Rethinking Information Synthesis in Multimodal Question Answering A Multi-Agent Perspective

Tejas Anvekar[*] **Krishna Singh Rajput**[*] **Chitta Baral** **Vivek Gupta**[†]
Arizona State University
`{tanvekar,krajput5,chitta,vgupt140}@asu.edu`
https://coral-lab-asu.github.io/MAMMQA/

Abstract

Recent advances in multimodal question answering have primarily focused on combining heterogeneous modalities or fine-tuning multimodal large language models. While these approaches have shown strong performance, they often rely on a single, generalized reasoning strategy, overlooking the unique characteristics of each modality ultimately limiting both accuracy and interpretability. To address these limitations, we propose MAMMQA , a multi-agent QA framework for multimodal inputs spanning text, tables, and images. Our system includes two Visual Language Model (VLM) agents and one text-based Large Language Model (LLM) agent. The first VLM decomposes the user query into sub-questions and sequentially retrieves partial answers from each modality. The second VLM synthesizes and refines these results through cross-modal reasoning. Finally, the LLM integrates the insights into a cohesive answer. This modular design enhances interpretability by making the reasoning process transparent and allows each agent to operate within its domain of expertise. Experiments on diverse multimodal QA benchmarks demonstrate that our cooperative, multi-agent framework consistently outperforms existing baselines in both accuracy and robustness.

1 Introduction

Multimodal question answering (MMQA) aims to answer complex queries by jointly reasoning over text, tables, and images, reflecting real-world information needs in domains such as scientific analysis, business intelligence, and education (Talmor et al., 2021; Hannan et al., 2020). Early MMQA systems typically linearized tables or generated image captions to cast all inputs into a text-only format, feeding them into pretrained text-only models (Luo et al., 2023a; Chen et al., 2020, 2021). While effective under certain settings, these unified approaches often obscure the unique structure and semantics of each modality, leading to degradation when inputs are missing or when fine-grained visual and tabular cues are critical.

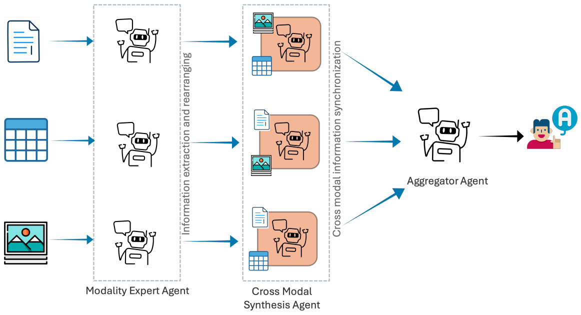
Information extraction and rearranging
Cross modal information synchronization
Modality Expert Agent
Cross Modal Synthesis Agent
Aggregator Agent
A

Figure 1: Depicting Illustration for our proposed MAMMQA , with three agents: 1) Modality Expert, that extracts modality specific insights; 2) Cross Modal Sythesis Agent, that synchronises information across modalities with insights from Modality Expert; 3) Aggregator Agent, that ground the answer using extracted cross modal information.

Recent advances in prompt-based reasoning have unlocked zero-shot and few-shot capabilities in large language models. Chain-of-Thought (CoT) prompting (Wei et al., 2022) and its multimodal extensions (Zhang et al., 2023; Zheng et al., 2023) guide a single LLM to generate intermediate steps, improving factual accuracy. Tree-of-Thoughts (ToT) further introduces search over multiple reasoning branches (Yao et al., 2023). However, these monolithic strategies still treat the LLM as a black box, entangling modality-specific extraction with cross-modal synthesis, which can obscure errors, hinder interpretability, and induce hallucinations when faced with ambiguous or partial inputs.

By contrast, multi-agent and ensemble techniques in NLP have demonstrated that specialized experts can collaborate to improve both accuracy and robustness (Chen et al., 2023; Puerto et al., 2023). Yet, such architectures remain underexplored in the multimodal setting. We identify a

[*] contributed equally
[†] primary supervisor

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3674–3686
December 20-24, 2025 ©2025 Association for Computational Linguistics

key opportunity: decoupling modality-specific evidence extraction from cross-modal integration and final answer adjudication can both leverage domain-specific strengths and provide transparent, verifiable reasoning traces.

In this work, we introduce MAMMQA , a fully prompt-driven, multi-agent framework for MMQA that dynamically allocates three types of agents modality experts, cross-modal synthesizers, and a consensus aggregator to decompose and solve complex queries without any fine-tuning. Our core contributions are:

- MAMMQA : A fully prompt-driven, multi-agent MMQA framework that splits reasoning into three interpretable stages modality experts, cross-modal synthesis, and evidence-grounded aggregation without any fine-tuning.
- **Unified, role-consistent agents**: A single prompt template reused across text, table, and image experts, enabling dynamic activation, efficient inference, and transparent error tracing.
- **State-of-the-art zero-shot performance**: Outperforms CoT, CapCoT, and ToT baselines and matches or exceeds several fine-tuned models on MULTIMODALQA and MANY-MODALQA, across both proprietary and open-source LLMs.
- **Robustness and calibration**: Static agents beat dynamic search methods (e.g. ToT) by over 10 %, maintain faithfulness under noise and irrelevant context, and avoid hallucinations via evidence-based abstention.

By structuring MMQA as a pipeline of specialized agents, MAMMQA not only achieves high accuracy but also provides end-to-end transparency and graceful failure modes in the face of ambiguous or incomplete inputs key properties for real-world deployment.

2 Our MAMMQA Framework

We propose MAMMQA , a Multi-Agent Multi-modal Question Answering framework designed to address core challenges in MMQA, such as modality ambiguity, fragmented evidence, and hallucinated reasoning. Rather than relying on monolithic prompting or fine-tuned end-to-end models, MAMMQA adopts a structured, agent-based architecture that decomposes the reasoning process into interpretable, well-defined steps. Each agent specializes in a narrow subtask, enabling systematic insight extraction, targeted cross-modal synthesis, and final answer adjudication through consensus.

2.1 Motivation and Design Principles

Multimodal question answering requires models to accurately interpret and integrate information across diverse modalities. However, conventional prompting approaches often struggle with modality disambiguation and fail to coordinate evidence coherently. To address this, our framework is inspired by the structure of expert committees, where domain specialists independently contribute insights that are later synthesized into a final decision. MAMMQA is guided by three core principles. Each agent is assigned a focused, well-defined role either modality-specific analysis, cross-modal synthesis, or answer aggregation mirroring real-world task delegation. The reasoning process is decomposed into a multi-stage pipeline that progresses from factual extraction to cross-modal interpretation and ultimately to consensus. The entire framework is prompt-driven, leveraging pre-trained language models without requiring any task-specific fine-tuning.

2.2 Overview of the MAMMQA Architecture

The MAMMQA framework consists of three sequential stages: modality-specific insight extraction, cross-modality synthesis and reasoning, and final answer aggregation. Each stage employs a set of pre-trained language models acting as agents, each prompted with a structured task definition. The number of active agents in the system dynamically adjusts depending on the available input modalities, ranging from five in the bi-modal case to seven for full tri-modal inputs. All agents interact through textual interfaces, ensuring that the pipeline remains fully modular and interpretable, best illustrated by Figure 1.

2.3 Stage I: Modality Expert Agent

The first stage applies a unified modality expert agent to each available input modality text, table, and image. Although executed independently for each modality, the same underlying system prompt is used across all instances; the only variation lies in the input modality. This agent is not tasked with answering the question but instead focuses on extracting key information relevant to the query, such as factual details, temporal markers, and contextual cues. It also flags gaps or ambiguities in the input that may impede complete reasoning. The agent's output is structured using a consistent templating format, producing modality-specific insights that form the foundation for the next reasoning stage.

2.4 Stage II: Cross-Modality Synthesis Agent

In the second stage, the same cross-modality synthesis agent is invoked once for each modality acting as the anchor context. Each instance of the agent takes as input the insights from a single modality-specific agent in Stage I, the raw data from the remaining two modalities, and the original question. This setup allows the agent to synthesize information across modalities while maintaining a consistent perspective grounded in one primary modality. The agent generates a complete answer to the question, supports it with structured reasoning, and identifies any remaining uncertainties. As with the first stage, the output adheres to a templated format comprising extracted insights, intermediate sub-answers, and a final answer. These outputs are then passed to the final aggregation stage.

2.5 Stage III: Aggregator Agent

The final stage introduces an aggregation agent responsible for synthesizing the outputs of the three synthesis agents. Crucially, this agent does not have access to the raw inputs. Instead, it receives only the generated responses and reasoning from the prior stage, along with the original question. Its task is to resolve disagreements, consolidate consistent answers, and produce a final, justified response. The agent follows a hierarchical decision process: it first checks for answer consistency among the agents; if two or more agree and present clear reasoning, it adopts that answer. If two agents express uncertainty while the third provides a confident, well-supported answer, the agent selects the confident one. In cases where all three answers differ, it evaluates them based on the clarity, coherence, and strength of their rationale, ultimately choosing the most convincing explanation. The final output includes both the chosen answer and a transparent summary of the reasoning process that led to its selection.

2.6 Advantages of the Multi-Agent Prompting Framework

The agent-based design of MAMMQA delivers interpretable and robust multimodal question answering. Its transparency stems from traceable intermediate steps, and robustness is achieved through redundant synthesis agents for answer cross-verification. Crucially, MAMMQA scales across LLM sizes and domains using only prompting, avoiding fine-tuning. This modularity supports generalization and faithful data grounding, establishing a framework for enhanced reliability, accuracy, and interpretability within MULTIMODALQA.

3 Experiments

We evaluate the effectiveness of our method on the Multimodal Question Answering (MMQA) task using exact match, demonstrating superior performance compared to prior state-of-the-art approaches, including UniMMQA (Luo et al., 2023a), AutoRouting (Talmor et al., 2021), ImplicitDecomp (Talmor et al., 2021), Binder (Cheng et al., 2023), SKURG (Yang et al., 2023a), PERQA (Yang et al., 2023b), Solar (Yu et al., 2023), UniRaG (Sharifymoghaddam et al., 2025), AETGA (Zhang et al., 2024), and PReasM-Large (Yoran et al., 2021). Additionally, we benchmark against standard in-context prompting baselines such as Chain-of-Thoughts (Wei et al., 2022) (CoT), Image-Captioning + CoT (CapCoT), and Tree-of-Thoughts (Yao et al., 2023) (ToT), across both proprietary (`gpt-4o-mini` (OpenAI et al., 2024), `Gemini-1.5-flash-8B` (Team and et al, 2024)) and open-source models (`Qwen2.5-VL-Instruct-7B/3B` (Qwen et al., 2025)). We further conduct endurance tests to assess the robustness of our approach under challenging scenarios. **Note:** All experiments involving `Qwen` models were conducted locally on a system equipped with 8× NVIDIA H200 GPUs.

Datasets We evaluate our approach on two prominent benchmark datasets designed to test key reasoning capabilities in multimodal question answering (MMQA).

MANYMODALQA (Hannan et al., 2020) contains 10,190 questions involving text, images, and tables distributed across 2,873 images, 3,789 passages, and 3,528 tables. The dataset is intentionally constructed with ambiguous questions where the relevant modality is not explicitly indicated. This

Modality	Img	Tb | Img	Tb | Txt	Tb	Txt | Img	Txt	Total
OpenAI 4o Mini							
CoT	33.15	53.81	66.67	84.55	55.95	77.67	64.60
CapCoT	53.91	64.98	69.05	84.14	**61.90**	77.33	70.39
ToT	54.97	63.35	64.37	67.70	61.11	69.65	64.88
Ours	**61.31**	**70.30**	**81.58**	**89.16**	59.75	**85.57**	**76.37**
Gemini 1.5-Flash 8B							
CoT	47.41	53.38	**58.88**	74.73	**46.43**	72.82	62.16
CapCoT	47.84	50.02	55.87	74.88	39.29	72.42	60.66
ToT	36.93	43.06	52.32	53.72	33.33	70.61	53.10
Ours	**51.23**	**54.12**	57.42	**83.69**	42.86	**79.47**	**65.84**
Qwen *2.5 VL 7B Instruct*							
CoT	29.11	32.58	30.66	38.75	17.86	38.28	33.84
CapCoT	48.10	53.94	60.56	71.52	41.67	71.31	61.54
ToT	**55.90**	47.82	52.50	60.83	41.64	64.44	57.12
Ours	50.74	**55.88**	**63.68**	**81.35**	**53.26**	**80.51**	**67.56**
Qwen *2.5 VL 3B Instruct*							
CoT	11.86	23.71	22.14	32.25	14.29	25.52	23.15
CapCoT	**48.10**	42.08	47.08	**64.94**	**39.29**	65.04	**53.98**
ToT	42.01	**43.65**	**48.40**	52.57	33.74	66.51	52.91
Ours	33.73	43.10	45.33	62.29	35.52	**67.73**	52.12

Table 1: Quantitative Analysis on MULTIMODALQA dataset

design tests a model's ability to perform modality disambiguation and selectively retrieve relevant information. It highlights whether the model can reason about what modality is needed and how to integrate it effectively. With 2,036 training and 3,055 development examples, it serves as a strong benchmark for evaluating modality selection under uncertainty.

MULTIMODALQA (Talmor et al., 2021) consists of 29,918 question-answer pairs across multiple modalities, with a significant 35.7% of questions requiring cross-modal reasoning that is, combining evidence from different types of inputs. This dataset assesses a model's ability to integrate heterogeneous information and reason jointly across modalities, not just within a single source. It is divided into training (23,817), development (2,442), and test (3,660) splits, and is particularly suited for evaluating compositional reasoning and information fusion in complex multimodal contexts.

Baselines To comprehensively evaluate our method, we compare it against several strong baselines spanning both finetuned and prompting-based approaches for multimodal question answering.

Finetuned Baseline. **UniMMQA** (Luo et al., 2023b) serves as a T5 (Raffel et al., 2020)-based finetuned baseline and represents a strong state-of-the-art model for MMQA. It is trained with supervised signals across modalities, enabling robust cross-modal representation learning.

Prompting-Based Agents. Our proposed prompting strategies adopt an agent in-context learning setup. Each agent specializes in a modality and collaborates to perform modality disambiguation and reasoning, allowing the model to retrieve, decompose, and integrate evidence dynamically without any parameter updates.

Reasoning Variants. **Chain-of-Thought (CoT)** guides the model to generate intermediate reasoning steps from the raw question and context.

CapCoT enhances CoT by incorporating detailed image and table captions generated via Gemini-2.0-Flash, providing richer modality cues in textual form.

Tree-of-Thought (ToT) extends CapCoT by simulating multiple reasoning paths using a depth-first search (DFS) strategy over caption-augmented inputs, introducing structured exploration for better answer synthesis.

LLM Configurations. All models, both proprietary and open-source, are evaluated with a `temperature` of 0.3 and `top-p` of 0.7 to ensure

Methods	Text	Table	Image	Total
Human	*92.00*	*89.60*	*94.00*	*91.60*
Voting	23.70	22.90	15.50	21.10
MMQA	48.60	40.40	27.20	39.70
MMQA†	59.30	46.30	29.00	46.30
UniMMQA Finetuned T5 Model				
Base	46.60	60.70	30.20	45.40
Large	48.50	67.50	34.90	50.00
3B	49.80	58.30	40.90	52.10
OpenAI 4o-mini				
CoT	87.20	94.23	57.33	81.21
CoT*	68.22	70.51	59.42	66.54
CapCoT	87.68	94.05	68.26	84.41
ToT	84.94	93.19	72.90	84.70
Ours	**92.50**	**96.78**	**78.02**	**89.90**
Gemini 1.5-Flash 8B				
CoT	86.05	91.52	68.77	82.81
CoT*	54.93	61.15	34.77	51.41
CapCoT	85.74	91.40	63.14	81.34
ToT	86.08	86.81	62.81	79.80
Ours	**89.76**	**94.52**	**77.33**	**87.91**
Qwen *2.5 VL 7B Instruct*				
CoT	59.84	68.71	45.47	58.87
CoT*	61.80	66.73	54.53	61.46
CapCoT	83.50	92.86	71.07	83.41
ToT	81.95	90.41	69.29	81.89
Ours	**87.11**	**96.31**	**77.56**	**87.61**
Qwen *2.5 VL 3B Instruct*				
CoT	70.08	75.61	50.70	66.54
CoT*	58.77	64.55	59.51	58.77
CapCoT	80.79	91.38	67.13	80.63
ToT	82.66	86.14	68.11	80.42
Ours	**88.79**	**94.90**	**72.67**	**86.37**

Table 2: Quantitative results on the MANYMODALQA dataset. Superscript † denotes the oracle setting, while * indicates the no-context (open-book QA) variant. Red highlights mark cases where CoT fails to abstain from answering without context, and occasionally outperforms its baseline in the [No Image] setting, suggesting potential data leakage.

consistent generation behavior. We employ the framework[1] to implement Tree-of-Thought (ToT) agents. Our proposed agents operate in a static and synchronous manner, removing the need for asynchronous execution. The corresponding prompts are detailed in Appendix C: Prompt A presents the Modality Expert Agent prompt, Prompt B illustrates the Cross-Modality Agent prompt, and Prompt C provides the Aggregator prompt.

3.1 Comparison with State-of-the-Art

A. MULTIMODALQA Results. As shown in Table 3, our method consistently outperforms prompting-based baselines across both proprietary and open-source models. On Qwen2.5-VL-7B, our agentic method achieves **76.37%**, surpassing CapCoT (**+5.98%**) and ToT (**+11.49%**). Largest gains are observed in cross-modal settings like *[table, text]* (+17.21%) and *image* (+7.4%), underscoring the benefit of grounded signal integration. On Gemini-1.5 Flash 8B, we achieve **65.84%**, outperforming CoT by **+3.68%**. While prompting methods show less spread here, our model remains notably better in *text* (+6.65%) and *table* (+8.81%), showing advantages in structurally complex inputs.

Model	Single	Multi	Overall
Finetuned Models			
AutoRouting	51.7	34.2	44.7
ImplicitDecomp	51.6	44.6	48.8
Binder	-	-	51.0
SKURG	66.1	52.5	59.8
PERQA	69.7	54.7	62.8
Solar	69.7	55.5	59.8
UniRaG	**71.7**	62.3	67.4
AETGA	69.8	**64.7**	68.8
PReasM L	-	-	59.0
MMQA-T5 L	-	-	57.9
UniMMQA (T5 B)	-	-	67.9
UniMMQA (T5 L)	-	-	71.3
UniMMQA (T5 3B)	-	-	**75.5**
Zero-Shot Models			
CoT Qwen 3B	23.75	22.24	23.15
CoT Qwen 7B	36.07	30.91	33.84
Our Agent 3B	57.72	43.39	52.12
Our Agent 7B	**73.16**	**58.93**	**67.56**

Table 3: Comparison of models on MULTIMODALQA dataset across single-modality, multi-modality, and overall performance. Note: B depicts Base Model and L depicts Large model.

Open-Source Scaling (Qwen). Our method yields strong results even on small open-source models. On Qwen2.5-VL-3B, it achieves **52.12%**, beating CoT by **+28.97%** and closely trailing ToT (**-0.79%**) despite significantly lower computation. On Qwen2.5-VL-7B, our method reaches **67.56%**, outperforming CapCoT (**+6.02%**), ToT (**+10.44%**), and CoT (**+33.72%**). These gains are especially prominent in multi-hop, hybrid modality tasks e.g., *[text, image]* (+90.53%) and *[table, text]* (+40.48%).

Model Scaling. Increasing model size from 3B to 7B brings a substantial **+29.62%** gain in overall performance. Improvements are concentrated in *image* (+50.43%) and *[text, image]* (+49.95%) modalities, suggesting that increased capacity amplifies our agentic system's ability to perform complex, cross-modal reasoning.

Efficiency Over Larger Baselines. Despite its smaller size, our 3B model surpasses Qwen-7B CoT on multiple modalities *text* (+76.94%), *[table, image]* (+32.31%) with an overall gain of **+54.02%**.

[1] https://github.com/openai/openai-agents-python

This highlights the architectural efficiency of structured agentic reasoning over naive pattern-based prompting.

B. MANYMODALQA Results. Our method generalizes well on this more challenging benchmark. On Qwen2.5-VL-7B, we score **89.90%**, outperforming CoT by **+8.69%**, ToT by **+5.20%**, and CapCoT by **+5.49%**. The largest gains occur in visual reasoning tasks e.g., *image* (+5.12%). On Gemini 1.5-Flash 8B, we obtain **87.91%**, with consistent improvements across modalities: **+5.10%** over CoT, **+8.11%** over ToT, and **+6.57%** over CapCoT. Our method shows stronger synergy between vision and language compared to captioning-heavy baselines.

Open-Source Performance On Qwen2.5-VL-7B we achieve **87.61%**, outperforming ToT (**+5.72%**) and CapCoT (**+4.20%**). Notably, we also surpass Gemini 1.5-8B in overall score (+0.30%) and structured modalities like *text* (+4.32%) and *table* (+5.38%). On Qwen2.5-VL-3B, our model reaches **86.37%**, significantly ahead of CoT (**+19.83%**) and CapCoT (**+5.74%**), and even outperforming Gemini-8B on both total score (+1.56%) and image (+8.56%).

Model Scaling. Scaling from Qwen 3B to 7B gives a modest **+1.24%** overall, but a pronounced boost in visual reasoning (**+6.72%**). This suggests that while our architecture is already strong at 3B, larger models especially enhance performance on ambiguous, visually grounded questions.

Finetuned vs. Zero-shot. Despite being zero-shot, our 7B method outperforms several finetuned baselines e.g., SKURG (59.8%), Solar (59.8%) and rivals AETGA (68.8%) and UniRaG (67.4%). Compared to Qwen-7B CoT, we observe large gains in both single-modality (**+37.09%**) and multi-modality (**+28.02%**) settings. Even our 3B variant exceeds Qwen-7B CoT by **+18.2%**, reaffirming that architecture not just size drives robust performance.

3.2 Robustness Analysis

MAMMQA **Mislabeling Robustness.** In our experiments with the MULTIMODALQA dataset, we evaluated multiple LLMs such as GPT and Qwen using baseline methods CoT, CapCoT, Tree-of-Thoughts, and our agentic method. During the analysis, we discovered discrepancies in the ground truth labels of the dataset, such as typos and outdated factual information. For instance, an answer labeled movie name "*laughin*" should have been "*laughing*." This led to certain models, particularly CoT, memorizing and reproducing these incorrect labels, thereby inflating their performance metrics artificially.

Model (Qwen 7B)	Old	New
TreeOfThoughts	57.12	59.06 (+1.94)
CoT	33.84	35.05 (+1.21)
CapCoT	61.54	64.52 (+2.98)
OurAgent	**67.56**	**71.58 (+4.02)**
Model (Qwen 3B)	**Old**	**New**
TreeOfThoughts	52.91	54.36 (+1.45)
CoT	23.15	24.07 (+0.92)
CapCoT	53.98	56.16 (+2.18)
OurAgent	**52.12**	**55.24 (+3.12)**

Table 4: Performance improvements with lable correction across model sizes on MULTIMODALQA .

After correcting these labels, we observed that the performance of our agentic method improved significantly more than the baseline methods as depicted in Table 4. This highlights the *robustness of our approach in extracting and synthesizing information from multiple modalities and grounding it accurately, even when faced with noisy or inconsistent data*. This correction process ultimately underscores the efficacy of our model in real-world scenarios.

MAMMQA **Pertubations Robustness.** Table 5 evaluates model robustness under two text-level perturbations: (1) sentence or paragraph shuffling and (2) injection of irrelevant context. In the *Text Shuffle* setting, baseline methods like TreeOfThoughts (7B: -42.21%, 3B: -6.97%) and CapCoT (7B: -39.11%, 3B: -8.82%) exhibit substantial drops but still attempt to answer suggesting reliance on memorized question patterns. In contrast, MAMMQA exhibits steep performance drops (7B: **-91.24%**, 3B: **-85.30%**), suggesting that it fails gracefully under broken contextual grounding, thereby reducing the risk of hallucinated answers.

Model (7B)	Original	Text Shuffle	Irrelevant Context
TreeOfThoughts	57.12	33.01 (-42.21%)	52.45 (-08.18%)
CoT	33.84	31.18 (-07.86%)	29.54 (-12.71%)
CapCoT	61.54	37.47 (-39.11%)	55.39 (-09.99%)
OurAgent	**67.56**	**05.92** (-91.24%)	**63.74** (-05.65%)
Model (3B)	**Original**	**Text Shuffle**	**Irrelevant Context**
TreeOfThoughts	52.91	49.22 (-06.97%)	47.11 (-10.96%)
CoT	23.15	20.48 (-11.53%)	19.62 (-15.25%)
CapCoT	53.98	49.22 (-08.82%)	47.12 (-12.71%)
OurAgent	**52.12**	**07.66** (-85.30%)	**48.05** (-07.81%)

Table 5: Robustness of different reasoning strategies under perturbations across model sizes.

In the *Irrelevant Context* setting, where unrelated text is appended, OurAgent remains the most stable (7B: **-5.65%**, 3B: **-7.81%**) compared to TreeOfThoughts (7B: -8.18%, 3B: -10.96%) and CapCoT (7B: -9.99%, 3B: -12.71%). This demonstrates that while OurAgent avoids overcommitting in incoherent contexts, it retains robustness when faced with extraneous information underscoring its grounding-driven reasoning approach.

MAMMQA **Calibration Robustness.** Chain-of-Thought (CoT) prompting, while effective in unimodal text reasoning, fails to generalize reliably in multimodal contexts. On MULTIMODALQA, CoT consistently produces high-confidence yet unfaithful answers when modality-specific evidence is absent intentionally. This prompts an important question: *Can LLMs, when operating under the* MAMMQA *framework, refrain from answering when provided with incomplete inputs?* As depicted in Table 2, on `Qwen-7B`, CoT achieves 58.87%, but this rises to 61.46% in a no-context setting (CoT*) indicating a reliance on pretraining priors rather than grounded inference.

This behavior suggests that CoT "recalls" plausible reasoning paths learned during training rather than "inferring" from the input analogous to a language model predicting the next sentence in a familiar story, even when the plot doesn't match. In contrast, our agent-based architecture enforces structured, evidence-grounded reasoning. The Modality Expert Agent first extracts information independently from text, table, and image inputs. A Cross-Modality Expert then integrates these signals with consistency checks. Crucially, if no relevant evidence is found, these agents abstain from answering propagating that abstention to the Aggregator, which itself is blind to the original question. This ensures that the final output is generated only when sufficient grounded evidence exists.

This setup explicitly separates extraction from generation, reducing hallucinations and enforcing cross-modal faithfulness. As a result, our method achieves 89.90% on OpenAI `Qwen2.5-VL-7B` and 87.61% on `Qwen-7B` outperforming CoT by **+8.69%** and **+28.74%**, respectively. *Unlike CoT, which confidently answers even in the absence of valid context, our agents are "evidence-seeking", "input grounded", rather than "answer-seeking," leading to more trustworthy and robust multimodal QA.*

3.3 Choices in MAMMQA Architecture

Dynamic vs. Static Agents . Dynamic agentic frameworks like Tree-of-Thoughts (ToT) (Yao et al., 2023) rely on explicit search typically via depth-first traversal to enumerate and rank multiple reasoning paths. In our setup, ToT is instantiated with 3 agents and a max depth of 3, generating an average of 12 thoughts per question. Despite this computational overhead, as shown in Table 1, ToT achieves 57.12% on `Qwen-7B`, while our static agentic method using only 3 sequential agents achieves **67.56%**, a **+10.44%** gain.

Beyond accuracy, ToT exhibits failure modes indicative of brittle search behavior: it frequently returns confidently incorrect answers (avg. confidence *0.93*) and often declares multi-hop questions "unanswerable," missing key compositional signals. In contrast, our static framework without iterative search or re-ranking demonstrates more grounded reasoning, better calibration, and robust handling of multi-modal, multi-hop queries. These findings challenge the assumption that dynamic search improves generalization, and highlight the efficacy of a lean, static agentic architecture in complex QA tasks.

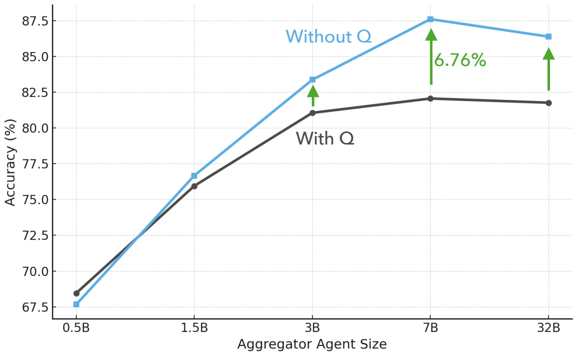

Figure 2: Aggregator Agent performance with and without question on MULTIMODALQA Dataset.

Question Agnostic Aggregator. As shown in Figure 2, our experiments reveal withholding the original question from the Aggregator Agent consistently improves performance across model scales. For instance, at 7B parameters, accuracy increases from 82.06% (with question) to **87.61%** (without question). This improvement arises because removing the question reduces reliance on linguistic priors and compels the Aggregator to synthesize answers solely from cross modality-grounded evidence provided by expert agents. *This, making the* MAMMQA *agent question agnostic* *make it's unbias, grounded, and more factual.*

Analogous to ensemble methods in NLP, where a meta-learner integrates outputs from base models without direct access to the input query, this separation mitigates bias and enhances answer faithfulness and factual consistency reducing hallucinations (Puerto et al., 2023).

4 Comparison with Related Work

We position our work within four key areas of multimodal question answering (MMQA): benchmark design, unified models, prompt-based reasoning, and multi-agent systems.

Benchmarks for Multimodal QA Datasets like MANYMODALQA (Hannan et al., 2020) and MULTIMODALQA (Talmor et al., 2021) challenge systems to reason across text, tables, and images, with a significant portion requiring cross-modal fusion. Others, such as HYBRIDQA (Chen et al., 2020), OTT-QA (Chen et al., 2021), and TAT-QA (Zhu et al., 2021), focus on structured-unstructured combinations. These benchmarks highlight core challenges in modality disambiguation and evidence integration issues MAMMQA explicitly tackles through specialized agents.

Unified Models Encoder-decoder architectures like AUTOROUTING and IMPLICITDECOMP (Talmor et al., 2021) embed all modalities into a shared space. More advanced models BINDER, SKURG, PERQA (Yang et al., 2023b), SOLAR (Yu et al., 2023), and UNIRAG (Sharifymoghaddam et al., 2025) incorporate retrieval and structural cues. While effective, these approaches often obscure modality roles and degrade with missing inputs. MAMMQA sidesteps these issues by activating only relevant agents per input.

Prompt-Based Reasoning Prompting strategies like Chain-of-Thought (Wei et al., 2022) and its multimodal variants (Zhang et al., 2023; Zheng et al., 2023), including Tree-of-Thoughts (Yao et al., 2023), offer zero-shot reasoning capabilities. However, they typically depend on a single LLM, making them prone to hallucinations and conflicts while MAMMQA distributes reasoning across modality-specific agents.

Multi-Agent Systems Agent-based approaches such as RECONCILE (Chen et al., 2023) leverage collective decision-making, often via voting. Though applied to math and planning (Yao et al., 2023), they remain underexplored in MMQA. MAMMQA adapts this paradigm by coordinating agents across modalities with structured synthesis, enabling verifiable and interpretable reasoning.

Unlike prior approaches that rely on monolithic models or single-agent prompting, MAMMQA introduces a multi-agent, interpretable architecture. It assigns specialized roles to agents based on modality and separates reasoning into distinct stages: extraction, synthesis, and aggregation. This design enables MAMMQA to match or exceed the performance of state-of-the-art MMQA models while offering improved interpretability, robustness, and zero-shot generalization.

5 Conclusion

We present MAMMQA , a modular, prompt-driven multi-agent framework for multimodal QA that performs structured reasoning through modality-specific extraction, cross-modal synthesis, and evidence-grounded aggregation entirely without finetuning. MAMMQA achieves state-of-the-art zero-shot results on both MULTIMODALQA and MANYMODALQA, outperforming prompting-based baselines and several finetuned models. On MULTIMODALQA, it achieves 76.37% with `Qwen2.5-VL-7B` and 67.56% with `Qwen2.5-VL-7B`, surpassing CapCoT and Tree-of-Thoughts by over 6% and 10%, respectively. On MANYMODALQA, it reaches 89.90% with `Qwen2.5-VL-7B` and 87.61% with `Qwen2.5-VL-7B`, outperforming CoT by up to 28.74%. In addition to strong performance, MAMMQA exhibits higher robustness and interpretability. It remains stable under irrelevant context and avoids hallucination in perturbed settings, while static agents outperform dynamic search-based methods like ToT with less complexity. These results underscore MAMMQA as a scalable, interpretable, and high-performing zero-shot solution for multimodal QA.

6 Limitations

MAMMQA 's reliance on separate LLM/VLM experts for each modality simplifies zero-shot generalization but incurs substantial inference latency, memory usage, and monetary cost. Extending the framework to additional modalities (e.g., audio, video, sensor data) would require equally capable foundation models or complex preprocessing pipelines, limiting applicability in resourceconstrained or real-time environments. The three-stage design (Aggregator) enhances transparency but makes the system brittle: mistakes in early extraction cannot be corrected downstream, and the Aggregator blind to raw inputs cannot recover missing or misinterpreted evidence. This fragility is reflected in our perturbation tests, where scrambled or incomplete context causes small accuracy drops. Incorporating iterative feedback or retrieval loops could improve robustness but would complicate the current prompt-driven simplicity.

7 Ethics Statement

The authors affirm that this work adheres to the highest ethical standards in research and publication. Ethical considerations have been meticulously addressed to ensure responsible conduct and the fair application of computational linguistics methodologies. Our findings are aligned with experimental data, and while some degree of stochasticity is inherent in black-box Large Language Models (LLMs), we mitigate this variability by maintaining fixed parameters such as temperature and top p for consistent generation across the models used. Furthermore, our use of LLMs, including `GPT-4o-mini`, `Gemini-1.5-flash-8B`, and `Qwen` models, complies with their respective usage policies. The research involved the analysis and correction of dataset mislabeling to ensure data integrity. We have used AI assistants (Grammarly and ChatGPT) to address the grammatical errors and rephrase the sentences. Finally, to the best of our knowledge, we believe that this work introduces no additional risk. To the best of our knowledge, this study introduces no additional ethical risks.

Acknowledgements

We thank the Complex Data Analysis and Reasoning Lab at Arizona State University for computational support, and Vihang Pancholi for his help during the submission process.

References

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *arXiv preprint arXiv:2309.13007*.

Wenhu Chen, Ming wei Chang, Eva Schlinger, William Wang, and William Cohen. 2021. Open question answering over tables and text. *Proceedings of ICLR 2021*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding language models in symbolic languages. *Preprint*, arXiv:2210.02875.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and qa over diverse inputs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7879–7886.

Haohao Luo, Ying Shen, and Yang Deng. 2023a. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8203–8213.

Haohao Luo, Ying Shen, and Yang Deng. 2023b. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9355–9367, Singapore. Association for Computational Linguistics.

OpenAI, :, and Aaron Hurst et. al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Haritz Puerto, Gözde Şahin, and Iryna Gurevych. 2023. MetaQA: Combining expert agents for multi-skill question answering. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3566–3580, Dubrovnik, Croatia. Association for Computational Linguistics.

Qwen, :, 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, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. 2025. UniRAG: Universal retrieval augmentation for large vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2026–2039, Albuquerque, New Mexico. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Gemini Team and Petko Georgiev et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang. 2023a. Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5223–5234, New York, NY, USA. Association for Computing Machinery.

Shuwen Yang, Anran Wu, Xingjiao Wu, Luwei Xiao, Tianlong Ma, Cheng Jin, and Liang He. 2023b. Progressive evidence refinement for open-domain multimodal retrieval question answering. *Preprint*, arXiv:2310.09696.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ori Yoran, Alon Talmor, and Jonathan Berant. 2021. Turning tables: Generating examples from semi-structured tables for endowing language models with reasoning skills. *CoRR*, abs/2107.07261.

Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Unified language representation for question answering over text, tables, and images. *Preprint*, arXiv:2306.16762.

Qing Zhang, Haocheng Lv, Jie Liu, Zhiyun Chen, Jianyong Duan, Hao Wang, Li He, and Mingying Xu. 2024. An entailment tree generation approach for multimodal multi-hop question answering with mixture-of-experts and iterative feedback mechanism. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4814–4822, New York, NY, USA. Association for Computing Machinery.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. In *arXiv preprint arXiv:2302.00923*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *arXiv preprint arXiv:2310.16436*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *Preprint*, arXiv:2105.07624.

A Additional Experimental Results

This section reports supplementary experiments extending our main evaluation. These include ablations analyzing model modularity, efficiency, and robustness.

Ablation on Synthesizer and Single Expert. Following reviewer feedback, we tested (a) removal of the *Cross-Modal Synthesizer* and (b) replacement of modality-specialized experts with a *single unified expert*. The results are presented in Table 6.

Model	MultiModalQA	Single Expert	ManyModalQA	Single Expert
`Qwen2.5-VL-7B`	76.37	72.17	89.90	85.70
Gemini 8B	65.84	61.64	87.91	83.71
`Qwen2.5-VL-3B`	67.56	53.36	87.61	63.41

Table 6: Ablation analysis on the Synthesizer and single-expert variants. Removing the Synthesizer or unifying experts both reduce accuracy by 4-25 points, demonstrating the necessity of modular specialization.

Removing the cross-modal Synthesizer causes significant drops (up to 24.2 pp on multi-hop tasks with `Qwen-7B`). Using a single unified expert consistently under performs relative to modality-specialized agents, confirming that modular decomposition is crucial not over-engineering.

Latency and Cost Efficiency. Although MAMMQA involves multiple agents, each executes a single prompt, unlike CoT, which relies on multi-sample decoding. Consequently, MAMMQA remains both interpretable and cost-efficient.

- `Qwen` 2.5-VL 3B + MAMMQA: **86.37%**
- `Qwen` 2.5-VL 7B + MAMMQA: **87.61%**

Both outperform stronger baselines such as `Qwen-7B` + CoT (66.54%) and Gemini 8B + CoT (84.81%), as well as GPT-4o + CoT (86.20%), demonstrating higher *performance per parameter*. The agentic overhead is minimal compared to CoT sampling-based reasoning.

Semantic Sensitivity: Text Shuffle Test. We additionally perform a "text shuffle" stress test, where words (not sentences) are randomly permuted to disrupt semantic coherence. CoT baselines show resilience (indicating reliance on superficial token associations), while MAMMQA suffers a larger drop (≈91.2%), highlighting stronger semantic sensitivity-an important trait for grounded multimodal reasoning.

These results empirically validate the design choices in MAMMQA: (i) modular specialization across experts is essential, (ii) the Synthesizer plays a key role in cross-modal composition, and (iii) the framework achieves both interpretability and cost efficiency without sacrificing accuracy.

B Qualitative Walkthrough Examples

We further illustrate MAMMQA's reasoning process through representative examples.

Example 1: "Which animal in the image is a mammal?"
Input: An image showing a frog, a dolphin, and a bird, with the caption "These animals live in different habitats."
Answer: *Dolphin.*

Stage 1 – Expert Agents: Image agent identifies animal regions; text agent extracts relevant concepts ("different habitats").

Stage 2 – Synthesizer: Fuses visual and textual evidence to reason over biological traits.

Stage 3 – Aggregator: Produces the reasoning trace: `[Visual: Dolphin → live birth] + [Text: warm-blooded] → Dolphin is mammal.`

Baseline (CoT) Comparison. CoT misclassifies "bird" due to token frequency bias. MAMMQA's modular design ensures factual grounding and interpretability.

Example 2: "According to the chart and text, which city had the highest rainfall?"
Input: A bar chart with rainfall data and a paragraph on weather patterns.
Answer: *Singapore.*

MAMMQA correctly identifies "Rainfall (mm)" as the relevant dimension, grounds the chart bars, aligns textual cues ("tropical"), and aggregates them for final reasoning.

These examples demonstrate how MAMMQA provides transparent, step-wise multimodal reasoning, maintaining interpretability while achieving competitive quantitative performance.

C Prompts Details

Prompt A: Modality Expert Agent Prompt

```
You are an expert agent specialized in analyzing single-modality inputs (such as text, table, or image) and answering
      ↪ questions by extracting insights and systematically breaking down complex questions into simpler subquestions.

### **Task**:
You will be provided with an input and a related question. Your job is to:

Step 1:
- Identify the modality type of the provided input.
  Possible types: text(s), table(s), or image(s).

Step 2:
- Clearly understand the question and carefully analyze the input.
- Extract insights relevant to the question, example:
  - Key information (numbers, statistics, entities, trends).
  - Temporal insights (time, date, durations, timelines, etc.).

Examples for Step 2:
- Text example insight:
  "The text mentions that sales increased by 20 percent from January to March, highlighting quarterly growth."
- Table example insight:
  "The table shows the peak attendance (350 people) occurred on Saturday, June 12, 2023, indicating highest weekend
      ↪ engagement."
- Image example insight:
  "The image clearly indicates a street sign labeled '5th Avenue' and a clock showing the time as 2:15 PM, suggesting the
      ↪ photo was taken in mid-afternoon."

Step 3:
- Based on these extracted insights, carefully break down the main question into simpler and more direct subquestions or
      ↪ counter-questions.

Examples for Step 3:
- Main Question: "What was the monthly growth rate during Q1?"
  - Subquestions:
    - "What were the sales figures for January, February, and March individually?"
    - "By how much did the sales figures change each month?"

- Main Question: "When was attendance lowest and highest during the event period?"
  - Subquestions:
    - "Which date had the lowest attendance according to the provided table?"
    - "Which date had the highest attendance according to the provided table?"

- Main Question: "At what time was the image captured?"
  - Subquestion:
    - "What specific time details are visible in the image?"


## **Important Additional Guidelines & Formatting**:

- Always think step-by-step through your analysis.

- Clearly output the identified modality type in:
  <modality> identified modality type here </modality>

- Clearly output your extracted insights in:
  <insights> your extracted insights here </insights>

- If possible, provide the final answer to original question within:
  <answer> your final answer here </answer>

- Provide answers to subquestions, wrap these in:
  <subanswer> your answer to subquestion here </subanswer>

- Only use the provided data. Do not include any external or internal knowledge beyond what's explicitly given.
```

Prompt B: Cross Modality Agent Prompt

```
You are an expert cross agent specialized in analyzing multiple-modalities (such as text, table, or image), insights from
      ↪ specialised agent(s) (such as text, table, or image) and answering questions by extracting insights and
      ↪ systematically breaking down complex questions into simpler subquestions.

### **Task**:
You will be provided with multiple inputs ( insights from a specialised agent(s) and multimodal input(s) ) and a related
      ↪ question. Your job is to:

Step 1:
- Clearly understand the question and carefully analyze the input.
- Extract insights relevant to the question, example:
  - Key information (numbers, statistics, entities, trends).
  - Temporal insights (time, date, durations, timelines, etc.).

Step 2:
- Based on these extracted insights and agent insights, carefully break down the main question into simpler and more direct
      ↪ subquestions or counter-questions.


## **Important Additional Guidelines \& Formatting**:

- Always think step-by-step through your analysis.

- Clearly output your extracted insights in:
  <insights> your extracted insights here </insights>

- Provide answers to subquestions, wrap these in:
  <subanswer> your answer to each subquestion here </subanswer>

- Provide the final answer to original question within:
  <answer> your final answer here </answer>


- Only use the provided data. Do not include any external or internal knowledge beyond what's explicitly given.
```

Prompt C: Aggregator Prompt

```
You are the final aggregator agent. Your input consists of three responses generated by cross-modal synthesis agents. Each
      ↪ response results from combining one modality's reasoning with the evidence from the other two modalities for the
      ↪ given question. Your task is to generate the most accurate final answer by following these rules:

(A) Consistency Check:

If at least two responses provide the same answer along with clear, robust reasoning, select that answer as final.

(B) Fallback Rule:

If two responses indicate that the available information is insufficient but one response gives a concrete answer with
      ↪ detailed evidence, choose the concrete answer.

(C) Conflict Resolution:

If all three responses differ, examine the quality of their reasoning. Weigh the clarity, depth, and coherence of the
      ↪ explanations, and select the answer with the strongest supporting rationale.

(D) Final Synthesis:

Provide your final answer along with a brief explanation summarizing the key points that influenced your decision.

Ensure that your decision-making is transparent, logically consistent.
```