Understanding Humorous Contradictions in Comics: A Multi-Agent Debate Approach

Suparna Das

Understanding Humorous Contradictions in Comics: A Multi-Agent Debate Approach

Dissertation submitted in partial fulfillment of the requirements for the degree of

Master of Technology
in
Cryptology and Security

by

Suparna Das
[Roll No: Crs2323]

under the guidance of

Malay Bhattacharyya
Associate Professor
Machine Intelligence Unit, Indian Statistical Institute

Anirban Mukhopadhyay
Professor
Department of Computer Science and Engineering, University Of Kalyani



Indian Statistical Institute
Kolkata – 700108, India

July 2025

CERTIFICATE

This is to certify that the Project titled **"Enhancing LLM's Power To Detect Humorous Context Through Multi-agent Debate"** submitted by **Suparna Das** to Indian Statistical Institute, Kolkata, in partial fulfillment for the award of the degree of **Master of Technology in Cryptology and Security** is a bonafide record of work carried out by her under my supervision and guidance. The project has fulfilled all the requirements as per the regulations of this Institute. The contents of this project, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Dr. Malay Bhattacharyya
Associate Professor, Machine Intelligence Unit
Indian Statistical Institute, Kolkata



Dr. Anirban Mukhopadhyay
Professor, Department of Computer Science and Engineering
Director, Internal Quality Assurance Cell (IQAC)
University of Kalyani

Acknowledgement

I would like to convey my great appreciation to my supervisors, Dr. Malay Bhattacharyya and Prof. Anirban Mukhopadhyay, for their guidance and encouragement throughout the duration of this project. I had a great curiosity toward the domain of LLM and Agentic AI but lacked knowledge and experience to tackle problems of my own. Dr. Bhattacharyya and Prof. Mukhopadhyay have aided me throughout the project and helped me immensely enter into this new domain that I always sought for. With his guidance I now have a deep appreciation for the field and the research sphere as a whole.

My deepest thanks to the faculties of Indian Statistical Institute, for their support. Special thanks to Bibhas da for his constant support throughout this journey. Thank you for keeping me motivated with your generous support.

Last but not the least, I would like to thank all my family, friends and peers for their continuous help and support. Finally I would like to thank all those, whom I have missed out from the above list.

Suparna Das
Roll No. CrS2323
Indian Statistical Institute
Kolkata – 700108, India.

Abstract

Recent advancements in large language models (LLMs) have demonstrated impressive reasoning capabilities, yet their ability to understand and generate nuanced humor—particularly in contexts involving juxtaposition and contradictory narratives—remains a significant challenge. This thesis explores the application of advanced reasoning techniques, specifically **multi-agent debate**, to enhance LLMs' performance in understanding and interpreting humor derived from juxtaposed comic panels. .

Leveraging the YesBut benchmark—a dataset designed to evaluate AI models on tasks ranging from literal description generation to deep narrative reasoning, we explore how multi-agent debate can address key limitations in current models, such as **visual misinterpretation, reasoning gaps, and hallucination**. By simulating collaborative critique among multiple AI agents, our approach encourages more robust reasoning, enabling models to better resolve contradictions, infer underlying philosophies, and generate coherent, contextually appropriate humor.

Our experiments demonstrate that multi-agent debate not only improves accuracy in humor comprehension tasks but also enhances the model's ability to generalize across diverse creative expressions. This work advances the development of socially intelligent AI systems capable of nuanced, human-like understanding of humor and creative narratives. The findings highlight the potential of multi-agent reasoning frameworks to bridge the gap between AI and human-level humor interpretation, paving the way for more sophisticated applications in automated content generation, entertainment, and human-AI interaction.

Contents

Acknowledgement **i**

Abstract **ii**

List of Tables **v**

List of Figures **vi**

1 Introduction **1**

2 Preliminary Studies **4**

2.1 The YESBUT dataset 4

2.2 Prior works on comic understanding 5

2.3 Benchmark results 5

3 Background and Related Tools **7**

3.1 Brief introduction on Agentic AI 7

3.1.1 Key features of Agentic AI 8

3.1.2 Traditional AI vs Agentic AI 9

3.2 Multiagent System: 9

3.2.1 Key characteristics of multi-agent systems 10

4 Methodology **12**

4.1 Motivation and Limitations of Prior Approaches 12

4.2 System Overview 13

4.3 System Configuration 14
4.3.1 Input format and preprocessing 15
4.3.2 Agents' roles and configuration 15
4.3.3 Corrective Feedback Method: 16
4.4 Complete workflow to execute the tasks 17
4.4.1 Generation Tasks 17
4.4.2 Title matching task 17
4.4.3 Underlying philosophy selection task 18
4.5 Evaluation Strategy 18

5 Experimental Results **19**

6 Conclusion and Future Work **21**

6.1 Future Work 21

List of Tables

2.1 Performance of various models on four comic understanding tasks: Literal Description, Contradiction, Philosophy, and Title Matching. 6

3.1 Comparison between Traditional AI and Agentic AI. 9

5.1 Comparison of evaluation metrics for each comic understanding task with and without GPT-4. 19

List of Figures

1.1 An instance of the YESBUT dataset for comic understanding of juxtaposed comic panels, with two panels; 'Yes' and 'But'. (Comic by Anton Gudim). . . . 2

3.1 The workflow of a single agentic system, ... 9

3.2 The workflow of a multiagentic system. 10

4.1 Workflow involving image input, vision-language model (VLM), two agents, and a refiner for output synthesis. 14

Chapter 1

Introduction

Comics are a unique blend of visual art and narrative that encapsulate a wide range of human experiences and emotions. They are intricate puzzles of visual narratives that merge images, text, and spatial relationships to convey layered meanings. While books depend entirely on written language and films usually unfold in a straightforward sequence, comics frequently employ fragmented, yet interlinked frames to build meaning through deliberate contrasts. This demands advanced analytical skills to connect together the implied associations.

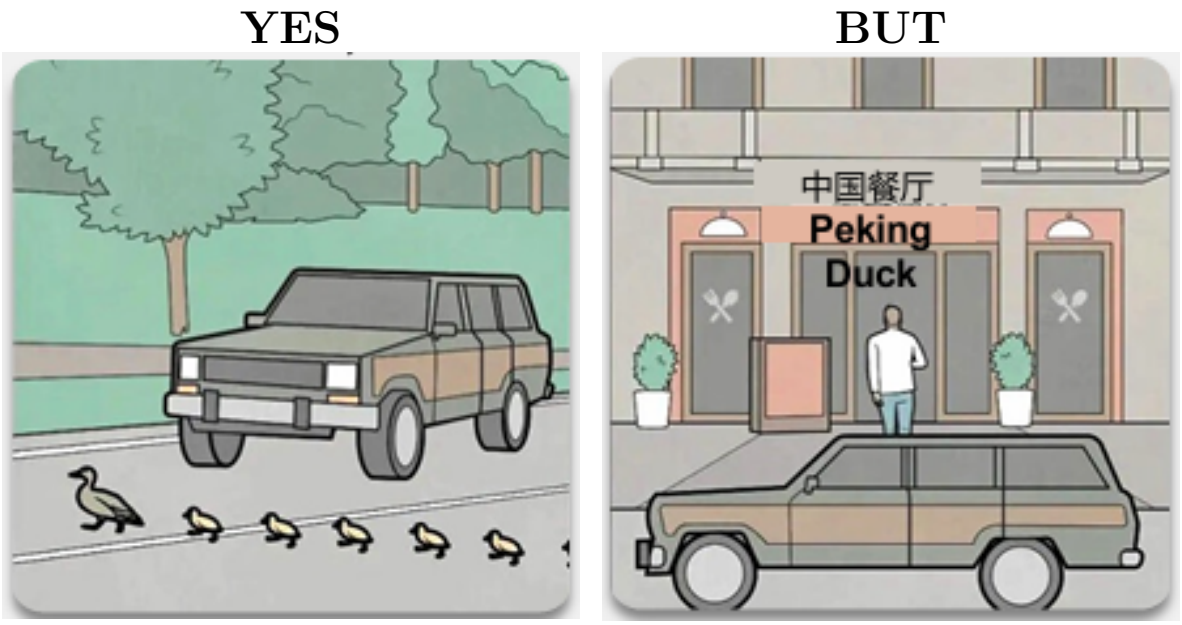
Human readers intuitively grasp this non-linear format, thanks to their innate capacity for comparative thinking and shared cultural context. However, existing AI systems designed to process both images and texts, still fall short in achieving this level of sophisticated interpretation. Current vision based language models face significant challenges in replicating the human ability to analyze layered comic narratives through comparative reasoning.

Understanding how artificial intelligence interprets comics represents a critical step forward in developing machines with deeper social awareness and meaning comprehension [2]. Comics, as a fundamental form of human artistic expression, provide unique perspectives into emotional nuances and cultural frameworks. Mastering this capability is vital for building AI systems with genuine social intelligence and fostering machine-driven creativity. Such advancements would directly benefit practical applications, including personalized content suggestions and AI-assisted creative tools, ultimately leading to more engaging user interactions.

Previous studies have applied vision language models (VLMs) to understand humor and deep semantics. However, these studies often focus on single-panel comics and do not investigate the more complex case of non-linear narratives created through juxtaposition, a fundamental element in comics. Juxtaposition involves placing two contrasting elements together to provoke thought or evoke humor. Moreover, we previously examined that VLMs' ability to understand comics, specifically focusing on humor derived from juxtaposition. The goal was to determine whether large models can accurately comprehend the complex and contradictory narratives present in comics. Such contradictions challenge conventional semantic interpretations and demand a deeper analysis. To examine this, they have introduced an annotated benchmark named "YESBUT" [4].

Each comic is annotated with a literal description, a contradiction illustration, the underlying philosophy it reveals or satirizes, and a title that summarizes the overall narrative, as shown

in figure 1.1.



Narrative Contradiction

The contradiction lies in the driver's display of compassion for live ducks by stopping and waiting for them to cross the road, yet this same respect does not extend to their treatment as food.

Underlying Philosophy Selection

✓ It highlights the contradictory relationship humans have with animals, where the value of an animal's life changes drastically based on context.
✗ The comic suggests there is no need to show empathy to the animals since they will be eventually consumed as food.

Title Matching

✓ Crossing Paths: From Compassion to Cuisine
✗ Food Empathy vs. Animal Protection

Figure 1.1: An instance of the YESBUT dataset for comic understanding of juxtaposed comic panels, with two panels; 'Yes' and 'But'. (Comic by Anton Gudim).

We then propose four tasks: (1) **literal description writing**– to produce a surface description of the comic narrative; (2) **contradiction generation**– where the model illustrates the narrative contradiction; (3) **underlying philosophy selection**– which targets at selecting the correct philosophy the comic reflects; and (4) **title matching**– where the model matches the comic with a proper title.

These tasks jointly cover different levels of comic understanding, from literal content comprehension to more in-depth narrative reasoning, providing a thorough evaluation of comic

understanding capabilities.

These techniques had been applied on a single model instance [2]. Instead, we propose an advanced technique through multi-agent settings, where multiple language model instances (or agents) individually propose and jointly debate their responses and reasoning processes to arrive at a single common answer. More specifically, given a query, multiple instances of a language model first generate individual candidate responses to a query. Then each individual model instance reads and critiques the responses of all other models and uses this content to update its own answer. This step is then repeated over several rounds. This process induces models to construct answers that are consistent with both their internal critic, as well as sensible in light of the responses of other agents. The resulting quorum of models can hold and maintain multiple chains of reasoning and possible answers simultaneously before proposing the final answer.

We observe that our debate approach outperforms single model baselines on the above 4 tasks. Using both multiple model agents and multiple rounds of debate is important to achieve the best performance. Given an initial query, after debating and examining the responses of other model instances, we find that the population almost always converges on a single and more accurate common answer. Debate results are also less likely to include false facts that models are internally uncertain of. This is because as the debate progresses, individual model instances tend to disagree on uncertain facts and omit them from the answer. Debate does not just act to amplify one correct answer in a model quorum; we find many cases where all the models initially make incorrect predictions, but then arrive at the correct answer as debate progresses.

Previously, some works mainly related to mathematical problem solving had been done through the debate system [1]. In this paper, we propose a methodology to make this multi-agent debate system, comprehensively used on the YESBUT dataset. We also measured the BERT score and ROUGE-2 score for literal description and contradiction generation tasks, and for philosophy selection and title matching tasks, we measured the accuracy score.

Chapter 2

Preliminary Studies

In this chapter, we will briefly go over the analysis done prior to the experiments.

2.1 The YESBUT dataset

The YESBUT dataset is a benchmark designed to evaluate the AI understanding of comics that exhibit humorous contradictions through juxtaposed narratives. Each sample in the dataset includes:

- A two-panel comic with contradictory storytelling, mostly sourced from Anton Gudim's "YES, BUT" series.
- A literal description narrating the visual elements.
- An explanation of the contradiction between the two panels.
- The underlying philosophy or message the comic conveys.
- A set of title options, including one correct and several plausible distractors.

The comics were scraped from social media (primarily Twitter and Pinterest), and after filtering out inappropriate, multi-panel, or duplicate samples, the final dataset includes 348 two-panel comics.

Each comic was annotated through a human-AI collaborative process involving GPT-4 and human editors. This ensured a high-quality and semantically consistent dataset. The annotations underwent multiple quality control phases including cross-verification by different annotators to minimize bias and subjectivity.

This dataset supports multiple tasks designed to probe different levels of comic understanding, ranging from literal comprehension to abstract reasoning, such as contradiction identification, philosophy selection, and title matching.

2.2 Prior works on comic understanding

Zhe Hu and Tuo Liang have introduced new techniques in their paper [2] to deal with juxtaposed images where they evaluated the model's performance in a zero-shot manner using both recent VLMs and LLMs. For VLMs, the comic image and the questions are both provided as inputs for output prediction. They included both commercial models such as GPT-4 and Claude-3, as well as open-source models including LLaVa, CogVLM, Qwen-VL, mPLUG-Owl2, and InstructBLIP.

In case of LLMs, since they cannot directly process images, they used the LLaVa-1.6 13B which generated literal descriptions as input due to its strong performance. They also included ChatGPT, Llama3, and Mistral 7B instruction model.They evaluated a wide range of commercial and open-source models on the full YESBUT dataset, which consists of 348 comics, across several tasks including literal description, contradiction generation, title matching, and underlying philosophy selection.

Table 2.1 summarizes the performance of selected models on those tasks. The relatively low score for certain models on ambiguous examples highlights the challenge of reasoning about humorous contradictions, especially for two panelled comic images in the YESBUT dataset — **a limitation we aim to address through a multi-agent debate approach**.

2.3 Benchmark results

They conducted comprehensive experiments on the YESBUT dataset, evaluted on both commercial and open sourced VLM models. From the result what they have shown from there the indication we got that commercial VLM's outperform their open sourced counterparts on most tasks. For VLMs, the original image is directly used as input, while for LLMs the generated comic description by a VLM (LLaVa-1.6 13B model) is used as a input.

For the philosophy and title understanding tasks, which are formulated as multiple-choice question answering, they used accuracy as the evaluation metric. For generation tasks including literal description and contradiction, they applied reference-based evaluation metrics commonly used in text generation studies, and reported ROUGE-2 (recall) and BERT Score (recall).

To complement automated assessment methods, which have inherent constraints in evaluating generated text quality, they incorporated manual evaluation by human experts. Three qualified evaluators were engaged to score system outputs using a 5-point Likert scale (1=lowest,5=highest) for two key tasks; for narrative description assessment, drawing from established methodology, They examined:

- Accuracy - Whether the generated text properly reflects the comic's storyline
- Thoroughness - If all significant narrative components are included
- Consistency - Verification that all statements align with visual content, eliminating fabricated elements

Setting	Model	Literal Description			Contradiction			Philosophy	Title
		BERT	R-2	GPT	BERT	R-2	GPT	Accuracy	Accuracy
VLMs	GPT-4	**88.32**	**87.46**	**3.76**	**87.64**	**83.21**	**4.03**	82.76	**60.25**
	Claude-3	87.68	80.30	3.28	86.93	80.63	3.79	**84.10**	56.42
	LLaVA-1.6-34B	86.45	67.62	2.86	84.06	75.95	3.51	78.83	**63.31**
	LLaVA-1.6-13B	81.34	75.95	2.69	85.66	76.19	3.16	69.16	55.08
	LLaVA-1.5-13B	78.77	58.21	2.51	86.48	67.57	3.36	66.22	48.75
	InstructBlip-13B	85.20	35.28	2.86	84.90	73.28	3.29	65.02	47.15
	CogVLM	80.80	55.51	2.65	87.07	69.31	3.10	63.10	49.52
	Qwen-VL-Chat	79.03	51.18	2.53	86.90	64.41	2.91	64.47	42.20
	mPlug-Owl2	78.26	47.38	2.57	86.08	55.09	2.59	61.27	43.10
	LLaVA-1.6-7B	80.71	70.36	2.79	86.71	63.76	2.87	59.80	39.83
	InstructBlip-7B	76.02	38.02	2.60	86.32	66.29	2.85	25.86	26.44
LLMs	ChatGPT	-	-	-	**87.78**	67.42	3.54	75.86	49.52
	Llama-3-8B-Instruct	-	-	-	87.41	70.52	3.59	72.13	49.71
	Mistral-7B-Instruct	-	-	-	87.01	67.70	3.64	66.00	45.98

Table 2.1: Performance of various models on four comic understanding tasks: Literal Description, Contradiction, Philosophy, and Title Matching.

They achieved 84.1% accuracy for underlying philosophy selection task by commercial VLM claude-3 and achieved 63.31% accuracy for title matching task by LLaVa-1.6-34B model. For the Literal description generation tasks GPT-4 achieved the highest BERT score 88.32% and ROUGE-2 (R-2) score 87.46% respectively. For contradiction generation task Chatgpt(commercial LLM) reported the highest BERT score 87.78% and GPT-4 reported the highest ROUGE-2 score 83.21%.

Chapter 3

Background and Related Tools

Despite their tremendous success in many applications and the real world fields, recent literature has highlighted the limitations of large language models (LLMs) in achieving consistent and reliable reasoning, particularly in tasks that require planning, strategic thinking, and modeling of complex environments or social dynamics. One notable position is presented by the LAW framework, which advocates for a more structured view of machine reasoning by integrating Language Models, Agent Models, and World Models. This perspective emphasizes deliberate, human-like reasoning — grounded in beliefs, goals, consequences, and inter-agent interactions — as essential for building more robust and versatile AI systems [3].

This aligns closely with the goals of our multi-agent debate framework, where agents simulate deliberate reasoning and critique each other's inferences to arrive at more accurate decisions in ambiguous or contradictory contexts.

In this chapter, we explore the foundational concepts of agentic AI and multi-agent systems, with a particular focus on how coordinated agent behaviors can simplify complex tasks and enhance reasoning capabilities.

3.1 Brief introduction on Agentic AI

Agentic AI refers to a class of artificial intelligence systems that operate with autonomy, intentionality, and adaptability, much like human agents. These systems are designed to pursue specific goals through a combination of planning, decision-making, and action execution, often with minimal or no human intervention. Unlike traditional **reactive systems**, which only respond to external stimulations like rule-based chatbots or prompt-only LLMs, agentic AI exhibits **proactive behavior** — initiating tasks, monitoring progress, handling obstacles, and dynamically revising its strategies to optimize outcomes.

At the core of agentic AI is the concept of goal-directed reasoning. An agent is given a high-level objective and is expected to autonomously determine:

- How to achieve it (via planning),
- When to act (via context-sensitive decision-making),

- What tools or resources to use (via environment modeling or tool to use), and

- How to adapt if conditions change (via feedback loops or memory)

3.1.1 Key features of Agentic AI

Agentic AI systems are characterized by a set of core capabilities that distinguish them from traditional reactive or narrowly supervised AI. These features collectively enable such systems to operate intelligently, flexibly, and with purpose in complex environments.

- **Autonomy:** Agentic AI operates independently, without requiring step-by-step human instruction. Once provided with a high-level goal, the agent can plan and execute actions on its own, making decisions dynamically based on context.

- **Proactive Goal Pursuit:** Unlike passive AI systems that respond only to prompts, agentic AI actively pursues its assigned goals. It identifies subgoals, formulates action plans, monitors progress, and adapts its strategy to fulfill its objectives.

- **Memory and Context Awareness:** An essential aspect of agentic reasoning is the ability to maintain and update internal memory or state. This includes both short-term context (current task status) and long-term context (past interactions or accumulated knowledge), enabling coherent, multi-step reasoning.

- **Tool Use and Multimodal Interaction:** Agentic systems can extend their capabilities by integrating external tools such as APIs, search engines, code interpreters, or databases. In some cases, they also support multimodal input/output, such as processing images, text, or even actions in physical environments.

- **Adaptivity and Feedback Integration:** Agentic AI systems can modify their behavior based on feedback, environmental changes, or internal evaluation. This allows them to recover from failure, optimize performance, and generalize across varying tasks — a trait that is vital for open-ended problem-solving.

3.1.2 Traditional AI vs Agentic AI

Feature	Traditional AI	Agentic AI
Initiative	Reactive (responds to prompts)	Proactive (plans/acts independently)
Scope	Narrow, single-task	Broad, multi-step workflows
Flexibility	Fixed behavior	Adapts to dynamic environments
Example	ChatGPT answering a question	AutoGPT researching a topic end-to-end

Table 3.1: Comparison between Traditional AI and Agentic AI.

3.2 Multiagent System:

A Multi-Agent System (MAS) is a decentralized computational framework comprising multiple autonomous agents that interact, collaborate, or compete to achieve individual or shared goals. These systems are particularly effective in solving complex problems that exceed the capacity of a single agent, by leveraging parallelism, specialization, and distributed reasoning.

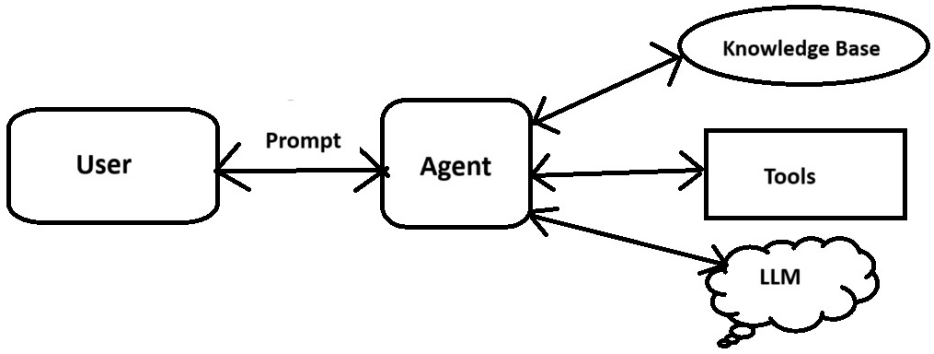


Figure 3.1: The workflow of a single agentic system,

Figure 3.1 illustrates the architecture of a traditional single-agent system, where one model handles the entire decision-making process. In contrast, Figure 3.2 depicts a multi-agent setup, in which distinct agents are assigned different roles or sub-tasks. These agents may communicate

and coordinate with each other to improve task performance, efficiency, and robustness. A Supervisor Agent may be integrated to oversee the interaction among agents, manages task flow, and ensures that the debate process adheres to the defined protocol without bias or interruption.

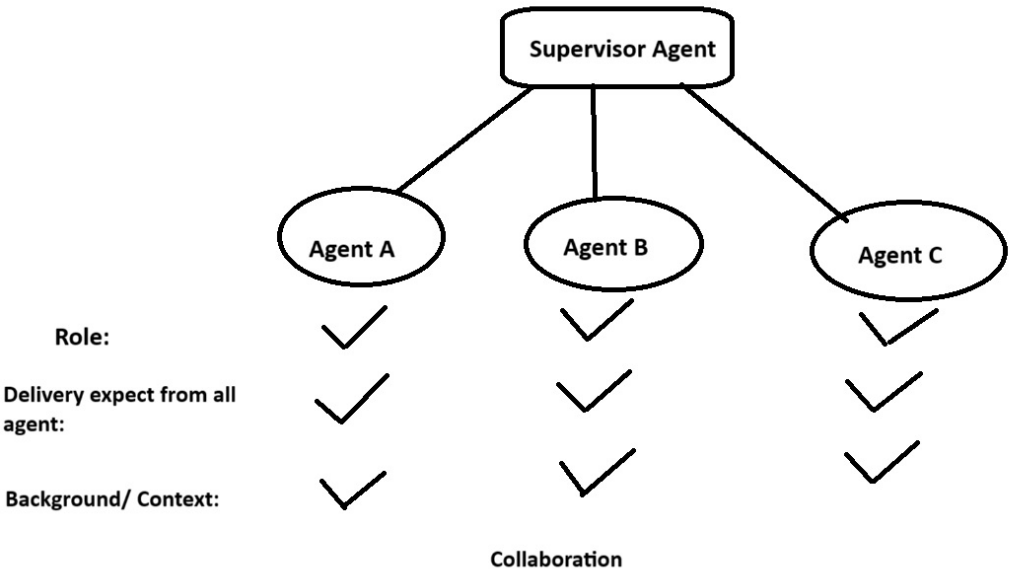


Figure 3.2: The workflow of a multiagentic system.

3.2.1 Key characteristics of multi-agent systems

Multi-agent systems are defined by several fundamental characteristics that distinguish them from single-agent or centralized AI architectures. These features enable them to operate effectively in dynamic, and complex environments:

- **Autonomy:** Each agent within a multi-agent system operates independently, making decisions based on its own internal state, knowledge, and goals. Agents are not micromanaged by external entities; rather, they are capable of initiating actions, responding to environmental stimuli, and pursuing objectives without direct human supervision. This autonomy allows for flexible and scalable system behavior, especially when agents encounter novel or evolving scenarios.
- **Decentralization:** Unlike centralized systems that rely on a single controller to manage the overall behavior, multi-agent systems are inherently decentralized. Control and decision-making are distributed among the agents, and global behavior emerges from the interactions between them. This decentralization improves system robustness, fault tolerance, and scalability, as the system can continue functioning even if some agents fail or operate with partial information.
- **Interaction:** Agents in a multi-agent system are designed to interact with one another to coordinate, collaborate, or even compete, depending on the nature of the task. These interactions may take the form of message passing, negotiation, consensus-building, or

conflict resolution. Effective communication is crucial in achieving coherent system behavior, especially when agents have overlapping or interdependent goals.

- **Heterogeneity:** Multi-agent systems often consist of heterogeneous agents — that is, agents with different roles, capabilities, decision-making strategies, or even goals. This diversity allows for specialization and division of labor, where each agent contributes to the overall system performance based on its strengths. Heterogeneity also enables the system to handle complex, multi-faceted tasks that require different perspectives or skills.

Chapter 4

Methodology

In this chapter, we describe the methodology and experimental setup used in our study. The primary focus is on the design and implementation of a multi-agent debate framework for understanding humorous contradictions in comics. We present the system architecture, agent roles, input processing steps, implementation of corrective feedback between two debaters, and evaluation strategy. The experiments are structured to demonstrate the effectiveness of agent collaboration in selecting appropriate comic titles, as compared to traditional single-model approaches.

An overview of agentic AI and the multi-agent debate system — which forms the backbone of our proposed workflow, has already been discussed in Chapter 3. In this chapter, we focus specifically on the implementation and impact of the corrective feedback mechanism between debating agents.

4.1 Motivation and Limitations of Prior Approaches

Recent work on comic understanding, particularly the YESBUT benchmark introduced by Huang et al.(2024) [2], has explored the ability of large language models (LLMs) to interpret humorous contradictions through tasks such as literal description generation, contradiction generation, moral inference, and title matching. These tasks require not just surface-level understanding but reasoning over juxtaposed perspectives embedded in two-panel comic structures.

In the YESBUT study, the authors evaluated the performance of several commercial and open-source LLMs — including GPT-4, Claude, Gemini and Mistral — on a dataset of 349 annotated comics. While top-tier models like GPT-4 and Claude showed relatively high performance, particularly in literal description and contradiction generation, their capabilities fell short in more nuanced tasks such as title selection and philosophical interpretation. These tasks involve abstract reasoning and subtle contextual understanding, which proved to be challenging even for state-of-the-art models.

A key limitation of these prior approaches is their reliance on single-model architectures, where a solitary LLM is expected to both analyze input and produce a final answer in isolation. Such

setups tend to suffer from:

- Overconfidence in wrong answers,
- Lack of self-correction or reflection, and
- Hallucination in justifications for choices, especially when options are semantically close or subtly misleading.

Moreover, without any form of interactive reasoning or multi-perspective evaluation, these systems often fail to detect inconsistencies or challenge their own conclusions. This is particularly problematic in tasks like title matching, where all options may appear superficially plausible, and success often hinges on detecting fine-grained contradictions between form and function, appearance and usability, or expectation and reality.

These limitations point to a critical need for **deliberative reasoning** — a process where competing viewpoints can be proposed, challenged, and refined. Inspired by human debate and collaborative problem-solving, our proposed method introduces a **multi-agent debate framework**, where two distinct agents independently evaluate a comic and its visual captions, then exchange justifications and refine their reasoning through corrective feedback. A third agent, the **refiner**, acts as a neutral judge, synthesizing the outputs and deciding the final answer.

This multi-agent setup not only addresses the lack of internal critique found in single-agent systems, but also encourages more robust decision-making through **perspective diversity, iterative refinement, and meta-level judgment**. In the following sections, we detail the components of this system, the debate protocol, and the experimental setup used to evaluate its effectiveness.

4.2 System Overview

To address the limitations identified in previous works, We construct an entire pipeline for the multiagent debate system along with VLM and agents with tools and LLM. For the entire work, we have used Microsoft **AutoGen** framework and **OpenAI API key**. The architecture consists of three core agents:

- **Debater A (based on GPT-3.5 Turbo)**,
- **Debater B (based on GPT-4 Turbo)**,
- **Refiner Agent (based on GPT-4 Turbo)**.

The system is designed to handle multiple-choice tasks—specifically, title selection for comics exhibiting visual or conceptual contradictions. Given a literal description of a comic and a set of four title options, each debater independently selects what they believe is the best title, accompanied by justification. The agents then exchange their reasoning and, if designed to do so, revise their answers based on critique. The refiner agent receives the final responses from both debaters and evaluates which one offers a more logically sound and semantically aligned argument.

This architecture draws inspiration from human deliberation: it encourages perspective-taking, error correction, and reasoning through structured disagreement. Unlike single-agent pipelines,

It promotes robustness through redundancy, disagreement resolution, and iterative feedback.

The AutoGen framework allows seamless orchestration of agent interactions, round-based conversations, and custom message routing. The core components of the architecture and their interaction protocol are detailed in the next sections.

4.3 System Configuration

In the figure 4.1 we can see the complete workflow of our multi-agentic debate system for performing the tasks.

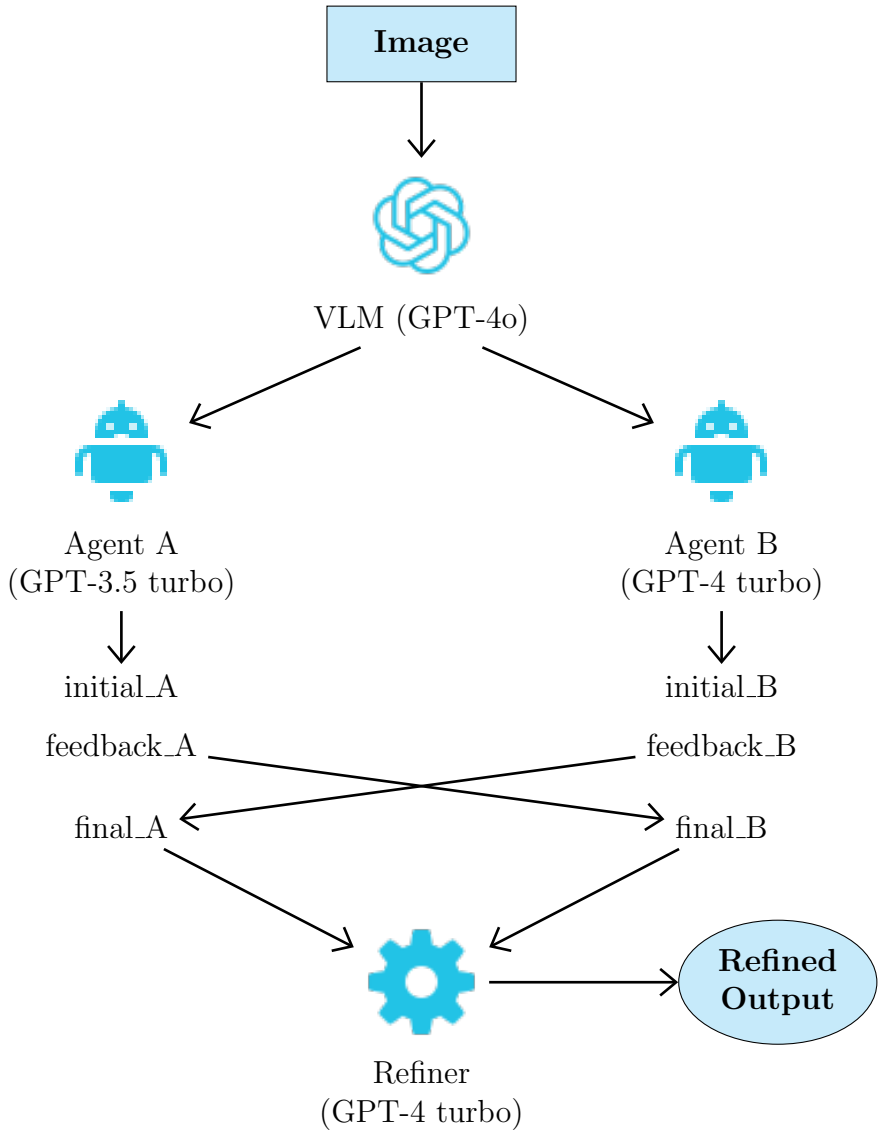


Figure 4.1: Workflow involving image input, vision-language model (VLM), two agents, and a refiner for output synthesis.

4.3.1 Input format and preprocessing

Our system operates on annotated samples from the YESBUT dataset, where each instance consists of a two-panel comic image accompanied by a structured JSON object containing multiple annotated fields. These include the literal description, contradiction explanation, moral/philosophical interpretation, and a multiple-choice list of four candidate titles, one of which is correct.

The image is passed to a VLM model (GPT-4o) using a multimodal prompt, which requests a literal, objective description of the two panels. The output is a paragraph of plain-text description detailing the visual components of each panel, such as characters, objects, expressions, and spatial arrangements.

Both Debater A and Debater B receive the same prompt, constructed by combining the literal image description (VLM output), the formatted title options from the JSON file, and explicit instructions to select the most appropriate title and justify their reasoning. This ensures that both agents are working from an identical information baseline, which is crucial for the fairness and interpretability of their debate.

4.3.2 Agents' roles and configuration

The proposed system is composed of three primary agents, each configured with a distinct large language model (LLM) backend and specialized responsibilities. These agents work collaboratively to analyze a given comic's description and title options through a structured debate process.

Debater A

Debater A is implemented using the GPT-3.5turbo model configured through OpenAI API access. This agent is responsible for initiating the reasoning process by:

- Analyzing the literal description of the comic,
- Identifying the central contradiction or thematic element,
- Selecting the most appropriate title option from a given multiple-choice list,
- Justifying its selection based on internal reasoning.

This agent is instructed to emphasize logical coherence and narrative alignment while maintaining brevity and clarity in its justification.

Debater B

Debater uses the GPT-4 turbo model, also from OpenAI. Like Debater A, it receives the same comic description from the VLM. However, its role is intentionally designed to:

- Offer an alternative perspective or interpretation,

- Challenge or supplement Debater A's reasoning,
- Defend or revise its own selected title with justification.

By using a different model backbone, the system encourages perspective diversity, exposing potential blind spots in a single-agent setup.

Refiner agent

The Refiner agent is implemented using GPT-4 turbo, a model specifically designed for enhanced logical reasoning. This agent plays the role of an impartial judge. It receives the final outputs from both debaters and performs the following:

- Compares the justifications provided by both agents,
- Evaluates which argument more accurately captures the comic's core contradiction or intent,
- Produces a final decision indicating the most appropriate title.

The Refiner operates without direct access to the ground-truth label and bases its decision solely on the strength of reasoning and alignment with the comic's theme.

4.3.3 Corrective Feedback Method:

We have applied the corrective feedback method to the debates that occur between the AI agents. There are five types of feedback correcting method.

- **Explicit Correction:** Requester provides the answer in correct form.
- **Recast:** Requester implicitly reformulates the errors of a crowd worker and provides correction.
- **Clarification Request:** Requester indicates that the annotation is not understood and ask for repitition or reformulation.
- **Metalinguistic Clues:** Requester poses questions or comments without providing the correct form.
- **Elicitation:** Requester asks question (not Yes/ No) and crowd workers give answers.

Specially for the all 4 tasks we have applied "Recast" corrective feedback method.

Why corrective feedback method works

Corrective feedback is effective because it:

- It targets specific mistakes.

- Offers scaffolded support (gradually reducing guidance).
- Encourages active learning.
- Builds long-term competence, not just short-term compliance.

4.4 Complete workflow to execute the tasks

Now we discuss the workflow of the debating system for performing the pre-defined tasks, which can be viewed in figure 4.1.

4.4.1 Generation Tasks

Goal: Generate the underlying contradiction and literal description for the image. In the YESBUT dataset we have 349 images. For the two generation tasks :

- First we used a VLM model(gpt-4o), with the prompt to generate the contradiction or literal description from the image.
- Now we made agent A and agent B along with the LLM's gpt-3.5 turbo and gpt-4o turbo respectively with some helping tools. After getting the output from VLM model we give it to the two agents.
- These two debate agents debated for 3 rounds. In the first round both of them proposed an output based on what they have got from VLM. After initial proposal they will get feedback from each other and will proposed a better version output than the initial output. Then Agent A will give its output to Agent B and vice versa. After getiing each others output both of them will propose a final solution.
- Now we made a refiner agent corresponding to gpt-4o turbo who gets all the final outputs from both of the agents and will give a refined version of contradiction or literal description corresponding to a juxtaposed comic image.

4.4.2 Title matching task

Goal: Select the most suitable title for a comic from multiple options.

- **Input:** Comic image + 4 candidate titles (1 correct, 3 distractors).
- **Debate Rounds:**
  - Each agent independently analyzes the comic and ranks titles.
  - Agents critique each other's rankings.
  - Refiner combines insights and finalizes the best title.

4.4.3 Underlying philosophy selection task

Goal: Select the philosophy/satire reflected in the comic from multiple options.

- **Input:** Comic image + 4 philosophy options (1 correct, 3 distractors).
- **Debate Rounds:**
  - Agents independently select and justify their choice.
  - Agents debate ambiguities
  - Refiner validates arguments and finalizes the best option.

4.5 Evaluation Strategy

To assess the effectiveness of the proposed multi-agent debate system, we evaluated model performance across all four YESBUT tasks: literal description generation, contradiction generation, title matching, and philosophy selection. Different evaluation metrics were used depending on the task type — generative or classification.

For contradiction generation and literal description, we employed the following standard text generation metrics:

- **BERT Score:** Measures the semantic similarity between generated and reference texts using contextual embeddings.
- **ROUGE-2 (Recall):** Captures bigram overlap between generated output and reference, with emphasis on recall to ensure key elements are not omitted.

These metrics were computed between the model-generated outputs and the human-authored annotations from the YESBUT dataset. Our results showed measurable improvement in both the BERT score and ROUGE-2 when using the multi-agent debate setup with corrective feedback, compared to single-agent baselines.

For title matching and philosophy selection, which are multiple-choice classification tasks, we used accuracy as the primary evaluation metric. A prediction was marked as correct if the selected option matched the annotated ground-truth answer.

Despite the simplicity of this metric, it effectively reflects the model's ability to resolve subtle visual or philosophical contradictions in a forced-choice setting. Our multi-agent system achieved a slight but consistent improvement in accuracy over single-agent systems, indicating the benefit of deliberative reasoning through agent interaction.

Chapter 5

Experimental Results

In this chapter, we take a look at the results obtained by applying our proposed methods to the YESBUT data.

We evaluate BERT and ROUGE-2 score for two generation task; literal description and contradiction generation, and accuracy measure in percentage for two selection tasks; underlying philosophy selection and title matching, through our multi-agent debate system, which is shown in table 5.1.

Task	**Evaluation on proposed sysytem**	**Evaluation on GPT-4**
Literal description generation	BERT: **0.9031**	BERT: 0.8832
	ROUGE-2: 0.8537	ROUGE-2: **0.8746**
Contradiction generation	BERT: **0.8972**	BERT: 0.8764
	ROUGE-2: **0.8413**	ROUGE-2: 0.8321
Philosophy selection	Accuracy: **0.8610**	Accuracy: 0.8276
Title matching	Accuracy: **0.7321**	Accuracy: 0.6025

Table 5.1: Comparison of evaluation metrics for each comic understanding task with and without GPT-4.

As shown in Table 5.1, our proposed multi-agent debate framework outperforms the single-agent GPT-4 (best-performing model in the previous literature 2.1) baseline across all four comic understanding tasks. In the literal description generation task, the BERT score improves from 0.8832 to 0.9031, indicating stronger semantic alignment between the generated descriptions and the ground-truth references. Although there is a slight decrease in ROUGE-2 recall for this task (from 0.8746 to 0.8537), the overall semantic quality — as captured by BERTScore — is notably higher, suggesting richer or more expressive generations that maintain core meaning even with less exact n-gram overlap.

For the contradiction generation task, both BERT score and ROUGE-2 show consistent im-

provements. The BERT score rises from 0.8764 to 0.8972, while ROUGE-2 increases from 0.8321 to 0.8413. These improvements highlight the ability of the multi-agent system to better capture nuanced contradictions between the two comic panels by incorporating diverse reasoning perspectives from multiple agents and refining them collaboratively.

The most significant gain is observed in the title matching task, where accuracy improves from 60.25% using GPT-4 alone to 73.21% with our multi-agent setup. This demonstrates the benefit of deliberation and critical feedback in resolving ambiguity among multiple plausible options. Similarly, the philosophy selection task shows a significant increase in accuracy, ranging from 82. 76% to 86. 10%. These classification tasks particularly benefit from agent disagreement, challenge-response dynamics, and final synthesis by a refiner agent, all of which contribute to more grounded and reasonable choices.

Overall, these results validate the effectiveness of the multi-agent debate approach in enhancing performance across both generative and discriminative tasks. The observed improvements are particularly impactful in tasks that involves subtle humor, contradiction, or abstract reasoning, where a single LLM may overlook important nuances or fail to self-correct. Our approach leverages disagreement between debaters as a productive mechanism, ultimately leading to more reliable and interpretable outcomes.

Chapter 6

Conclusion and Future Work

In this work, we introduced a multi-agent debate framework to enhance AI's ability to understand humor through juxtaposed comics, addressing key challenges in nonlinear reasoning, visual interpretation, and abstract comprehension. Our experiments demonstrated that:

- Collaborative agents (Debater_A, Debater_B, Refiner) significantly improved performance on literal description and contradiction generation, as measured by BERT Score and ROUGE-2 recall.
- Title matching and philosophy selection saw marginal but meaningful accuracy gains, highlighting the role of adversarial critique in refining ambiguous decisions.

By integrating agentic AI principles, we bridged critical gaps identified in prior work, particularly in visual humor understanding and narrative abstraction.

6.1 Future Work

To extend this research, we propose the following directions:

- Integrate long-term context (memory augmented agents) to retain debate histories and improve consistency.
- If the number of debaters and the number of debate rounds could have been increased, then chances of getting a good score will also increase.
- Increasing the number of images can be effective.
- Integrate sentiment analyzer. It will enhance agents' understanding of humor by analyzing emotional tone in comics.

Bibliography

[1] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023. URL: https://openreview.net/pdf?id=zj7YuTE4t8.

[2] Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. Cracking the code of juxtaposition: Can ai models understand the humorous contradictions. *arXiv preprint arXiv:2405.19088*, 2024. URL: https://arxiv.org/abs/2405.19088.

[3] Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023. URL: https://arxiv.org/pdf/2312.05230.

[4] Tuo Liang, Zhe Hu, Jing Li, Hao Zhang, Yiren Lu, Yunlai Zhou, Yiran Qiao, Disheng Liu, Jeirui Peng, Jing Ma, et al. When'yes'meets' but': Can large models comprehend contradictory humor through comparative reasoning? *arXiv preprint arXiv:2503.23137*, 2025.