Task-Specific Multimodal Question Answering Agents via Confidence Calibration and Incremental Reasoning for QANTA 2026

Nirjhar Das [1] **Md. Al-Mamun Provath** [1]

Abstract

We present our submission to the QANTA 2026 shared challenge at the ICML 2026 Workshop on Efficient Multimodal Question Answering (EMMQA). Quanta evaluates multimodal quizbowl systems that answer pyramid-style questions from incrementally revealed text and accompanying images while operating under realistic efficiency constraints. The challenge consists of two distinct tasks: Tossup questions, which require deciding when to answer under uncertainty, and Bonus questions, which emphasize accurate answer selection and human adoption. To address these differing objectives, we develop a task-specific two-agent architecture. Our Tossup agent uses GPT-4.1-mini with confidence-calibrated answering and a domain-specific numeric reasoning policy that reduces overconfident predictions from isolated quantitative clues. Our Bonus agent uses GPT-4.1 with leadin-aware reasoning, structured relational reasoning, and multimodal evidence integration to improve exact answer selection. Rather than relying on a retrieval pipeline or model ensembles, our approach emphasizes efficient reasoning policies and confidence calibration within a hosted-only environment. Our system achieved the highest overall leaderboard score of 0.402, including a Tossup score of 0.238 and a Bonus Effect score of 0.164. The results demonstrate that lightweight, task-specific reasoning strategies can provide strong performance on resource-constrained multimodal question answering benchmarks.

1. Introduction

The QANTA 2026 shared challenge studies efficient multimodal question answering in the quizbowl setting, where systems answer pyramid-style questions from incrementally revealed text and accompanying images. Multimodal question answering has been extensively studied in visual question answering and multimodal reasoning settings (Antol et al., 2015). However, quizbowl-style question answering presents a unique challenge because systems must continuously update their beliefs and decide when sufficient evidence exists to commit to an answer (Boyd-Graber et al., 2012; Wallace et al., 2019). Unlike conventional question answering benchmarks that provide complete context, quizbowl requires systems to reason under uncertainty and determine when sufficient evidence exists to commit to an answer. This setting makes confidence calibration and efficient reasoning as important as answer accuracy.

QANTA consists of two complementary tasks. Tossup questions require systems to decide when to answer as clues are revealed, balancing the benefit of early correct answers against penalties for incorrect predictions. Bonus questions provide complete context and emphasize accurate answer selection, calibration, and usefulness to human teammates. The inclusion of multimodal evidence further requires systems to integrate textual and visual information while remaining computationally efficient.

The distinct objectives of the two tasks motivate a task-specific design. We develop a two-agent architecture consisting of a GPT-4.1-mini-based Tossup agent and a GPT-4.1 based Bonus agent. The Tossup agent focuses on confidence-calibrated answering and robust quantitative reasoning, while the Bonus agent emphasizes leadin-aware reasoning, structured answer selection, and multimodal evidence integration. Our system achieved the highest overall leaderboard score of 0.402, including a Tossup score of 0.238 and a Bonus Effect score of 0.164.

The primary contribution of this work is not a novel model architecture but a collection of task-specific reasoning and calibration strategies that improve performance under the efficiency constraints of QANTA 2026. Specifically, we make the following contributions:

- A task-specific two-agent architecture for efficient multimodal question answering
- Confidence-calibrated reasoning policies for incremen-

[1]Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Chattogram, Bangladesh. Correspondence to: Nirjhar Das <nirjhardasami@gmail.com>.

Preprint. June 22, 2026.

tal answering and quantitative reasoning.
- An empirical analysis of multimodal reasoning and calibration in the QANTA 2026 shared challenge.

2. Task Overview

QANTA 2026 is a shared challenge organized as part of the ICML 2026 Workshop on Efficient Multimodal Question Answering (EMM-QA). The competition evaluates systems on quizbowl-style question answering, a setting that combines incremental reasoning, confidence calibration, and multimodal evidence integration under realistic efficiency constraints. Questions span a wide range of domains, including literature, history, science, fine arts, and current events. Unlike conventional question answering benchmarks that provide complete context before inference, QANTA requires systems to operate on partially revealed information and make decisions under uncertainty.

2.1. Tossup Task

Tossups are pyramidal questions revealed incrementally from early, highly distinctive clues to later, more accessible clues. Systems may answer at any point during the clue stream by issuing a buzz. This creates a trade-off between speed and accuracy: answering earlier can increase reward, while incorrect answers incur penalties. Consequently, successful systems must balance answer quality with calibrated decision-making.

Some tossups additionally contain images, requiring systems to integrate visual evidence with incrementally revealed textual clues. Images may include artworks, photographs, maps, diagrams, scientific figures, or OCR-readable text. The central challenge is therefore not only identifying the correct answer, but also determining when sufficient evidence exists to answer confidently.

2.2. Bonus Task

Bonuses are multi-part questions that evaluate depth of knowledge and structured reasoning. Each bonus begins with a leadin that defines a shared theme, category, or relationship across constituent parts. Unlike tossups, bonus questions provide complete context before answering and emphasize precise answer selection rather than timing.

Bonus questions may also include images that provide complementary evidence for one or more parts. In the human-AI collaboration setting, AI outputs serve as recommendations to a human captain who makes the final submission decision. As a result, answer accuracy, confidence calibration, and interpretability are all important factors influencing downstream utility.

2.3. Evaluation Metrics

Evaluation reflects the decision-making requirements of competitive quizbowl. For tossups, primary metrics include expected score, buzz precision, and average buzz position. These metrics jointly measure answer correctness, calibration, and the ability to answer early without excessive risk.

For bonuses, evaluation emphasizes effect, part-level accuracy, question-level accuracy, calibration, and adoption. In addition, systems are assessed under practical efficiency considerations including inference cost, latency, and token usage, aligning with the workshop's focus on efficient multimodal question answering.

The differing objectives of tossup and bonus questions motivate distinct reasoning and decision-making strategies. This observation guided the design of our task-specific architecture, described in the following section.

3. System Architecture

The QANTA 2026 competition presents two distinct question-answering tasks with substantially different objectives. Tossup questions require rapid decision-making under uncertainty, where systems must determine not only the correct answer but also the optimal time to answer. In contrast, Bonus questions prioritize answer accuracy, structured reasoning, and usefulness in a human-AI collaboration setting. We therefore adopt a task-specific architecture consisting of two specialized agents rather than a single unified model.

Figure 1 provides an overview of the proposed system. Incoming multimodal inputs, consisting of textual clues and optional images, are routed according to task type. Tossup questions are processed by a GPT-4.1-mini-based agent optimized for low-latency incremental reasoning and calibrated buzzing decisions. Bonus questions are processed by a GPT-4.1 based agent optimized for precise answer selection, relational reasoning, and human adoption.

3.1. Design Philosophy

Our design is guided by three principles: task specialization, confidence calibration, and efficient multimodal reasoning. First, we treat tossups and bonuses as fundamentally different decision-making problems. A single prompting strategy, that performs well on one task may not be optimal for the other. Second, we emphasize calibrated confidence estimates because both competition tracks depend on reliable strategies that leverage visual information when useful while avoiding unnecessary computational overhead.

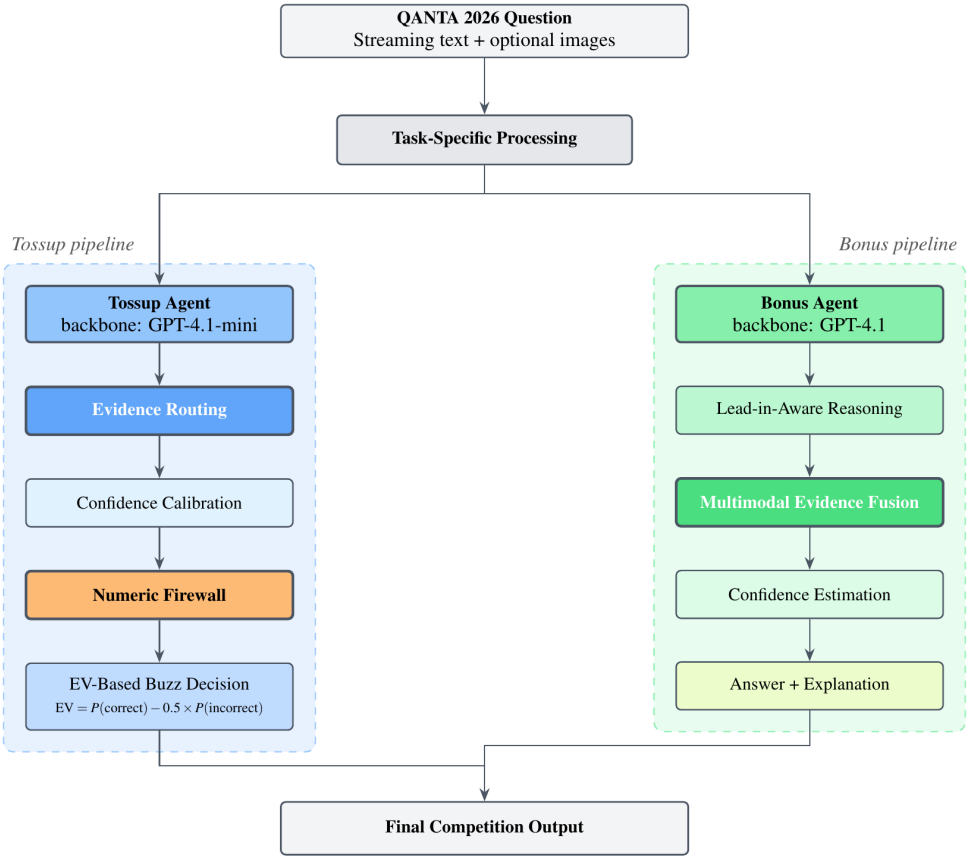

Figure 1. Overview of the proposed QANTA 2026 system. The Tossup agent combines evidence routing, confidence calibration, a Numeric Firewall, and expected-value-based buzzing, while the Bonus agent performs leadin-aware multimodal reasoning and evidence fusion.

3.2. Overall Architecture

The Tossup agent receives incrementally revealed clues and produces both an answer prediction and an associated confidence estimate. These outputs are passed through a confidence-based decision policy that determines whether sufficient evidence exists to issue a buzz. To improve robustness, the system employs a calibration policy that discourages overconfident predictions based solely on isolated numbers, equations, or measurements.

The Bonus agent operates on complete question context and receives the leadin, question text, and any accompanying images. The agent performs leadin-aware reasoning to identify shared themes, categories, or relationships across bonus parts before generating final answers. In addition to answer prediction, the agent produces concise evidence-based explanations intended to facilitate rapid human verification and adoption.

3.3. Multimodal Reasoning Strategy

Rather than treating text and images as equally important at all times, our system adopts an evidence-routing strategy. Textual clues are used to generate the initial hypothesis set, while images serve primarily as supporting or disambiguating evidence. Visual information is therefore incorporated selectively when it provides information not already available in the text.

This process is implemented through a late-fusion strategy with cross-checking between modalities. Candidate answers generated from textual evidence are evaluated against available visual evidence before confidence is increased. This approach aims to reduce unnecessary multimodal processing while maintaining the ability to exploit informative images when present.

The resulting architecture aims to provide an effective balance between accuracy, calibration, and computational efficiency, allowing the system to adapt its reasoning strategy to the differing requirements of tossup and bonus question answering.

4. Tossup Agent

The Tossup task requires systems to answer pyramid-style questions under uncertainty while simultaneously deciding when sufficient evidence exists to commit to an answer. Because incorrect buzzes incur penalties, successful performance depends not only on answer accuracy but also on calibrated confidence estimation. To address these requirements, we employ a GPT-4.1-mini-based Tossup agent optimized for efficient incremental reasoning, confidence calibration, and low-latency inference.

4.1. Agent Design

We selected GPT-4.1-mini as the primary Tossup model due to its favorable balance between accuracy, latency, and inference cost. Unlike bonus questions, tossups require repeated evaluation of incrementally revealed clues, making computational efficiency an important consideration. The agent is instructed to behave as an expert quizbowl player, continuously updating its hypothesis as new evidence becomes available while avoiding overcommitment based on generic or weak clues.

The Tossup agent produces two outputs: an answer prediction and an associated confidence estimate. Images are treated as supporting evidence rather than primary information sources. Textual clues drive initial hypothesis generation, while visual evidence is used selectively for disambiguation when it provides unique information not already present in the text.

4.2. Confidence-Calibrated Buzzing

A key design decision is the separation of answer generation from buzz decisions. Rather than directly trusting model confidence, the system employs a calibrated confidence gate that determines whether answering is beneficial under the competition scoring rules. Recent work has shown that language model confidence estimates can provide useful signals for uncertainty-aware decision making (Kadavath et al., 2022).

The buzzing policy is based on Expected Value (EV) maximization. Specifically, the system seeks to maximize $EV = P(\text{correct}) - 0.5 \times P(\text{incorrect})$, reflecting the reward for correct answers and the penalty for incorrect buzzes. This formulation encourages the agent to balance speed and accuracy rather than aggressively answering at the earliest opportunity. The buzzing decision can be viewed as a form of selective prediction, where the system chooses whether sufficient confidence exists to commit to an answer (Geifman & El-Yaniv, 2017).

During development, we observed that static confidence thresholds often resulted in over-buzzing on generic clue fragments. To address this issue, we adopted position-aware confidence calibration strategies and selected a conservative confidence threshold of 0.90 for GPT-4.1-mini. Earlier experiments using lower thresholds, particularly 0.85, produced substantially more incorrect buzzes and reduced overall expected score.

4.3. Numeric Reasoning Guardrails

A recurring failure mode involved quantitative and scientific questions. The model frequently produced overconfident predictions when exposed to isolated numbers, equations, or measurements without sufficient contextual evidence. These errors often resulted in premature buzzes and unnecessary penalties.

To mitigate this behavior, we introduced a Numeric Firewall, a domain-specific calibration policy designed for mathematics and science questions. The Numeric Firewall was implemented as a prompt-level calibration policy rather than a separate post-processing module. The core principle is that a single numerical clue should not substantially increase confidence unless it is supported by a named entity, formula context, or multiple consistent pieces of evidence. Numeric answers remain valid when explicitly requested by the question or when several independent clues converge on the same conclusion.

This guardrail significantly reduced science-related negations and improved overall calibration, particularly in quantitative domains where numerical patterns alone are often insufficient for reliable identification.

4.4. Development and Error Analysis

The final Tossup system emerged through three major iterations. First, the optimization objective was shifted from raw answer accuracy to expected-value maximization, improving the trade-off between reward and risk. Second, the Numeric Firewall was introduced to address quantitative reasoning failures. Third, confidence calibration was refined through position-aware thresholding and adoption of a more conservative decision threshold.

Common failure mode included over-buzzing on generic clue fragments, confusion between closely related entities, and excessive confidence on isolated numerical evidence. These issues were mitigated through calibrated confidence gating, stronger emphasis on distinctive clues, and the introduction of domain-specific reasoning guardrails.

The resulting system achieved a Tossup score of 0.238 and contributed substantially to the highest overall leaderboard score obtained during the competition.

5. Bonus Agent

Unlike tossups, bonus questions prioritize answer accuracy, structured reasoning, and usefulness in a human-AI collaboration setting. Bonus questions provide complete context before answering and often requires reasoning over shared themes, categories, or relationships introduced through a leadin. To address these requirements, we employ a dedicated GPT-4.1 based Bonus agent optimized for leadin-aware reasoning, multimodal evidence integration, and human adoption.

5.1. Agent Design

We selected the GPT-4.1 model for bonus question answering due to its stronger reasoning, calibration, and disambiguation capabilities. While GPT-4.1-mini provided a favorable efficiency profile for tossups, bonus questions place greater emphasis on answer quality and structured reasoning than on low-latency decision making. In practice, the full model exhibited more reliable behavior on ambiguous questions, relational reasoning tasks, and multimodal examples.

The Bonus agent is prompted as an answer retrieval and disambiguation system rather than an open-ended reasoning assistant. Outputs are intentionally short, structured, and confidence-calibrated. For each question, the model produces an answer, a confidence estimate, and a concise evidence-based explanation intended to support rapid human verification.

5.2. Leadin-Aware Reasoning

A central component of the Bonus agent is leadin-aware reasoning. Bonus questions frequently contain a leadin that establishes a shared theme, category, or relationship across multiple parts. Ignoring this context can lead to inconsistent answer selection or failures in relation reasoning.

The Bonus agent therefore processes the leadin before evaluating individual parts. The leadin is used to constrain the candidate answer space and guide subsequent reasoning. This design encourages consistency across bonus parts while improving answer precision in cases where multiple plausible entities remain under consideration.

5.3. Multimodal Evidence Integration

The Bonus agent adopts an evidence-routing strategy for multimodal reasoning. Textual clues are used to generate the initial hypothesis set, while visual information is incorporated only when it contributes unique evidence. Images are therefore treated as evidence rather than decoration and may provide OCR text, diagrams, maps, scientific structure, artistic styles, or other identifying signals.

To combine modalities efficiently, we employ a late-fusion strategy with cross-checking. Candidate answers are first generated from textual evidence and subsequently evaluated against visual evidence. The image is used to determine whether it supports, refines, or contradicts the text-derived hypothesis. Only candidates that remain consistent across both modalities receive increased confidence.

This approach reduces premature multimodal anchoring while preserving the ability to exploit highly informative visual evidence when available.

5.4. Human Adoption Optimization

In the human-AI competition setting, answer quality alone is insufficient. Prior work has shown that effective human-AI collaboration depends not only on predictive performance but also on the interpretability and trustworthiness of model recommendations (Bansal et al., 2021). Human captains must be able to quickly evaluate and trust the system's recommendations. Consequently, the Bonus agent was explicitly optimized for adoption.

During development, we observed that verbose explanations and hedged responses reduced human willingness to use the model's suggestion. To address this issue, outputs were constrained to concise answer-centered formats with calibrated confidence estimates and short supporting rationales. Confidence scores serve as a lightweight uncertainty signal, while explanations provide sufficient evidence for rapid auditing without overwhelming the user.

5.5. Development and Error Analysis

The final Bonus agent emerged through a series of iterative refinements. Early versions often produced overly verbose explanations that negatively affected adoption. Subsequent revisions emphasized concise answer retrieval, confidence calibration, and structured outputs. Additional improvements focused on reducing entity confusion by prioritizing distinctive clues and delaying commitment when multiple closely related candidates remained plausible.

A second class of failures involved multimodal reasoning and quantitative questions. These issues were addressed through evidence routing, cross-modal verification, and the incorporation of domain-specific reasoning policies inherited from the overall system design. The resulting agent achieved a Bonus Effect score of 0.164 and contributed substantially to the highest overall leaderboard score obtained during the competition.

6. Experiments and Results

6.1. Experimental Setup

Experiments were conducted within the official QANTA 2026 shared challenge hosted as part of the ICML 2026 Workshop on Efficient Multimodal Question Answering (EMM-QA). The competition evaluates hosted multimodal question-answering systems on both Tossup and Bonus tasks. Systems receive streaming textual clues and, for selected questions, accompanying images. Performance is evaluated using task-specific metrics that jointly capture answer quality, confidence calibration, buzzing behavior, and human-AI collaboration effectiveness. The results reported in this paper correspond to the official challenge submissions `QANTA41Mini_V1` for Tossup and `QANTA41_Bonus_V1` for Bonus.

6.2. Tossup Results

Table 1 summarizes the performance of the proposed Tossup agent. The system achieved a Tossup score of 0.238, the highest score on the leaderboard at evaluation time.

Table 1. Final Tossup Results.

Metric	Value
E[Score]	0.238
Buzz Precision	72.5%
Buzz Frequency	94.2%
Buzz Position	60.558
Win Rate	71.1%
Total Evaluation Cost (USD)	$0.14

The agent maintained a Buzz precision of 72.5% while buzzing on 94.2% of questions and achieved a Win Rate of 71.1%. Despite operating under hosted-agent constraints, the system achieved competitive buzzing behavior at an inference cost of only $0.14. The reported cost corresponds to the total API expenditure incurred during the official Tossup evaluation.

6.3. Bonus Results

Table 2 presents the performance of the Bonus agent. The system achieved a Bonus Effect of 0.164 together with 89.1% Part Accuracy and 72.7% Question Accuracy.

The agent additionally achieved strong Calibration (88.2%) and Adoption (33.8%), indicating that its responses were both reliable and useful in the human-AI collaboration setting. Adoption measures the frequency with which human participants accepted the AI recommendation under the official evaluation protocol.

Table 2. Final Bonus Results.

Metric	Value
Bonus Effect	0.164
Part Accuracy	89.1%
Question Accuracy	72.7%
Calibration	88.2%
Adoption	33.8%

6.4. Overall Leaderboard Performance

Table 3 summarizes the final leaderboard performance. Combining the Tossup and Bonus agents yielded an Overall Score of 0.402, corresponding to the top-ranked position on the competition leaderboard at evaluation time. The closest competing system obtained an Overall Score of 0.370, resulting in a margin of 0.032.

Table 3. Overall Leaderboard Performance.

Metric	Value
Overall Rank	1
Overall Score	0.402
Tossup E[Score]	0.238
Bonus Effect	0.164
Bonus Part Accuracy	89.1%
Bonus Adoption	33.8%

These results suggest that task-specific reasoning policies, confidence calibration, and efficient multimodal evidence integration can provide a substantial advantage in resource-constrained multimodal question answering.

7. Error Analysis and Discussion

Although the proposed system achieved the highest overall leaderboard score, several recurring failure modes were observed during development and evaluation. Analyzing these failures proved critical for improving expected score, confidence calibration, and multimodal reasoning performance. This section discusses the primary sources of error, the design modifications introduced to address them, and the broader lessons learned from developing efficient hosted multimodal question-answering agents for QANTA 2026.

7.1. Confidence Calibration

Confidence calibration emerged as one of the most important factors affecting Tossup performance. Confidence calibration has been widely recognized as an important component of reliable machine learning systems, particularly when predictions are used to drive downstream decisions under uncertainty (Guo et al., 2017). In the QANTA setting, incorrect buzzes incur explicit penalties, making overconfident

predictions substantially more costly than delayed but correct answers. Consequently, the objective is not simply to maximize answer accuracy, but rather to maximize expected score through calibrated decision-making.

During early development, the system employed more aggressive buzzing thresholds. While these configurations occasionally produced earlier correct answers, they also resulted in frequent incorrect commitments to semantically related entities. Representative examples included confusion between *Frodo* and *Bilbo*, *Nick Carraway* and *Jay Gatsby*, and *Mormonism* and *Catholicism*. In these cases, the model successfully identified the correct semantic neighborhood but lacked sufficient evidence to distinguish among nearby candidates.

Table 4 summarizes the impact of confidence threshold selection during development using a small set of playground Tossup questions. Although these experiments do not reflect the official leaderboard evaluation, they were useful for understanding the trade-off between buzzing aggressiveness and precision. Lower thresholds increased buzzing frequency but substantially reduced precision and expected score due to incorrect early commitments. In contrast, more conservative thresholds reduced catastrophic errors and improved expected value, motivating the final confidence-gating strategy used in the submitted system.

Table 4. Effect of confidence threshold on a development set of playground Tossup questions.

Threshold	Precision	Buzz Position	Expected Score
0.75	25%	25	-0.2
0.85	75%	39	0.0
0.90	100%	48	+0.5

A key observation was that confidence estimates generated by GPT-4.1-mini were not uniformly reliable across all ranges. Confidence values 0.80-0.85 frequently corresponded to plausible but unstable hypotheses that were later revised as additional clues became available. By contrast, confidence values above 0.90 were significantly more likely to correspond to the final correct answer. This observation motivated the final confidence threshold used in the submitted system.

The final submitted system achieved a Buzz Precision of 72.5% on the official Tossup leaderboard, indicating that the trends observed during development largely transferred to the full competition setting. Overall, these experiments demonstrate that confidence calibration is a critical component of incremental question answering systems. Accurate answer generation alone is insufficient; systems must also determine when available evidence is strong enough to justify an irreversible buzzing decision.

7.2. Related-Entity Confusion

A recurring challenge in incremental question answering is distinguishing between closely related entities that share substantial contextual overlap. In quizbowl, early clues are intentionally obscure and often describe a broader narrative, historical period, scientific domain, or thematic category before revealing uniquely identifying information. As a result, multiple plausible candidates may remain consistent with the available evidence for a significant portion of the question.

This behavior was particularly evident in literary, historical, and religious questions. For example, clues describing a fictional narrative may initially support several characters from the same work, while clues about a religious movement may remain compatible with multiple denominations until a distinctive identifier appears. In such cases, the model often identified the correct answer family early but lacked sufficient evidence to confidently distinguish among neighboring entities.

We found that generic clues were substantially less informative than distinctive clues. Early commitment based on thematic similarity frequently produced unstable hypotheses, whereas rare names, unique events, specific quotations, and highly distinctive attributes were much stronger predictors of the final answer. This observation motivated the final prompting strategy, which emphasized distinctive evidence and encouraged uncertainty when multiple related candidates remained plausible.

These findings suggest that a significant portion of quizbowl difficulty arises not from knowledge gaps but from fine-grained entity disambiguation under incomplete information. Improving robustness to related-entity confusion remains an important direction for future multimodal question-answering systems.

7.3. Lessons Learned

Several practical lessons emerged during system development. First, confidence calibration proved more important than raw answer accuracy in the Tossup task because incorrect buzzes incur explicit penalties. Second, simpler prompting strategies consistently outperformed heavily engineered reasoning protocols for both Tossup and Bonus agents. Third, multimodal evidence was most effective when used selectively for disambiguation rather than as a primary source of information. Finally, domain-specific guardrails, such as the Numeric Firewall used for quantitative questions, improved robustness by preventing overconfidence from limited numerical evidence.

These observations suggest that efficient multimodal question answering benefits from task-specific reasoning policies, calibrated decision-making, and lightweight evidence

integration strategies rather than increasingly complex prompting pipelines.

8. Conclusion

This paper presented a task-specific multimodal question-answering system for the QANTA 2026 shared challenge. Rather than relying on a single unified strategy, the proposed approach employed separate agents for the Tossup and Bonus tasks, allowing each component to optimize for its respective objective. The Tossup agent combined GPT-4.1-mini with confidence calibration, expected-value-based buzzing, and domain-specific reasoning guardrails, while the Bonus agent utilized GPT-4.1 with leadin-aware reasoning and multimodal evidence routing.

Experimental results demonstrated that these design choices were effective under hosted-agent resource constraints. The final system achieved a Tossup E[Score] of 0.238, a Bonus Effect of 0.164, and an Overall Score of 0.402, securing first place on the competition leaderboard at the time of submission. Error analysis further highlighted the importance of confidence calibration, selective multimodal fusion, and robust entity disambiguation for incremental question answering.

Overall, our findings suggest that efficient multimodal question answering benefits more from calibrated decision-making, task-specific reasoning policies, and targeted evidence integration than from increasingly complex prompting pipelines. Future work may explore improved entity disambiguation, stronger quantitative reasoning capabilities, and adaptive confidence estimation methods for incremental multimodal question answering.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

Boyd-Graber, J. et al. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*, 2012.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems Workshops*, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Kadavath, S. et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Wallace, E. et al. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.