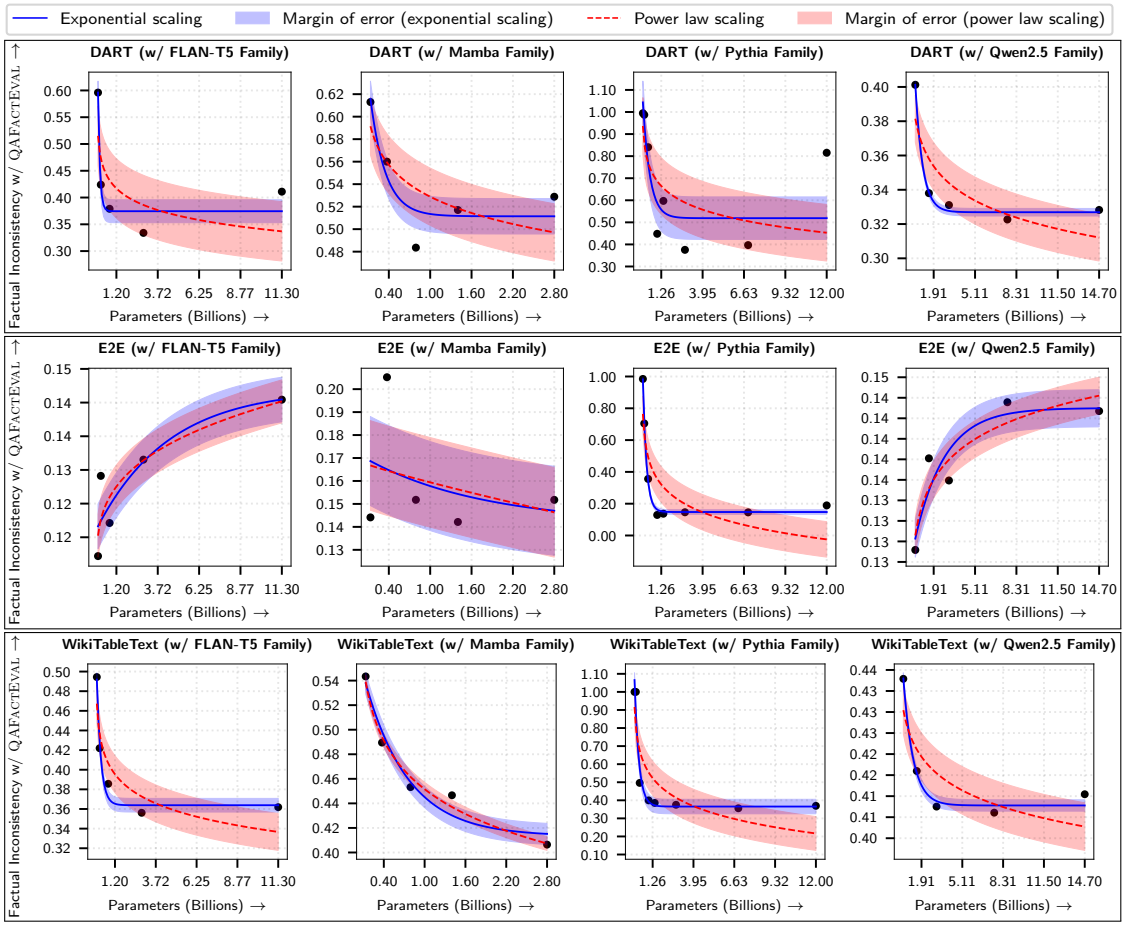


FIGURE A.13: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by QAFACTEVAL, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families.

TABLE A.14: Results from *VaCScal* corresponding to Figure A.13. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	FLAN-T5	Mamba	Pythia	Qwen2.5
DART	Exponential	1.12E-03/✗	1.05E-03/✗	2.02E-02/✗	5.48E-04/✓
	Power Law	7.57E+19/✗	1.48E-03/✗	2.71E-02/✗	3.05E-04/✓
E2E	Exponential	1.60E-05/✗	6.69E-03/✗	3.63E-04/✓	1.38E-05/✗
	Power Law	3.76E-04/✗	3.23E+26/✗	4.55E-02/✗	2.49E-04/✗
WikiTableText	Exponential	9.30E-04/✓	3.43E-04/✓	2.46E-02/✓	1.51E+01/✓
	Power Law	1.68E-03/✗	9.86E-06/✗	5.40E+09/✗	3.13E-05/✓

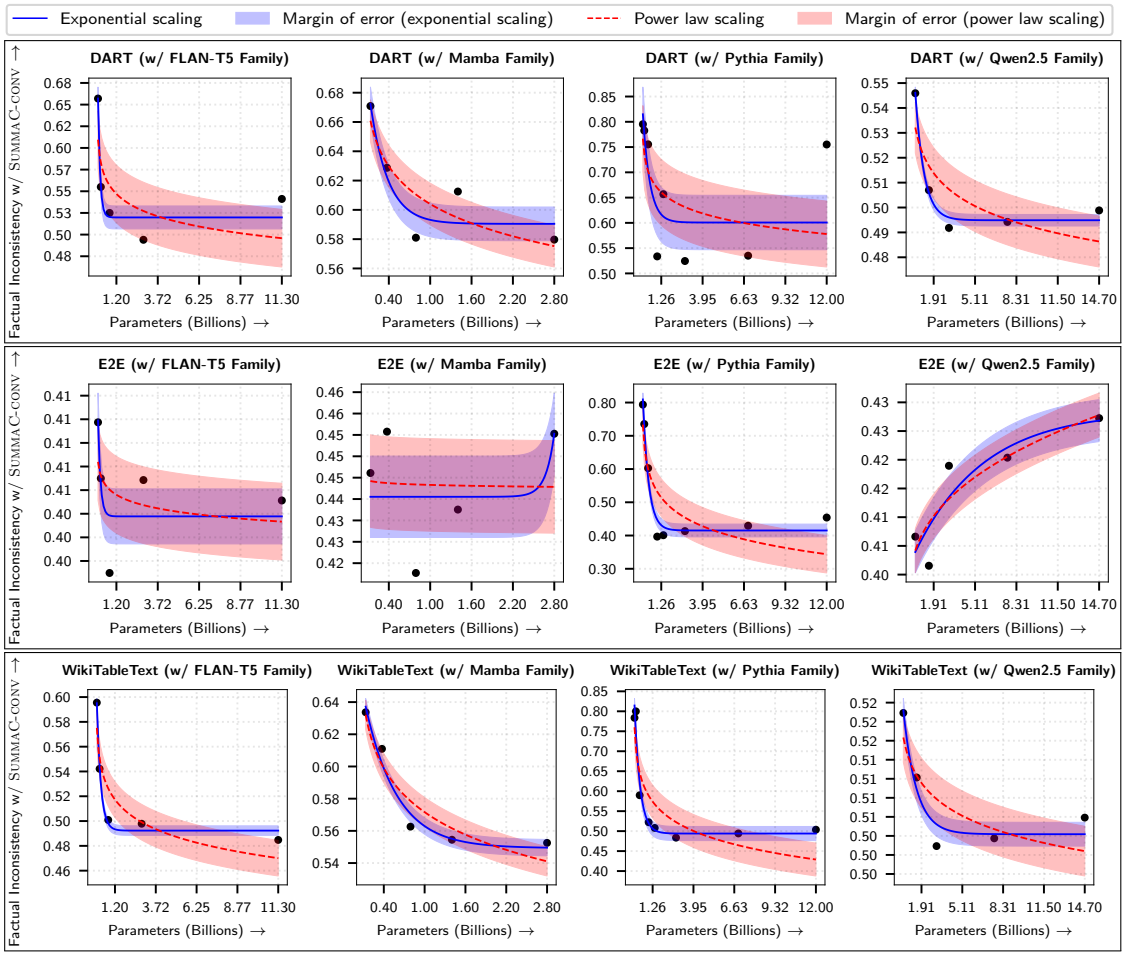


FIGURE A.14: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by SUMMAC-CONV, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families.

TABLE A.15: Results from *VaCScal* corresponding to Figure A.14. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	FLAN-T5	Mamba	Pythia	Qwen2.5
DART	Exponential	1.19E-03/✓	6.45E-04/✗	5.15E-03/✗	1.18E+02/✓
	Power Law	1.23E-03/✗	4.38E-04/✗	1.38E+00/✗	6.58E-05/✓
E2E	Exponential	2.03E-05/✗	3.17E+01/✗	1.08E-03/✓	2.07E-04/✗
	Power Law	2.76E-06/✗	5.18E-04/✗	9.26E+04/✗	8.35E-03/✗
WikiTableText	Exponential	3.79E-05/✓	1.34E-03/✓	3.26E-03/✓	3.17E+01/✓
	Power Law	1.26E-02/✗	8.00E+14/✗	6.00E+06/✗	1.03E-04/✗

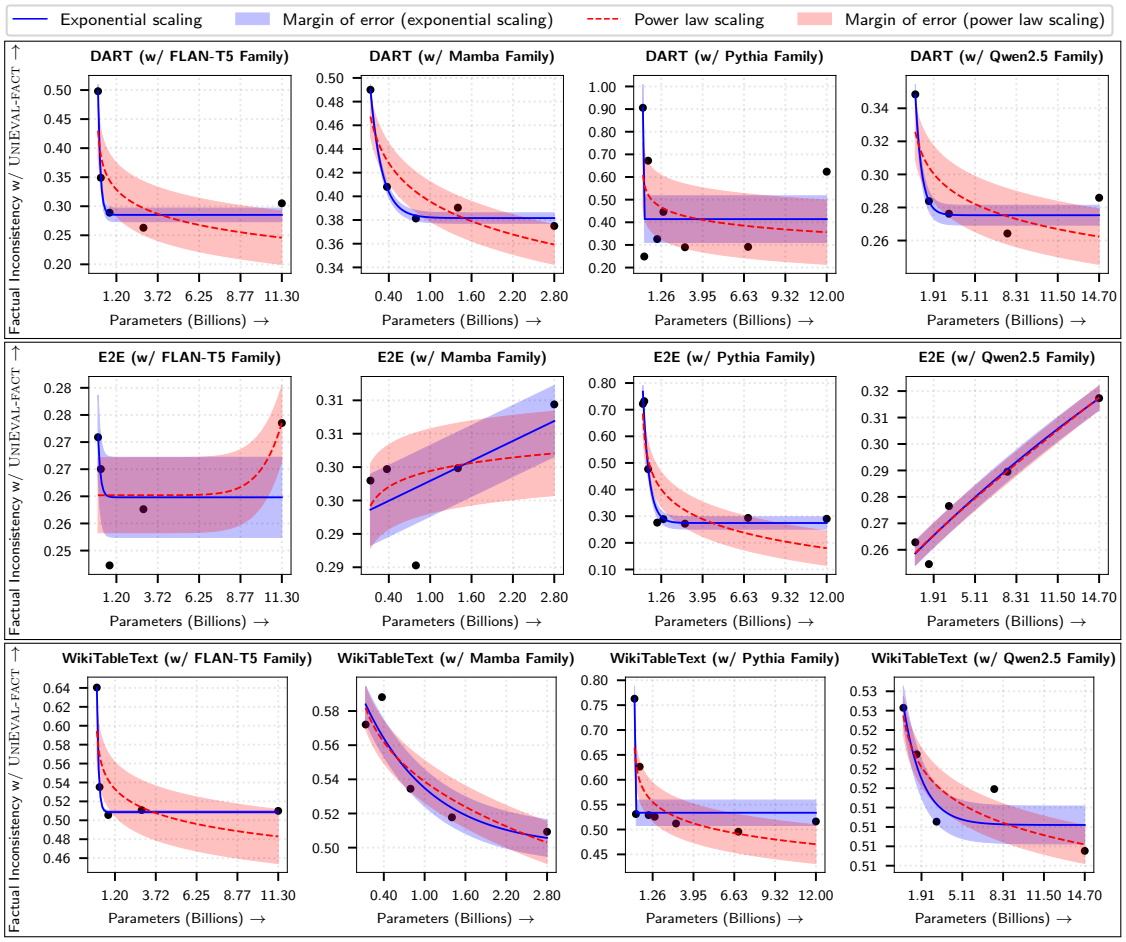


FIGURE A.15: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by UniEval-fact, plotted against model size across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families.

TABLE A.16: Results from *VaCScal* corresponding to Figure A.15. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	FLAN-T5	Mamba	Pythia	Qwen2.5
DART	Exponential	1.17E-03/✓	2.49E+01/✓	6.10E-02/✗	2.45E-04/✓
	Power Law	2.40E+19/✗	3.64E-03/✓	4.99E+89/✗	2.03E-04/✓
E2E	Exponential	1.43E-04/✗	9.19E-05/✗	1.79E-03/✓	1.06E-05/✓
	Power Law	1.02E-03/✗	1.36E-03/✗	5.24E+136/✗	3.18E-05/✓
WikiTableText	Exponential	1.28E-04/✓	9.08E-04/✗	3.63E-04/✓	1.02E+01/✗
	Power Law	1.32E-04/✓	1.13E-03/✗	1.38E-03/✓	3.04E+134/✗

A.7.2 Scaling Behavior Based on FLOPs of Full Fine-Tuned LLM Families

We present the scaling behavior of four LLM families—FLAN-T5, Mamba, Pythia, and Qwen2.5—on three D2T datasets (DART, E2E, and WikiTableText) based on model size, using two automatic factual consistency metrics: AlignScore and QAFactEval. Our analysis reveals that the scaling behavior with respect to FLOPs under full fine-tuning closely mirrors the trends observed with model size. As shown in Figure A.16 and Figure A.17, exponential scaling consistently outperforms power law scaling across all four LLM families in terms of lower margin of error (MoE), indicating more stable and reliable fits. The outcomes of the *VaCScal* framework (Table A.17 and Table A.18) provide additional empirical support for the exponential scaling formulation under the evaluated settings when FLOPs are used as the scaling factor.

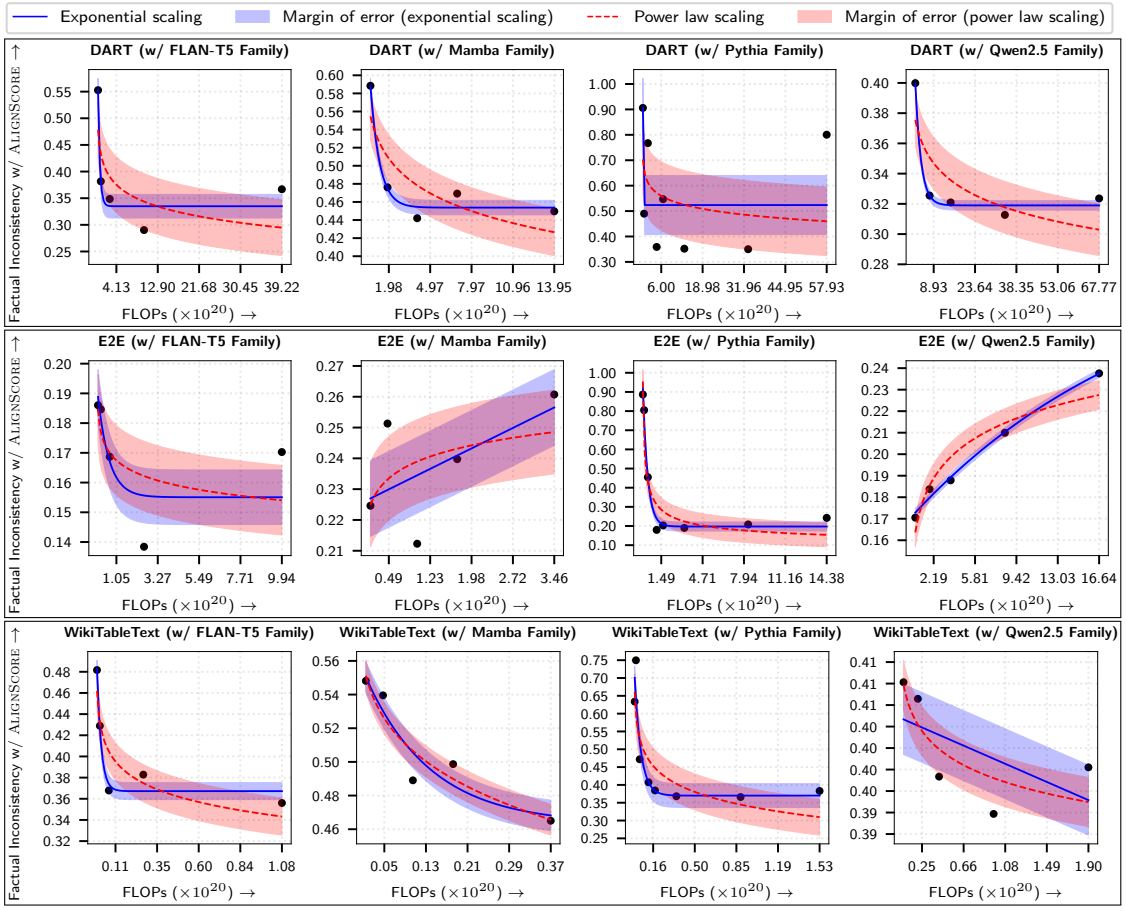


Figure A.16: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by AlignScore, plotted against FLOPs across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families.

Takeaway. *In the full fine-tuning setting, across all four widely used automatic metrics and LLM families (FLAN-T5, Mamba, Pythia, and Qwen2.5), exponential scaling*

TABLE A.17: Results from *VaCScal* corresponding to Figure A.16. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	FLAN-T5	Mamba	Pythia	Qwen2.5
DART	Exponential	3.36E-03/✗	1.96E-04/✓	8.98E+01/✗	4.58E-05/✓
	Power Law	3.30E+11/✗	8.52E-04/✓	4.16E-01/✗	4.65E-05/✓
E2E	Exponential	2.42E-05/✗	1.15E+01/✗	2.87E-03/✓	1.19E-05/✓
	Power Law	2.27E-04/✗	5.22E-04/✗	1.71E+137/✗	1.13E-05/✓
WikiTableText	Exponential	3.39E+00/✓	3.06E-04/✗	6.08E-03/✓	8.81E-05/✗
	Power Law	1.23E-02/✗	6.19E+03/✗	4.28E+07/✗	7.24E+22/✗

receives greater support as a model for the decline of factual inconsistency in D2T—with respect to FLOPs—based on ALIGNSCORE *and* QAFACTEVAL*, offering a more stable and practically reliable scaling model compared to power law scaling.*

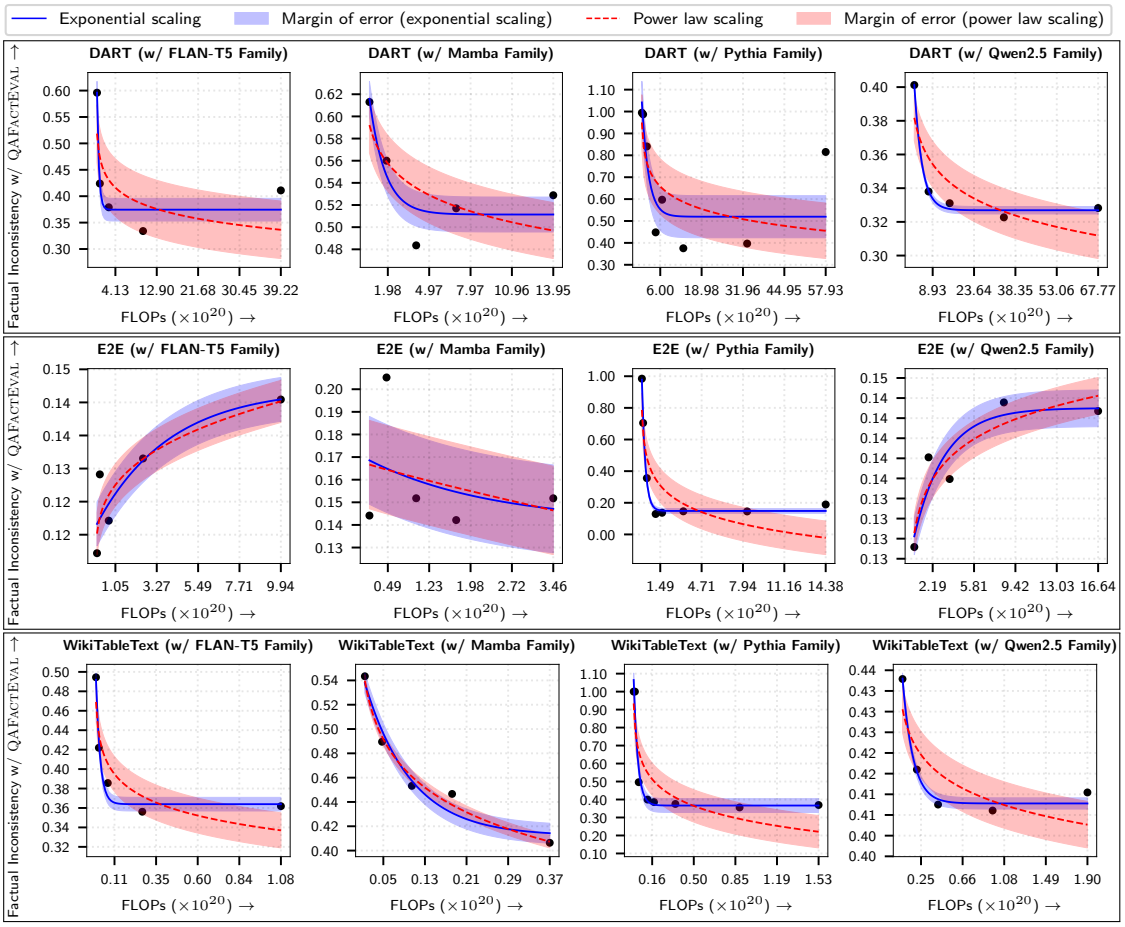


FIGURE A.17: Fitted power law and exponential scaling models (with 95% confidence intervals) for factual inconsistency in D2T, measured by QAFACTEVAL, plotted against FLOPs across full fine-tuned FLAN-T5, Mamba, Pythia and Qwen2.5 LLM families.

TABLE A.18: Results from *VaCScal* corresponding to Figure A.17. Numerical values, along with (✓/✗) and blue/red highlights, indicate the outcomes of Stages I, II, and III of *VaCScal*.

Dataset	Scaling	FLAN-T5	Mamba	Pythia	Qwen2.5
DART	Exponential	3.63E-03/✗	4.82E-04/✗	1.86E-02/✗	3.23E-05/✓
	Power Law	1.47E-03/✗	5.92E-04/✗	2.84E-02/✗	5.59E-05/✓
E2E	Exponential	2.24E-05/✗	1.54E-04/✗	1.59E+02/✓	1.03E-05/✗
	Power Law	3.56E-04/✗	1.22E+27/✗	3.53E+133/✗	7.00E-06/✗
WikiTableText	Exponential	8.80E-05/✓	3.81E-04/✓	2.52E-02/✓	2.63E-04/✓
	Power Law	6.07E-03/✗	1.51E-01/✗	1.17E+09/✓	1.08E-05/✓

A.8 On the Aberrant Scaling Patterns of E2E and ViGGO

Throughout our experiments, we consistently observed aberrant behavior from the E2E dataset, and to a lesser extent, the ViGGO dataset. Specifically, factual inconsistency in these datasets tends to increase with scaling factors such as model size, FLOPs, and fine-tuning data size—contrary to expected trends. Although we could not definitively identify the root cause of this behavior, we hypothesize that the low lexical diversity and closed-domain nature of these datasets may be contributing factors. Limited diversity can lead to overfitting during fine-tuning, resulting in degraded output quality and increased factual inconsistency. Additionally, the closed-domain nature of the E2E and ViGGO datasets (see Table 6.1) may hinder LLMs from effectively leveraging their pre-trained knowledge. To support this hypothesis, we present the type-token ratio (TTR), computed using the following formula, for all five datasets used in our study:

$$\text{TTR} = \frac{\text{number of unique words or tokens}}{\text{total number of words and tokens}}$$

A lower TTR indicates reduced lexical diversity. As shown in Table A.19, the E2E and ViGGO datasets exhibit the lowest lexical richness—both in the input source data and the corresponding references—compared to the other three datasets.

TABLE A.19: Type-Token Ratio (TTR) for all five D2T datasets, calculated separately for input source data and corresponding reference texts.

	DART	E2E	ViGGO	WebNLG	WikiTableText
TTR (of source-input)	0.0178	0.0001	0.0029	0.0057	0.0621
TTR (of reference text)	0.03	0.0058	0.0297	0.0209	0.1297

Bibliography

[1] Linyong Nan, Dragomir R. Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. DART: open-domain structured data record to text generation. In *Proceedings of the NAACL-HLT*, pages 432–447, 2021. URL https://doi.org/10.18653/v1/2021.naacl-main.37.

[2] Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. Table-to-text: Describing table region with natural language. In *Proceedings of the AAAI*, pages 5020–5027, 2018. URL https://doi.org/10.1609/aaai.v32i1.11944.

[3] Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech and Language*, 59:123–156, 2020. URL https://doi.org/10.1016/j.csl.2019.06.009.

[4] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Proceedings of the EMNLP*, pages 1203–1213, 2016. URL https://doi.org/10.18653/v1/d16-1128.

[5] Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. Challenges in data-to-document generation. In *Proceedings of the EMNLP*, pages 2253–2263, 2017. URL https://doi.org/10.18653/v1/d17-1239.

[6] Michela Lorandi and Anya Belz. High-quality data-to-text generation for severely under-resourced languages with out-of-the-box large language models. In *Proceedings of the EACL Findings*, pages 1451–1461, 2024. URL https://aclanthology.org/2024.findings-eacl.98.

[7] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. The webnlg challenge: Generating text from RDF data. In *Proceedings of the INLG*, pages 124–133, 2017. URL https://doi.org/10.18653/v1/w17-3518.

[8] Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the SIGDIAL*, pages 201–206, 2017. URL https://doi.org/10.18653/v1/w17-5525.

[9] Anja Belz. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455, 2008. URL https://doi.org/10.1017/S1351324907004664.

[10] David L. Chen and Raymond J. Mooney. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of the ICML*, pages 128–135, 2008. URL https://doi.org/10.1145/1390156.1390173.

[11] Adela Schneider, Andreas Madsack, Johanna Heininger, Ching-Yi Chen, and Robert Weißgraeber. Data-to-text systems as writing environment. In *Proceedings of the In2Writing*, pages 1–10, Dublin, Ireland, 2022. URL https://aclanthology.org/2022.in2writing-1.1.

[12] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the EMNLP*, pages 1711–1721, 2015. URL https://doi.org/10.18653/v1/d15-1199.

[13] Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. Sticking to the facts: Confident decoding for faithful data-to-text generation. *CoRR*, abs/1910.08684, 2019. URL http://arxiv.org/abs/1910.08684.

[14] Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, Elena Lloret, Elena Simona Apostol, Ciprian-Octavian Truica, Branislava Sandrih, Sanda Martincic-Ipsic, Gábor Berend, Albert Gatt, and Grazina Korvel. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207, 2022. URL https://doi.org/10.1613/jair.1.12918.

[15] Wei Er Luo, Xi Yue, and Chao Tie Zhong. Data-to-text generation with data control and multi-loss fusion. In *Proceedings of the FAIML*, pages 244–247, 2023. URL https://doi.org/10.1145/3616901.3616995.

[16] Ratish Puduppully and Mirella Lapata. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527, 2021. URL https://doi.org/10.1162/tacl_a_00381.

[17] Shujie Li, Liang Li, Ruiying Geng, Min Yang, Binhua Li, Guanghu Yuan, Wanwei He, Shao Yuan, Can Ma, Fei Huang, et al. Unifying structured data as graph for data-to-text pre-training. *Transactions of the Association for Computational Linguistics*, 12:210–228, 2024. ISSN 2307-387X. URL https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00641/2346090/tacl_a_00641.pdf.

[18] Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. STAR: boosting low-resource information extraction by structure-to-text data generation with large language models. In *Proceedings of the AAAI*, pages 18751–18759, 2024. URL https://doi.org/10.1609/aaai.v38i17.29839.

[19] Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992. URL https://aclanthology.org/J92-4003.pdf.

[20] Steffen Bickel, Peter Haider, and Tobias Scheffer. Predicting sentences using n-gram language models. In *Proceedings of the EMNLP-HLT*, pages 193–200, 2005. URL https://aclanthology.org/H05-1025/.

[21] Anja Belz. Probabilistic generation of weather forecast texts. In *Proceedings of the NAACL-HLT*, pages 164–171, 2007. URL https://aclanthology.org/N07-1021/.

[22] Wei Lu and Hwee Tou Ng. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the EMNLP*, pages 1611–1622, 2011. URL https://aclanthology.org/D11-1149/.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[24] Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. In *Proceedings of the ICASSP*, pages 6645–6649, 2013. URL https://doi.org/10.1109/ICASSP.2013.6638947.

[25] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997. URL https://doi.org/10.1017/S1351324997001502.

[26] Sidi Lu, Yaoming Zhu, Weinan Zhang, Jun Wang, and Yong Yu. Neural text generation: Past, present and beyond. *CoRR*, abs/1803.07133, 2018. URL http://arxiv.org/abs/1803.07133.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT*, pages 4171–4186, 2019. URL https://doi.org/10.18653/v1/n19-1423.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. URL https://doi.org/10.48550/arXiv.2302.13971.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the NeurIPS*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[31] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the ACL*, pages 7871–7880, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.703.

[32] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. URL https://doi.org/10.48550/arXiv.2310.06825.

[33] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang

Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. URL https://doi.org/10.48550/arXiv.2412.15115.

[34] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023. URL https://doi.org/10.1145/3571730.

[35] Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449, 2024. URL https://doi.org/10.1109/TKDE.2023.3304385.

[36] Çaglar Gülçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. Pointing the unknown words. In *Proceedings of the ACL*, 2016. URL https://doi.org/10.18653/v1/p16-1014.

[37] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the ACL*, pages 1073–1083, 2017. URL https://doi.org/10.18653/v1/P17-1099.

[38] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the ACL*, pages 2673–2679, 2019. URL https://doi.org/10.18653/v1/p19-1256.

[39] Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the ACL*, pages 4884–4895, 2019. URL https://doi.org/10.18653/v1/p19-1483.

[40] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. In *Proceedings of the ACL*, pages 889–898, 2018. URL https://aclanthology.org/P18-1082/.

[41] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. In *Proceedings of the ACL*, pages 1638–1649, 2018. URL https://aclanthology.org/P18-1152/.

[42] Min Zhang, Jianfeng He, Shuo Lei, Murong Yue, Linhan Wang, and Chang-Tien Lu. Can llm find the green circle? investigation and human-guided tool manipulation for compositional generalization. In *Proceedings of the ICASSP*, pages

11996–12000, 2024. URL https://www.doi.org/10.1109/ICASSP48485.2024.10446355.

[43] Dan Hendrycks. Introduction to ai safety, ethics, and society. *Taylor & Francis*, forthcoming. URL www.aisafetybook.com.

[44] Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the ACL*, pages 1072–1086, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.101.

[45] Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. Evaluating factual accuracy in complex data-to-text. *Computer Speech and Language*, 80:101482, 2023. URL https://doi.org/10.1016/j.csl.2023.101482.

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the ICML*, pages 2048–2057, 2015. URL http://proceedings.mlr.press/v37/xuc15.html.

[47] Heng Gong, Xiaocheng Feng, Bing Qin, and Ting Liu. Table-to-text generation with effective hierarchical encoder on three dimensions (row, column and time). In *Proceedings of the EMNLP*, pages 3141–3150, 2019. URL https://doi.org/10.18653/v1/D19-1310.

[48] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the ACL*, pages 1906–1919, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.173.

[49] Saar Hommes, Chris van der Lee, Felix J. Clouth, Jeroen K. Vermunt, Xander Verbeek, and Emiel Krahmer. A personalized data-to-text support tool for cancer patients. In *Proceedings of the INLG*, pages 443–452, 2019. URL https://aclanthology.org/W19-8656/.

[50] Ruslan Yermakov, Nicholas Drago, and Angelo Ziletti. Biomedical data-to-text generation via fine-tuning transformers. In *Proceedings of the INLG*, pages 364–370, 2021. URL https://doi.org/10.18653/v1/2021.inlg-1.40.

[51] Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. Making effective use of healthcare data using data-to-text technology. *Data science for healthcare: methodologies and applications*, pages 119–145, 2019. URL https://doi.org/10.1007/978-3-030-05249-2_4.

[52] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018. URL https://doi.org/10.1613/jair.5477.

[53] Anya Belz, Simon Mille, and David M. Howcroft. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the INLG*, pages 183–194, 2020. URL https://doi.org/10.18653/v1/2020.inlg-1.24.

[54] Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *CoRR*, abs/2203.05227, 2022. URL https://doi.org/10.48550/arXiv.2203.05227.

[55] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25 (70):1–53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

[56] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the COLM*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

[57] Dimitra Gkatzia. Content selection in data-to-text systems: A survey. *CoRR*, abs/1610.08375, 2016. URL http://arxiv.org/abs/1610.08375.

[58] Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. Findings of the E2E NLG challenge. In *Proceedings of the INLG*, pages 322–328, 2018. URL https://doi.org/10.18653/v1/w18-6539.

[59] Chris van der Lee, Emiel Krahmer, and Sander Wubben. Automated learning of templates for data-to-text generation: Comparing rule-based, statistical and neural methods. In *Proceedings of the INLG*, pages 35–45, 2018. URL https://doi.org/10.18653/v1/w18-6504.

[60] Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, and Lei Li. Variational template machine for data-to-text generation. In *Proceedings of the ICLR*, 2020. URL https://openreview.net/forum?id=HkejNgBtPB.

[61] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the AAAI*, pages 4881–4888, 2018. URL https://doi.org/10.1609/aaai.v32i1.11925.

[62] Juraj Juraska, Kevin Bowden, and Marilyn A. Walker. Viggo: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the INLG*, pages 164–172, 2019. URL https://aclanthology.org/W19-8623/.

[63] Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. Order-planning neural text generation from structured data. In *Proceedings of the AAAI*, pages 5414–5421, 2018. URL https://doi.org/10.1609/aaai.v32i1.11947.

[64] Chris Kedzie and Kathleen R. McKeown. Controllable meaning representation to text generation: Linearization and data augmentation strategies. In *Proceedings of the EMNLP*, pages 5160–5185, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.419.

[65] Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn A. Walker. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the NAACL-HLT*, pages 152–162, 2018. URL https://doi.org/10.18653/v1/n18-1014.

[66] Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Proceedings of the EMNLP Findings*, pages 2223–2236, 2020. URL https://doi.org/10.18653/v1/2020.findings-emnlp.202.

[67] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the EMNLP*, pages 9332–9346, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-main.750.

[68] Saad Obaid ul Islam, Iza Skrjanec, Ondrej Dusek, and Vera Demberg. Tackling hallucinations in neural chart summarization. In *Proceedings of the INLG*, pages 414–423, 2023. URL https://doi.org/10.18653/v1/2023.inlg-main.30.

[69] Joy Mahapatra and Utpal Garain. Divergence-aware selective data-to-text generation with llms for factual consistency. In *Pattern Recognition and Machine Intelligence*, pages 564–571, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-18480-1.

[70] Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng-Ann Heng, and Wai Lam. Unveiling the generalization power of fine-tuned large language models. *CoRR*, abs/2403.09162, 2024. URL https://doi.org/10.48550/arXiv.2403.09162.

[71] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024. ISSN 2666-6510. URL https://www.sciencedirect.com/science/article/pii/S2666651023000141.

[72] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. URL https://doi.org/10.48550/arXiv.2211.05100.

[73] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the ACL*, pages 328–339, 2018. URL https://aclanthology.org/P18-1031/.

[74] Wilson L. Taylor. 'cloze procedure': A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953. ISSN 0022-5533. URL https://doi.org/10.1177/107769905303000401.

[75] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[76] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[77] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the NeurIPS*, pages 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

[78] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the NAACL-HLT*, pages 2227–2237, 2018. URL https://doi.org/10.18653/v1/n18-1202.

[79] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of the ICLR*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

[80] Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, NUI Galway, 2019. URL https://aran.library.nuigalway.ie/bitstream/handle/10379/15463/neural_transfer_learning_for_nlp.pdf.

[81] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *Proceedings of the ICLR*, 2016. URL http://arxiv.org/abs/1511.06391.

[82] Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Proceedings of the NeurIPS*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/59404fb89d6194641c69ae99ecdf8f6d-Abstract-Conference.html.

[83] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data only. In *Proceedings of the NeurIPS*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/fa3ed726cc5073b9c31e3e49a807789c-Abstract-Datasets_and_Benchmarks.html.

[84] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[85] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Proceedings of the*

NeurIPS, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

[86] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the ACL/IJCNLP*, pages 4582–4597, 2021. URL https://doi.org/10.18653/v1/2021.acl-long.353.

[87] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):195:1–195:35, 2023. URL https://doi.org/10.1145/3560815.

[88] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311–318, 2002. URL https://aclanthology.org/P02-1040/.

[89] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. URL https://aclanthology.org/W05-0909/.

[90] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. URL https://aclanthology.org/W04-1013.

[91] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. In *Proceedings of the ACL*, pages 7881–7892, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.704.

[92] Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the ACL*, pages 3544–3552, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.326.

[93] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232, 2023. URL https://doi.org/10.48550/arXiv.2311.05232.

[94] Joy Mahapatra and Utpal Garain. A comprehensive performance evaluation of llms for data-to-text generation and divergence-weighted training. *ACM Transactions on Intelligent Systems and Technology*, 17(2), 2026. ISSN 2157-6904. URL https://doi.org/10.1145/3795137.

[95] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the EMNLP*, pages 632–642. The Association for Computational Linguistics, 2015. URL https://doi.org/10.18653/v1/d15-1075.

[96] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. URL https://doi.org/10.1162/tacl_a_00453.

[97] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the ACL*, pages 3214–3252, 2022. URL https://doi.org/10.18653/v1/2022.acl-long.229.

[98] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the EMNLP*, pages 9004–9017, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-main.557.

[99] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. URL https://doi.org/10.48550/arXiv.2205.01068.

[100] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023. URL https://doi.org/10.48550/arXiv.2309.01219.

[101] Gabor Angeli, Percy Liang, and Dan Klein. A simple domain-independent probabilistic approach to generation. In *Proceedings of the EMNLP*, pages 502–512, 2010. URL https://aclanthology.org/D10-1049/.

[102] Joy Mahapatra, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. Statistical natural language generation from tabular non-textual data. In *Proceedings of the INLG*, pages 143–152. The Association for Computer Linguistics, 2016. URL https://doi.org/10.18653/v1/w16-6624.

[103] Adam Pauls and Dan Klein. Faster and smaller n-gram language models. In *Proceedings of the ACL-HLT*, pages 258–267, 2011. URL https://aclanthology.org/P11-1027/.

[104] Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the EACL*, 2006. URL https://aclanthology.org/E06-1040/.

[105] Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3), 2018. URL https://doi.org/10.1162/coli_a_00322.

[106] Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the ACL*, pages 643–653, 2018. URL https://aclanthology.org/P18-1060/.

[107] Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of the Interspeech*, pages 1045–1048, 2010. URL https://doi.org/10.21437/Interspeech.2010-343.

[108] Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *Proceedings of the NAACL-HLT*, pages 720–730, 2016. URL https://doi.org/10.18653/v1/n16-1086.

[109] Preksha Nema, Shreyas Shetty, Parag Jain, Anirban Laha, Karthik Sankaranarayanan, and Mitesh M. Khapra. Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In *Proceedings of the NAACL-HLT*, pages 1539–1550, 2018. URL https://doi.org/10.18653/v1/n18-1139.

[110] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of the ACL/IJCNLP*, pages 1577–1586, 2015. URL https://doi.org/10.3115/v1/p15-1152.

[111] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the ICLR*, 2015. URL http://arxiv.org/abs/1409.0473.

[112] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the EMNLP*, pages 1412–1421, 2015. URL https://doi.org/10.18653/v1/d15-1166.

[113] Feng Nie, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. An encoder with non-sequential dependency for neural data-to-text generation. In *Proceedings of the INLG*, pages 141–146, 2019. URL https://aclanthology.org/W19-8619/.

[114] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI*, pages 6908–6915, 2019. URL https://doi.org/10.1609/aaai.v33i01.33016908.

[115] Joy Mahapatra and Utpal Garain. Exploring structural encoding for data-to-text generation. In *Proceedings of the INLG*, pages 404–415. Association for Computational Linguistics, 2021. URL https://doi.org/10.18653/v1/2021.inlg-1.44.

[116] Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When LLM meets domain experts. In *Proceedings of the NeurIPS*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1190733f217404edc8a7f4e15a57f301-Abstract-Datasets_and_Benchmarks.html.

[117] Joy Mahapatra and Utpal Garain. A comprehensive study on factual consistency in llms for data-to-text generation. *Sādhanā*, 50(3):196, 2025. URL https://doi.org/10.1007/s12046-025-02860-5.

[118] Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity and quality in natural language generation. *CoRR*, abs/2004.10450, 2020. URL https://arxiv.org/abs/2004.10450.

[119] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the ACL/IJCNLP*, pages 7319–7328, 2021. URL https://doi.org/10.18653/v1/2021.acl-long.568.

[120] Joy Mahapatra, Soumyajit Roy, and Utpal Garain. Exponential scaling of factual inconsistency in data-to-text generation with fine-tuned LLMs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=xPaPd6g5WG.

[121] Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the EACL*, pages 2734–2744, 2021. URL https://doi.org/10.18653/v1/2021.eacl-main.236.

[122] Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *Proceedings of the EMNLP*, pages 670–680, 2014. URL https://doi.org/10.3115/v1/d14-1074.

[123] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. URL http://arxiv.org/abs/1411.2539.

[124] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017. URL https://doi.org/10.1109/TPAMI.2016.2598339.

[125] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the CVPR*, pages 770–778, 2016. URL https://doi.org/10.1109/CVPR.2016.90.

[126] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the EMNLP*, pages 670–680, 2017. URL https://doi.org/10.18653/v1/d17-1070.

[127] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the NeurIPS*, pages 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[128] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the ICLR*, 2015. URL http://arxiv.org/abs/1412.6980.

[129] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[130] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov,

Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. URL https://doi.org/10.48550/arXiv.2307.09288.

[131] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *Proceedings of the ICLR*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

[132] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the EMNLP*, pages 563–578, 2019. URL https://doi.org/10.18653/v1/D19-1053.

[133] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In *Proceedings of the NeurIPS*, pages 27263–27277, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html.

[134] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *CoRR*, abs/2211.09110, 2022. URL https://doi.org/10.48550/arXiv.2211.09110.

[135] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the NeurIPS*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[136] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In

Proceedings of the EMNLP, pages 6449–6464, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-main.397.

[137] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the EMNLP*, pages 38–45, 2020. URL https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[138] Zdenek Kasner, Ekaterina Garanina, Ondrej Plátek, and Ondrej Dusek. Tabgenie: A toolkit for table-to-text generation. In *Proceedings of the ACL*, pages 444–455, 2023. URL https://doi.org/10.18653/v1/2023.acl-demo.42.

[139] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[140] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the ACL*, pages 1383–1392, 2018. URL https://aclanthology.org/P18-1128/.

[141] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the ACL*, pages 157–165. Association for Computational Linguistics, 2022. URL https://doi.org/10.18653/v1/2022.acl-short.18.

[142] Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models. In *Proceedings of the ICLR*. OpenReview.net, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

[143] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the WMT*, pages 186–191, 2018. URL https://doi.org/10.18653/v1/w18-6319.

[144] Zdenek Kasner and Ondrej Dusek. Beyond reference-based metrics: Analyzing behaviors of open llms on data-to-text generation. *CoRR*, abs/2401.10186, 2024. URL https://doi.org/10.48550/arXiv.2401.10186.

[145] Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the NAACL-HLT*, pages 2587–2601, 2022. URL https://doi.org/10.18653/v1/2022.naacl-main.187.

[146] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the ACL*, pages 11328–11348, 2023. URL https://doi.org/10.18653/v1/2023.acl-long.634.

[147] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Proceedings of the NIPS*, pages 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

[148] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

[149] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. URL https://doi.org/10.48550/arXiv.2203.15556.

[150] David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham M. Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=1Avb4jYjLb. Featured Certification.

[151] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the EMNLP*, pages 3045–3059, 2021. URL https://doi.org/10.18653/v1/2021.emnlp-main.243.

[152] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *Proceedings of the ICLR*, 2024. URL https://openreview.net/forum?id=5HCnKDeTws.

[153] Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. Llms on the line: Data determines loss-to-loss scaling laws. *CoRR*, abs/2502.12120, 2025. URL https://doi.org/10.48550/arXiv.2502.12120.

[154] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory F. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep

learning scaling is predictable, empirically. *CoRR*, abs/1712.00409, 2017. URL http://arxiv.org/abs/1712.00409.

[155] Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. Data scaling laws in NMT: the effect of noise and architecture. In *Proceedings of the ICML*, pages 1466–1482, 2022. URL https://proceedings.mlr.press/v162/bansal22b.html.

[156] Elias Frantar, Carlos Riquelme Ruiz, Neil Houlsby, Dan Alistarh, and Utku Evci. Scaling laws for sparsely-connected foundation models. In *Proceedings of the ICLR*, 2024. URL https://openreview.net/forum?id=i9K2ZWkYIP.

[157] Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *CoRR*, abs/2405.18392, 2024. URL https://doi.org/10.48550/arXiv.2405.18392.

[158] Alexander B. Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *CoRR*, abs/2405.00592, 2024. URL https://doi.org/10.48550/arXiv.2405.00592.

[159] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the NAS*, 121(27):e2311878121, 2024. URL https://www.pnas.org/doi/abs/10.1073/pnas.2311878121.

[160] Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Proceedings of the NeurIPS*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6fcb1afcc1e9c2c82c8ddddf03bcf0f6-Abstract-Conference.html.

[161] Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, and et al. Language models scale reliably with over-training and on downstream tasks. In *Proceedings of the ICLR*, 2025. URL https://openreview.net/forum?id=iZeQBqJamf.

[162] S Weisberg. *Applied Linear Regression*. John Wiley & Sons, 2005. URL https://www.doi.org/10.1002/0471704091.

[163] Andrew F Siegel. *Practical Business Statistics*. Academic Press, 2016. URL https://doi.org/10.1016/C2015-0-00463-4.

[164] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: journal of the Econometric Society*, pages 307–333, 1989. URL https://doi.org/10.2307/1912557.

[165] S Shaphiro and MBJB Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965. URL https://doi.org/10.2307/2333709.

[166] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the ICML*, pages 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

[167] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *CoRR*, 2020. URL https://arxiv.org/abs/2101.00027.

[168] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[169] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the EMNLP*, pages 2023–2038, 2022. URL https://doi.org/10.18653/v1/2022.emnlp-main.131.

[170] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. URL https://scipy.org/.

[171] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A

survey on in-context learning. In *Proceedings of the EMNLP*, pages 1107–1128. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.emnlp-main.64.

[172] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[173] Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23:9:1–9:34, 2022. URL https://jmlr.org/papers/v23/20-1111.html.