CEDA: Cross-modal Evaluation through Debate Agents for Robust Hallucination Detection

Susmit Neogi *
Indian Institute of Technology Bombay
susmit.neogi@iitb.ac.in

Yun Wang
Amazon
yyunwan@amazon.com

Abstract

We present **CEDA**, a novel multimodal framework for detecting hallucinations in large language model outputs through a multi-agent debate approach. While existing methods for black-box LLMs often rely on response sampling and self-consistency checking, our framework leverages a three-fold approach: a multi-agent debate setting to critically examine and debate the authenticity of generated content, a lightweight classifier head on top of LLM-as-a-judge for more calibrated detection, and a confidence estimation to quantify uncertainty in hallucination detection. This debate-based architecture enables a more nuanced and contextual evaluation of potential hallucinations across multiple modalities. Through extensive experiments on five different benchmarks – TruthfulQA, Natural Questions, FEVER, HallusionBench and HaluEval Summarisation, we demonstrate that our approach achieves significant improvements over baseline methods. Our framework also provides interpretable debate traces that help explain the reasoning behind hallucination determinations. These results suggest that structured multi-agent debate systems offer a promising direction for improving the reliability and trustworthiness of language model outputs.

Keywords LLM-as-a-Judge, Retrieval Augmented Generation, Multi-Agent Debate, Hallucination Detection, Uncertainty Quantification, Multimodal Reasoning

1 Introduction

Multimodal Large Language Models (MLLMs) have become the backbone of modern AI systems, powering applications from open-domain question answering to image captioning and visual question-answering. Despite their impressive fluency and versatility, a critical vulnerability persists: hallucinations - generated content that is syntactically plausible but factually incorrect or ungrounded in the input modalities [25]. This issue is especially severe in high-stakes or multi-modal settings, where misleading outputs can propagate falsehoods, undermine user trust, and cause real-world harm.

Despite its critical importance, hallucination detection remains an open and under-defined challenge. Hallucinations can arise from multiple sources: model overconfidence, training data bias, retrieval mismatches, or modality misalignment [27]. As these models are increasingly deployed in real-world systems without guaranteed access to external knowledge sources, detecting hallucinations becomes a central challenge. One key difficulty lies in the lack of ground truth references, particularly in open-ended generation tasks where multiple plausible outputs exist, complicating supervised training, especially in zero-resource settings. In multimodal large language models (MLLMs), hallucinations can arise from complex interactions such as visual misinterpretation or misalignment between visual and textual modalities [26], rendering simple text-only consistency checks inadequate.

*Work done during internship at Amazon

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

Moreover, language models often exhibit overconfidence, generating outputs with high certainty even when incorrect [23]. Techniques based on self-consistency or repeated sampling are not only computationally expensive but also frequently ineffective. Adding to these issues, many real-world applications impose black-box constraints where only final model output is accessible, excluding internal activations, training data, or retrieval logs, making many existing detection strategies impractical in deployment scenarios.

2 Related Works

Detecting hallucinations in large language models (LLMs) is an increasingly pressing challenge, especially as these systems are deployed in open-ended, knowledge-intensive applications. Traditional methods for hallucination detection often involve repeated sampling, response self-consistency, and external fact-checking modules. However, these approaches suffer from high computational costs and limited effectiveness in scenarios with subtle factual distortions, as well as lack of access to model architecture and internal logits in black-box settings. Often, these methods are difficult to be adapted to multimodal settings.

Self consistency and sampling approaches include SelfCheckGPT [3], which detects hallucinations in black-box LLMs by sampling multiple generations and checking semantic agreement. While effective in zero-resource settings, it is computationally expensive and limited to textual outputs. Entailment and fact-verification approaches like FactScore [20] evaluate hallucination at the atomic fact level by decomposing model outputs into claims and verifying them against references. Similarly, FEVER [12] established large-scale fact-checking benchmarks, which inspired methods that ground detection in textual entailment. Retrieval augmented and tool based approaches like FacTool [21] integrates retrieval with verification steps, while REFCHECKER [22] introduces reference-based fine-grained hallucination detection. These approaches are designed for knowledge-intensive tasks where external context is available. There are structured detection techniques like G2LDetect [17], which introduces a global-to-local framework that aggregates evidence across the input to better capture subtle hallucinations. HalluciNot [19] incorporates common-knowledge verification for robust factuality detection. UQLM [16] ensembles multiple scorers (black-box, white-box, LLM-judge) with uncertainty quantification. Recent work has explored debate among multiple LLMs to improve factuality and robustness [4]. While debate has been shown to improve reasoning in both textual and multimodal settings, prior works typically use only single turn debate using two well-defined agents. This makes this method very grounded to a particular modality or task.

Our Contribution We propose a three-fold contribution to the hallucination detection literature in multi-modal large language models (MLLMs). First, we introduce a novel multi-agent debate framework that leverages dynamic prompting, information cross-referencing, and state updates through agents, enabling context-aware reasoning over both visual and textual modalities. This architecture allows agents to iteratively extract, validate, and argue over fine-grained multimodal evidence. Second, we integrate a lightweight classifier head over the LLM-generated judgment to calibrate hallucination predictions more robustly. Third, we use the classifier confidence as a metric to quantify uncertainty in the predictions. This dual-layer verdict system—combining human-like debate with structured classification—offers improved generalizability, interpretability, and reliability. Together, our method presents a scalable, modular pipeline for hallucination detection that can operate effectively in real-world multimodal settings.

3 Methodology

In contrast to prior work, our approach integrates multi-agent debate with a central LLM judge, along with a sentence classifier, enabling structured evaluation of factual consistency across modalities and argument viewpoints. Unlike semantic entropy or zero-resource methods, we do not require token-level entropy modeling or hand-crafted perturbations. Compared to previous multi-agent frameworks, our design includes a more complex agentic interaction and a formal judging mechanism that learns to score arguments and a classifier for final calibrated evaluation. Our method involves cross referencing information from images, texts, documents and all other modalities, resulting in a context-enhanced prompting. This yields superior generalization, reasoning, and factual sensitivity across both vision-language as well as purely textual tasks.

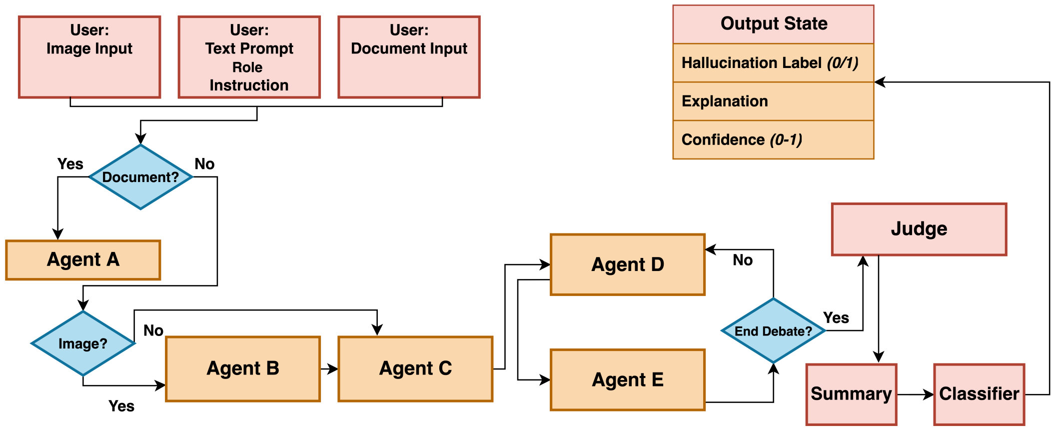

Figure 1: Agentic workflow in CEDA for hallucination detection in multimodal setting

3.1 Agents

Our CEDA framework implements a multi-agent architecture orchestrated through specialized agents with distinct functionalities. As illustrated in Figure 1, the framework employs five key agents operating in a sequential pipeline. Initially, Agent A performs Retrieval Augmented Generation (RAG) [7] to extract relevant context from input documents. The entire document is divided into m chunks, where each chunk typically corresponds to a meaningful unit of text, such as a sentence. Each chunk is encoded into a vector embedding space. The text prompt is also processed as a single chunk. This is done using a text embedding model. Using cosine similarity as the distance metric, we choose top n chunks, closest to the vector embedding of the prompt in the embedding space, to form the context. This is followed by Agent B, which utilizes a multimodal LLM to process and integrate visual information from input images. We generate prompt dynamically, based on the system instructions, to extract more meaningful and aligned information from the images. Agent C then synthesizes the accumulated context with the user query to generate an enriched contextual representation.

The core debate mechanism is implemented through agents D and E, which engage in structured argumentation based on the enriched context. The debate progression is governed by a conditional termination criterion that evaluates inter-agent consensus. Specifically, the framework monitors a two-turn conversation window for complete agreement between agents. Debate termination occurs either upon reaching consensus or exceeding a predefined maximum number of turns.

The final stage involves a judge agent that analyzes the complete debate history to generate a verdict, accompanied by an explanatory summary that ensures interpretability. This output feeds into a binary classifier that determines the hallucination prediction (0: hallucinated, 1: non-hallucinated) with an associated confidence score. This hierarchical architecture enables both robust hallucination detection and transparent decision-making through explicit debate traces.

3.2 Workflow

Our framework implements a dynamic state-based architecture for multimodal hallucination detection. The system operates through three primary state transitions: initial, intermediate, and output, as shown in Figure 2. The initial state incorporates multimodal inputs including images, documents, contextual passages, and system instructions. This information is processed by specialized agents: an image agent for visual analysis and a Retrieval-Augmented Generation (RAG) agent for textual and document processing. These agents update the intermediate (mutable) state, which consists of an evolving system prompt, extracted information, and an ongoing debate history. The intermediate state serves as input to a multi-agent debate framework, where the debating agents engage in structured argumentation, continuously enriching the context and updating the debate history. Finally, a judge agent evaluates the intermediate state to produce the output state, comprising a binary hallucination label (0/1), a confidence score (0-1), and an explanatory rationale. This cascading architecture enables comprehensive cross-modal analysis and facilitates transparent, interpretable hallucination detection.

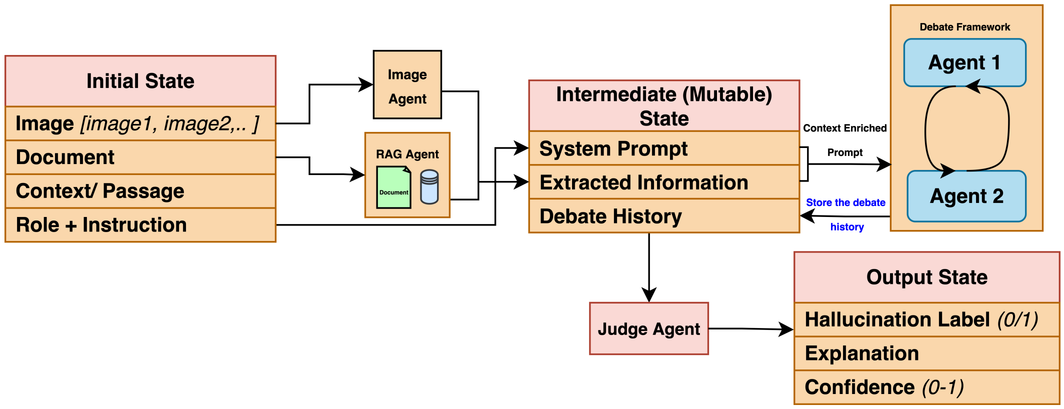

Figure 2: State update workflow in CEDA

3.3 Context Enhanced Dynamic Prompting

A central innovation in CEDA lies in the design of a context-enhanced dynamic prompting mechanism that couples retrieval, multimodal evidence extraction, and debate orchestration. Unlike static prompting strategies, like in case of repeated sampling techniques [3], that provide the same fixed context to all agents, CEDA adapts the prompt dynamically based on intermediate signals from upstream agents. Specifically, the RAG module and image agent first extract heterogeneous evidence, which is synthesized by the context agent into structured claims and supporting facts. These synthesized contexts are then injected into the debate prompts in a selective and role-specific manner: debaters receive potentially conflicting subsets of the evidence to maximize critical examination, while the judge receives an aggregated, conflict-aware prompt that highlights areas of disagreement. This dynamic conditioning encourages agents to argue from complementary perspectives, reduces redundancy, and surfaces subtle inconsistencies that static prompts often miss. As a result, CEDA transforms raw multimodal information into a debate-ready context that is both richer and more adversarially informative. This novel prompting strategy enables the debate to focus on contentious factual claims rather than focusing on redundant context, thereby increasing both the accuracy of hallucination detection and the interpretability of the resulting debate traces.

3.4 Classifier

Following the multi-agent debate process, the final structured response—along with the supporting dialogue and extracted information—is processed by the judge agent, which generates a concise verdict indicating whether the original claim is hallucinated or supported. This verdict is then passed to a downstream classification module to generate a final hallucination label and associated confidence score. In our framework, we employ the SetFit [6] classifier, a parameter-efficient few-shot learning method built upon sentence transformers. In our case we use paraphrase-mpnet model [2] for the sentence transformer. SetFit finetunes sentence embeddings using contrastive learning without full backpropagation through the transformer. Let $f_\theta(\cdot)$ denote the sentence encoder with parameters θ. Given a pair of semantically similar sentences (x_i, x_j^+) and a dissimilar negative sample x_k^-, the sentence transformer is finetuned using contrastive learning.

After embedding alignment, classification head (in this case, logistic regression) is trained on the labeled hallucination dataset. Let $z_i = f_\theta(x_i)$ be the embedding of the i-th verdict and $y_i \in \{0, 1\}$ be its hallucination label. The classifier is trained using binary cross-entropy loss:

$$\mathcal{L}_{\text{classification}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\sigma(A_i)) + (1 - y_i)\log(1 - \sigma(A_i))] \tag{1}$$

where $A_i = w^\top z_i + b$, $\sigma(\cdot)$ is the sigmoid function, and (w, b) are the classifier weights and bias.

[2] huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

This two-stage architecture enables highly data-efficient training and provides a calibrated confidence score via the sigmoid output. This makes it especially suitable for hallucination detection in settings where supervised data is limited. The classifier thus transforms abstract debate verdicts into interpretable hallucination predictions, supporting robust downstream applications.

3.5 Confidence

The uncertainty in CEDA's hallucination predictions can be quantified by a confidence scoring mechanism. This score is derived from the posterior probabilities obtained from the SetFit classifier's output layer. When there is strong evidence of hallucination through clear agent consensus, the logits probability is close to 1 for the hallucinated class. Conversely, for clearly non-hallucinated content, the probability distribution peaks near 1 for the non-hallucinated class. In cases of agent disagreement or ambiguous evidence leading to uncertain judge verdicts, the probability distribution approaches uniformity (≈ 0.5), indicating low confidence in the classification. More specifically, if p and $1-p$ are the class probabilities of hallucinated and non-hallucinated class respectively, the confidence score can be defined as $\max(p, 1-p)$. This probabilistic formulation provides a natural measure of prediction uncertainty that aligns with the framework's ability to detect hallucinations, effectively capturing the degree of confidence in the model's decisions.

4 Experiments

4.1 Agent Model Selection

In our multi-agent framework, we use different models for different agents. For parsing the documents using RAG, we use the base model of Amazon Titan Embed Text V1 for text embeddings. For extracting information from images, we use Claude 3.5 Sonnet. For the debating agents, we use Mistral 7B [28] and Claude 3.5 Haiku. Two different LLMs are used for the agents to reduce any information or inductive bias present in a single LLM. For the debate summarization and judgment by the judge agent, we use Claude 3.5 Sonnet.

4.2 Training

For the classification component, we employ the SetFit classifier, that has demonstrated robust generalization capabilities across diverse domains. The classifier is trained on debate summaries generated by our framework using a training set of 1000 samples from the FEVER [12] dataset. These samples are processed through our multi-agent debate pipeline to generate structured verdicts, which serve as training instances for the SetFit classifier (0: hallucinated, 1: non-hallucinated). Notably, this single classifier, trained on a small, domain-specific sample, generalizes effectively across all benchmark datasets without fine-tuning, demonstrating the framework's domain-agnostic capabilities and efficient few-shot learning performance.

4.3 Datasets and Experimental Setup

We evaluate our hallucination detection framework across five diverse datasets: TruthfulQA, Natural Questions (NQ), FEVER, HallusionBench, and HaluEval-Summarisation. For each, we benchmark binary classification performance using F1 score and accuracy. We also report the average time per evaluation sample.

The TruthfulQA [8] dataset focuses on detecting false beliefs and misinformation across domains like health, law, and common misconceptions. Using 1,000 validation samples, we create positive (truthful) and negative (hallucinated) question-answer pairs. Each pair is verified using our framework, and binary F1 score is computed against the labeled ground truth. The Natural Questions [11] dataset consists of real search queries and associated short answers. We evaluate 1,000 samples from the validation set, where ground truth only contains non-hallucinated answers. Our model classifies hallucination presence, and we report the percentage of correct non-hallucinated predictions in terms of the F1 score. The HaluEval-Summarisation [14] dataset includes document-summary pairs labeled as factual or hallucinated based on human assessments. It evaluates a model's ability to detect both subtle and overt factual inconsistencies in generated summaries, providing a robust benchmark for hallucination detection. The FEVER [12] dataset is a fact verification dataset containing claims labeled as supported or refuted. For the purpose of hallucination detection, we drop the class of non verifiable data. We sample 1,000 test instances, treating supported as non-hallucinated and refuted as hallucinated. The model checks factuality, and we report binary classification F1 score. The HallusionBench [13] dataset assesses hallucinations in vision-language models using 1,129 yes/no questions over 346 image pairs. Questions are categorized as Visual Dependent (VD) or Visual Supplement (VS). The model classifies each as hallucinated or not, and we report accuracy.

4.4 Results

We evaluate our framework on all five benchmarks using the best performing parameters. These parameters are (i) Maximum number of debate turns (= 5); (ii) Maximum number of output tokens (= 100); (iii) Number of document chunks for context using RAG ($n = 5$)(iii) Combination of different LLMs for different agents as mentioned in previous section. The experiments and results are discussed later.

4.4.1 TruthfulQA

We report the zero-shot prompting performance of LLMs like Qwen, GPT-4o, and GPT-4o mini. The HDM model [19] uses context and common knowledge verification for hallucination detection. The MAD (Multi-agent debate) model [15] implements a multi-agent debate approach for hallucination detection. Our framework performs almost at par with the best performing HDM model. However, our framework performs better than the current multi-agent debate based approach. We report the baseline metrics in table 1. The baseline results are reported from original papers [19][15]. The average evaluation time is reported to be **10.39 seconds** per evaluation sample.

Table 1: TruthfulQA performance comparison

Model	F1 Score (%)↑
GPT-4o	53.8
GPT-4o mini	56.2
Qwen	61.6
MAD	61.9
HDM	**83.7**
CEDA	81.3

Table 2: NQ performance comparison

Model	F1 Score (%)↑
Gemini	48.00
GPT-3.5	55.00
UQLM (Gemini)	57.00
UQLM (GPT-3.5)	65.00
CEDA	**82.10**

4.4.2 NQ

We report the zero-shot prompting performance of LLMs like GPT 3.5 and Gemini. The ensemble scoring method introduced in the paper Uncertainty Quantification for Language Models [16] involves aggregating black box, white box as well as LLM-as-a-judge methods for scoring and using a threshold to classify a question-answer pair as hallucinated or not. The baseline results are reported from original UQLM paper [16]. Our framework outperforms the existing baseline methods. We report the baseline metrics in table 2. The average evaluation time is reported to be **8.74 seconds** per evaluation sample.

4.4.3 HaluEval-S

We evaluate our framework on the HaluEval-S (summarisation) dataset and compare its performance against a range of hallucination detection baselines, as reported in the G2LDetect [17] paper. The methods include zero-shot prompting of GPT-4 LLM as well as chain-of-thought (CoT) prompting. Other methods include self-consistency like SelfCheckGPT [3], entailment scoring like FactScore [20], and retrieval-augmented tools like FacTool [21] and REFCHECKER [22]. We also report the G2LDetect model performance, which uses global to local information extraction from the text to detect hallucination. Our method outperforms the best performing setting of G2LDetect (using GPT-4). The baseline results are reported from original papers [17]. We report the baseline metrics in table 3. The average evaluation time is reported to be **13.78 seconds** per evaluation sample.

Table 3: HaluEval-S Performance Comparison

Model	F1 Score (%)↑
GPT-4	65.88
GPT-4 (CoT)	66.77
REFCHECKER	69.00
SelfCheckGPT	69.81
FactScore	70.33
FacTool	70.93
G2LDetect	72.98
CEDA	**75.00**

Table 4: FEVER Performance Comparison

Methods	F1 Score (%)↑
FactScore	66.58
GPT-4	70.11
FacTool	70.21
GPT-4 (CoT)	71.13
SelfCheckGPT	72.64
REFCHECKER	72.76
G2LDetect	75.55
CEDA	**85.00**

4.4.4 FEVER

We evaluate our framework on the FEVER dataset and compare its performance against the hallucination detection baselines mentioned in previous section, as reported in the G2LDetect [17] paper. Our method outperforms the best performing setting of G2LDetect (using GPT-4). The baseline results are reported from original papers [17]. We report the baseline metrics in table 4. The average evaluation time is reported to be **8.85 seconds** per evaluation sample.

4.4.5 HallusionBench

We report the accuracy score of state-of-the-art LLMs on the entire dataset, as provided in their official GitHub [3] repository. Our model outperforms the baseline performance. We report the metrics in table 5. The average evaluation time is reported to be **15.10 seconds** per evaluation sample.

Table 5: HallusionBench Performance Comparison

Model	Accuracy (%)↑
mPLUG-Owl	43.93
InstructBLIP	45.26
LLaVA-1.5	46.94
mPLUG-Owl2	47.30
Claude 3	56.86
GPT-4o	65.28
Human Eval	67.58
CEDA	**72.36**

Table 6: Ablation study results across five benchmarks. We report F1 scores (%) and computation time (seconds) for different configurations of debate turns and agent combinations. N denotes the maximum number of debate turns; debating agents A_1 and A_2 denote Agent 1 and Agent 2 respectively.

N	$\mathbf{A_1}$	$\mathbf{A_2}$	TruthfulQA		FEVER		NQ		Hallusion		HaluEval	
			F1	Time	F1	Time	F1	Time	F1	Time	F1	Time
2	Mistral	Haiku	75.0	8.69	80.0	7.60	87.0	6.95	63.0	14.92	72.0	10.43
4	Mistral	Haiku	78.0	9.91	82.0	8.19	88.0	8.22	66.0	15.02	73.0	10.88
6	Mistral	Haiku	**81.0**	10.47	**85.0**	9.43	90.0	8.47	**71.0**	15.13	**74.0**	13.88
8	Mistral	Haiku	80.0	11.39	85.0	10.22	91.0	8.82	70.0	16.09	74.0	13.91
6	Mistral	Mistral	71.0	12.72	82.0	10.11	88.0	9.12	66.0	17.07	57.0	12.22
6	Haiku	Haiku	73.0	14.48	83.0	9.19	**94.0**	7.30	62.0	16.19	62.0	17.27

[3] github.com/tianyi-lab/HallusionBench

5 Ablation Studies

We conduct extensive ablation experiments to analyze the impact of different components and configurations in CEDA. Specifically, we investigate (1) the optimal number of debate turns, (2) the effect of agent model combinations, and (3) their impact on computational efficiency across five diverse benchmarks. The analysis was done to assess relative performance among different settings and not against the state-of-the-art models. The experiments were performed on a small subset of 1000 samples of each benchmark dataset and the results differ from the absolute results reported in earlier sections. These results are presented in table 6.

5.1 Impact of Number of Debate Turns

We first examine how the number of debate turns affects model performance while keeping other parameters constant (Mistral-Haiku agents, temperatures 0.2/0.6). As shown in table 6, increasing the number of turns from 2 to 6 consistently improves performance across all benchmarks. On FEVER, the F1 score improves from 80% to 85%, while HallusionBench shows an improvement from 63% to 71%. However, further increasing to 8 turns shows no significant change in performance. This shows, in most settings, the agents reach to conclusive agreement within atmost 6 iterations.

Different tasks show varying sensitivity to the number of turns. Text-based benchmarks like NQ show more stable improvements (87% $\rightarrow$ 90% $\rightarrow$ 91%), while multimodal tasks like HallusionBench demonstrate more pronounced gains (63% $\rightarrow$ 71%) before plateauing. This suggests that visual reasoning benefits more from extended debates up to a certain point.

5.2 Temperature Analysis

Across all experiments, we keep temperature of one of the agents to be 0.2 and other to be 0.6. The hypothesis is, a higher temperature provides more variation in output probabilities [24], according to equation 2:

$$p(x_i|\mathbf{x}_{<i}) = \frac{\exp(z_i/T)}{\sum_{j=1}^{N}\exp(z_j/T)} \tag{2}$$

where z_i represents the logit for token i, T is the temperature parameter. This often leads to more creative and diverse responses. On the other hand, lower temperature often results in more straightforward responses. As an attempt to balance these attributes, we maintain this combination of temperatures.

5.3 Agent Model Combinations

We compare three 6-turn debate setups: (1) heterogeneous (Mistral–Haiku), (2) homogeneous (Mistral–Mistral), and (3) homogeneous (Haiku–Haiku).

As shown in table 6, the heterogeneous configuration consistently outperforms homogeneous setups in most benchmarks. On TruthfulQA, Mistral-Haiku achieves 81% F1 score compared to 71% and 73% for homogeneous configurations. However, we observe interesting exceptions: on NQ, Haiku-Haiku achieves the best performance (94% F1), suggesting that model selection should consider task-specific characteristics. This can be attributed, as mentioned earlier, to the inherent biases in individual LLMs, which get mitigated when using the ideal combination of LLMs as debating agents against each other.

5.4 Computational Efficiency Analysis

Our experiments reveal important efficiency trade-offs. While increasing debate turns generally improves performance, it comes with linear increase in computation time. For instance, on FEVER, time increases from 7.6s to 10.22s when increasing turns from 2 to 8. The efficiency impact varies by task complexity - HallusionBench shows consistently higher computation times ($\sim$20s) due to its multimodal nature. Notably, agent combination choices significantly affect efficiency. The Mistral-Haiku configuration generally offers the best performance-efficiency trade-off, while homogeneous configurations sometimes require longer processing times without corresponding performance benefits. These findings validate our design choices while highlighting important considerations for deploying CEDA in different scenarios.

5.5 Agents Ablation

Individual Agent Contribution A typical ablation study cannot be performed by removing some agents and analysing the model performance. This is because each agent performs a strong role in the unified multimodal framework, attending to different modalities of data. However, number of debate turns, which we discussed in previous section, controls the number of times an agent is invoked in the overall framework. We clearly see that as debate turns increase, effective number of agents involved in the framework increases, resulting in improvement in performance. The invocation of agents is completely dynamic based on conditional statements, meaning, in a simpler condition with easy convergence of debate, the number of turns is lesser than a difficult condition to asses, where agents require multiple rounds of debate to converge to a conclusion. Hence, for simpler use cases, the number of debate turns can be one, working as a simple LLM-as-a-judge.

Multi-Agent Debate vs LLM-as-a-Judge From the performance on benchmarks, discussed in the experiments section, we see that our multi-agent debate framework outperforms a single LLM-as-a-judge based on commercial LLMs like GPT 4o, Claude, Gemini, as well as white-box open source LLMs like Mistral, Qwen, LLaVa. This can be attributed to the corollary A.0.1, which states that increasing the number of agents improves the performance of the predictive model. This empirical and theoretical argument explains why our unified multimodal framework outperforms single multimodal LLM-as-a-judge. Moreover, in our framework, each agent, specialized in a particular task/modality, presents its views in a structured debate, which is aggregated by the judge to reach the final conclusion. This is different from multimodal LLMs, where multimodal heads use a shared backbone for prediction, limiting diversity and causing redundancy in bias. Thus, multiple agents are more than just modality processing units; they help in removing biases which are one of the major reasons for hallucination in LLMs.

Classifier Head on top of LLM-as-a-Judge In our unified framework, we have a classifier head on top of LLM-as-a-Judge for better calibrated response. This is to bring a degree of determinism in the outputs of the LLM-as-a-Judge, which is a black-box model. To reduce the computational cost, we cap the number of debate turns to an upper limit, this may cause the end of debate before the agents reach a conclusion. As a result, the debate summary may be ambiguous and lead to incorrect judgments by the judge LLM. To solve this, we add a trainable classifier layer on top, which is trained on both clear consensus as well as ambiguous situations, thus bringing more determininism and calibrated responses. This is vital for the success of downstream tasks dependent on the response from our framework.

6 Conclusion

This work introduces CEDA, a novel multi-agent debate framework for detecting hallucinations in large language model outputs across multiple modalities. Through extensive experiments on five diverse benchmarks, we demonstrate that CEDA achieves strong performance in identifying factual inconsistencies and hallucinations, outperforming existing approaches and establishing new state-of-the-art performance. The framework's unique combination of structured multi-agent debate, calibrated classification, and confidence estimation enables robust hallucination detection while providing interpretable reasoning and explanation. The framework's ability to generalize across different tasks and modalities, while maintaining reasonable evaluation times, demonstrates its practical utility for real-world applications. Key strengths of our approach include (1) domain-agnostic architecture that effectively handles both visual and textual modalities, (2) interpretable debate traces that explain hallucination determinations, (3) lightweight classifier head for calibrated hallucination detection, (4) confidence scoring to quantify uncertainty in predictions.

While CEDA demonstrates promising results in hallucination detection, several important directions remain for future exploration. These include scaling and extending the framework to handle streaming data, incorporating active learning for continuous improvement, and developing more efficient debate strategies. We believe CEDA represents a promising direction for improving the reliability and trustworthiness of language model outputs through structured multi-agent evaluation.

References

[1] S. Farquhar, J. Kossen, L. Kuhn, et al. 2024. Detecting hallucinations in large language models using semantic entropy. Nature 630, 625–630.

[2] X. Qiu and R. Miikkulainen. 2024. Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS '24).

[3] P. Manakul, A. Liusie, and M. J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 9004–9017.

[4] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. 2024. Improving factuality and reasoning in language models through multi-agent debate. In Proceedings of the 41st International Conference on Machine Learning (ICML '24), Vol. 235. JMLR.org, 11733–11763.

[5] Z. N. Lin, Z. Wang, Z. Xu, and Yinghui. 2024. Interpreting and Mitigating Hallucination in MLLMs through Multi-agent Debate. arXiv preprint. DOI:10.48550/arXiv.2407.20505.

[6] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg. 2022. Efficient Few-Shot Learning Without Prompts. arXiv preprint arXiv:2209.11055.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, and S. Riedel. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS '20).

[8] S. Lin, J. Hilton, and O. Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 3214–3252.

[9] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 1601–1611.

[10] J. Welbl, N. F. Liu, and M. Gardner. 2017. Crowdsourcing multiple choice science questions. arXiv preprint arXiv:1707.06209.

[11] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. 2019. Natural Questions: A Benchmark for Question Answering Research. Transactions of the Association for Computational Linguistics 7, 452–466.

[12] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 809–819.

[13] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, et al. 2024. HALLUSIONBENCH: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. arXiv preprint arXiv:2403.09653.

[14] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. arXiv preprint arXiv:2305.11747.

[15] R. Li, T. Bagade, K. Martinez, F. Yasmin, G. Ayala, M. Lam, and K. Zhu. 2024. A Debate-Driven Experiment on LLM Hallucinations and Accuracy. CoRR abs/2410.19485.

[16] D. Bouchard and M. S. Chauhan. 2025. Uncertainty Quantification for Language Models: A Suite of Black-Box, White-Box, LLM Judge, and Ensemble Scorers. arXiv preprint arXiv:2504.19254.

[17] X. Cheng, Z. Tan, Z. Zheng, and W. Lu. 2025. G2LDetect: A Global-to-Local Approach for Hallucination Detection. Proceedings of the AAAI Conference on Artificial Intelligence 39, 1 (2025), 102–109. DOI:10.1609/aaai.v39i1.31985.

[18] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 6449–6464.

[19] B. Paudel, A. Lyzhov, P. Joshi, and P. Anand. 2025. HalluciNot: Hallucination Detection Through Context and Common Knowledge Verification. arXiv preprint arXiv:2504.07069.

[20] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv preprint arXiv:2305.14251.

[21] I.-C. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, and P. Liu. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. arXiv preprint arXiv:2307.13528.

[22] X. Hu, D. Ru, L. Qiu, Q. Guo, T. Zhang, Y. Xu, Y. Luo, P. Liu, Y. Zhang, and Z. Zhang. 2024. RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models. arXiv preprint arXiv:2405.14486.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML '17), Vol. 70. PMLR, 1321–1330.

[24] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous. 2024. Is Temperature the Creativity Parameter of Large Language Models? arXiv preprint arXiv:2405.00492.

[25] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys 55, 12 (2023), 1–38. DOI:10.1145/3571730.

[26] Q. Yu, J. Li, L. Wei, L. Pang, W. Ye, B. Qin, S. Tang, Q. Tian, and Y. Zhuang. 2024. HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), 12944–12953. DOI:10.1109/CVPR52733.2024.01230.

[27] M. Cleti and P. Jano. 2024. Hallucinations in LLMs: Types, Causes, and Approaches for Enhanced Reliability. DOI:10.13140/RG.2.2.12184.61445.

[28] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, W. El Sayed, et al. 2023. Mistral 7B. arXiv preprint arXiv:2310.06825.

A Appendix

A.1 Convergence of Multi-Agent Debate

Let X be the event "statement is non-hallucinated" and let $\mu = \Pr(X)$ be its true probability (for a fixed input x). For $i = 1, \ldots, n$ let agent i output a probability estimate

$$p_i := \sigma(f_i(x)) \in [0, 1],$$

where $\sigma(\cdot)$ is the sigmoid and f_i denotes the LLM backbone used by agent i. Define the ensemble (average) estimator

$$\widehat{p}_n := \frac{1}{n} \sum_{i=1}^{n} p_i.$$

Corollary A.0.1 *For any $\varepsilon > 0$,*

$$\Pr\left(|\widehat{p}_n - \mu| \geq \varepsilon\right) \ \leq \ 2\exp\big(-2n\varepsilon^2\big).$$

Consequently $\widehat{p}_n \xrightarrow{p} \mu$ (convergence in probability) as $n \to \infty$, and for fixed ε the Hoeffding upper bound decays exponentially in n. In particular, for any fixed ε the ensemble average achieves strictly tighter concentration than a single agent: the analogous single-agent Hoeffding bound is $2\exp(-2\varepsilon^2)$, so the ensemble improves the exponent by a factor n.

We make the standard assumptions that the LLM based agents are bounded and unbiased estimators of the task (hallucination detection, in this case). Hence, $\mathbb{E}[p_i] = \mu$ for all i (or equivalently $\mathbb{E}[\widehat{p}_n] = \mu$). Also, the above corollary requires independence of the variables. This may not completely hold in our case, as the agents are involved in a structured debate, hence the response of agent 1 (R_1) depends on the response of agent 2 (R_2). So, $\Pr(R_1|R_2) \neq \Pr(R_1)$. But, depending on the degree of correlation, n can be replaced by an effective n_{eff} number of independent variables.

A.2 Comparative Analysis

Entailment-based Detection Entailment-based detectors work by decomposing outputs into claims and checking them against the input or retrieved evidence using pretrained entailment models. While effective in narrow domains like FEVER, these methods often break down when claims are paraphrased, spread across long contexts, or when the entailment model faces out-of-distribution phrasing. They are also brittle across tasks, as the underlying entailment models are typically trained on specific datasets. CEDA surpasses these approaches because the debate mechanism allows agents to challenge claims with contextually relevant counterarguments, while the judge and classifier distill disagreements into more robust binary decisions. Empirically, this leads to large gains — for instance, CEDA beats G2LDetect by nearly 10 F1 points on FEVER — showing its capacity to generalize beyond what static entailment classifiers can achieve.

Prompt-based Detection Prompt-based methods like SelfCheckGPT and HDM attempt to detect hallucinations by re-querying the same model or a verifier model and then measuring consistency across answers or having the verifier judge correctness. These methods are attractive because they require no training, but they suffer from self-confirmation bias: if a model hallucinates consistently, repeated queries or a same-family verifier may reinforce the error. HDM incorporates commonsense verification and performs very well on TruthfulQA, but it remains a single-agent framework. CEDA's debate architecture instead introduces heterogeneity by using agents with different models and temperatures, ensuring diverse perspectives are represented. This increases the likelihood that a hallucination is identified and challenged. While HDM edges out CEDA on TruthfulQA, CEDA clearly outperforms across broader, less structured tasks such as NQ, FEVER, and HallusionBench, highlighting that debate yields stronger generalization beyond benchmark-specific tuning.

Retrieval-based Detection Retrieval-augmented verifiers improve hallucination detection by grounding claims against external corpora, comparing outputs with retrieved passages for consistency. This works well when retrieval captures clean, relevant evidence, but in practice retrieval often brings in partial, noisy, or conflicting context. Most RAG-based baselines use shallow entailment checks, which fail when interpretation or multi-step reasoning is required. CEDA integrates retrieval more effectively by placing it within a debate: one agent specializes in evidence retrieval, while the debaters use this evidence to argue for or against claims. This allows the system to confront noise, weigh evidence, and refine interpretations, rather than blindly accepting retrieval results. The empirical edge is clear on FEVER and NQ, where retrieval is central: CEDA dominates baselines that rely on retrieval alone, confirming the added value of adversarial reasoning layered over evidence.

Multimodal Detection Multimodal baselines typically use a vision-language model (VLM) to generate answers conditioned on both image and text, and hallucination detection is attempted either by re-querying the same model or by checking consistency across image-text pairs. While powerful, VLMs have characteristic blind spots: they over-rely on textual priors, misinterpret subtle visual cues, and struggle with adversarial or ambiguous images. As a result, they often hallucinate objects or relations absent from the visual scene. CEDA improves on this by combining a vision-specialized agent with text-specialized debaters, ensuring that visual evidence is explicitly cross-examined during the debate.

Uniqueness of CEDA CEDA's core distinction is that it does not rely on a single static heuristic — whether entailment, entropy, re-prompting, or retrieval, but instead frames hallucination detection as a process of adversarial dialogue. By forcing heterogeneous agents to present and contest claims, CEDA uncovers inconsistencies that would otherwise be invisible to single-agent or static baselines. The LLM judge provides an interpretable synthesis of the debate, while the classifier ensures calibrated, generalizable predictions. Theoretically, this combination mitigates bias, leverages diversity, and strengthens cross-domain robustness. Empirically, this design translates into state-of-the-art results on four out of five benchmarks, especially in tasks requiring reasoning over noisy retrievals or multimodal inputs.

A.3 Prompts

A.3.1 System Prompt

The system prompt is generated by appending the *Role* and *Instruction* inputs obtained from the user. The system prompt can be formatted as:

```
System: You are an expert debate agent. Your role is [Role].
You are assigned the following task: [Instruction]
```

A.3.2 Image Agent Prompt

As discussed earlier, we generate prompt dynamically based on the task and the input text (fact) to be verified. This prompt is generated by passing a meta-prompt to the LLM. This generated prompt – the image prompt, specifies what information is needed to answer the visual question, and guides the image agent to extract meaningful information from the image. The meta-prompt is:

```
System: You are an image extractor and visual question-answering agent.
Your system prompt is [system prompt].
Based on an image provided you have to answer/verify the following:
[input text].
Your response should be:
Generate a detailed prompt to extract the necessary information from the image
to answer this question.
```

Now the output of the LLM using meta prompt is, say, *image prompt*. The final prompt passed to the image agent is formatted as following:

```
System: [image prompt]
Your response should be:
1. Answered in a single paragraph
2. Do not ask leading questions
3. Try to give a conclusive response
```

The extracted information from the image, as well as from the text documents, are stored in the intermediate state, as shown in figure 2. We have discussed these in detail, with examples in later sections.

A.3.3 Debate Agent Prompt

The debate agents argue about the correctness of the fact based on the extracted information and the previous opinion by opposite agent. This can be accessed from the state history and we call it *last message*.

```
System: [system prompt]. You are a debating agent, debating the question-answer
↪  pairs. Respond very briefly in 1-2 sentence. Think analytically and
↪  rationally.
Current statement or last point: [last message]
Original fact to be checked: [input text]
Extracted context from documents and images: [extracted_info]
Your response should be:
First state in ONE word whether you 'AGREE' or 'DENY' with the last message,
then continue with a brief justification. If needed, refer back to facts from
↪  the context above.
Respond with your very brief argument or counter-argument.
Before answering, try to evaluate your response analytically/ numerically if
↪  possible.
Don't ask leading questions at end.
Try to give a very short conclusive answer"
```

A.3.4 Judge Agent Prompt

After the end of debate, the entire history is appended to form a *conversation*. This is passed to the judge agent to summarise and provide final verdict. The judge agent prompt is formatted as follows:

```
System: You are the neutral judge of a debate between two agents discussing a
factual statement.
Your task: Conclude from the debate whether the fact is correct or wrong.
Debate Transcript: [conversation]
Your response should be:
Has consensus been reached? If so, very briefly give a concluding summary based
on your judgement.
If the agents disagree, being neutral judge, you give your conclusion very
briefly.
No need to individually summarise every agent response
Give a brief summary of how the agents agree or disagree over every point
discussed.
Final verdict whether the statement is (1) correct; (0) wrong"
```

Example 1

We will demonstrate the prompts and LLM responses at each step, while evaluating a fact based on a text reference provided.

We use a sample text document from Huggingface. [4] It contains trivia about cats.

The fact to be verified is:

```
Female cats tend to be right pawed.
```

Without the reference document, the LLM cannot provide confident response. However, with reference document, better context is extracted from the document and hence, LLM provides correct response.

For this we pass the following system prompts – Role:

```
Fact Verifier
```

[4] huggingface.co/ngxson/demo_simple_rag_py/blob/main/cat-facts.txt

Instruction:

```
Verify from the document if the fact is correct or hallucinated
```

The extracted information from the image is:

```
Info extracted from documents: Female cats tend to be right pawed, while male
cats are more often left pawed. Interestingly, while 90% of humans are
right handed, the remaining 10% of lefties also tend to be male. Cats have
five toes on each front paw, but only four toes on each back paw. A female
cat is called a queen or a molly. All cats have claws, and all except the
cheetah sheath them when at rest. The claws on the cat’s back paws aren’t as
sharp as the claws on the front paws because the claws in the back don’t
retract and, consequently, become worn.
```

The following is the debate structure:

```
Agent 1- agree: the document provides evidence that female cats have a
preference for using their right paw more often than their left paw. the
statement is supported by the information presented.

Agent 2- agree: the extracted context directly supports the statement,
explicitly stating that "female cats tend to be right pawed, while male
cats are more often left pawed." the information is clear and provides a
direct confirmation of the original claim about female cats' paw
preference.
```

The verdict provided by the judge after debate:

```
Hallucinated: False
Judgement Summary: Based on the debate transcript, both agents AGREE that the
statement "Female cats tend to be right pawed" is correct. They both cite
supporting evidence from the document that directly confirms the claim.
Verdict: CORRECT The fact is verified and supported by the context provided in
the debate.
Confidence: 0.9601831532567048
```

Example 2

In this example, we will demonstrate visual hallucination detection using CEDA. Figure 3 contains a chart and involves answering question referring to the same. Figure 2 contains an illusion, often leading to hallucination by LLMs. These images are taken from HallusionBench dataset.

Chart-based Image

This fact is referenced to the figure 3. The fact to be verified is:

```
Is China, Hongkong SAR, the leading importing country of gold, silverware, and
jewelry with the highest import value in 2018?
```

For this we pass the following system prompts –

Role:

```
Fact Verifier
```

Instruction:

```
Verify from the image if the fact is correct or hallucinated
```

The extracted information from the image is:

```
Info extracted from image: image_0: Based on the image provided, we can verify
the gold, silverware, and jewelry import statistics for
```

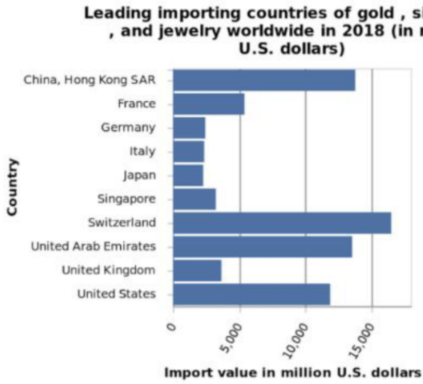

Figure 3: Chart-based Image

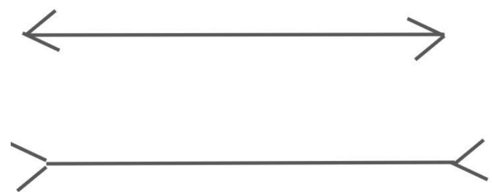

Figure 4: Illusion-based Image

```
China (Hong Kong SAR) in 2018.
The chart clearly shows that China, Hong Kong SAR was one of
the leading importing countries for these products worldwide in 2018.
The import value for China (Hong Kong SAR) is represented by a blue bar that
extends to approximately 11,000 million U.S. dollars, making it the
second-highest importer among the countries listed. This places China
(Hong Kong SAR) behind Switzerland, which shows the highest import value, and
ahead of other major economies such as the United States, France, and
Germany. The data is presented in a clear, comparative format, allowing for
easy verification of China\'s position relative to other importing
countries in the specified year.
```

The following is the debate structure:

```
Agent 1- agree: the image and document provide clear evidence that china
(hong kong sar) was the second-highest importer of gold, silverware, and jewelry
in 2018, with an import value of approximately 11,000 million u.s. dollars.

Agent 2: deny: the statement is incorrect. while china (hong kong sar) is the
second-highest importer, this does not match the original claim of being
the leading importing country. the data clearly shows switzerland as the
top importer, making the original statement false.
```

The verdict provided by the judge after debate:

```
Hallucinated: True
Judgement Summary: Based on the debate transcript, the fact is partially
incorrect. While China (Hong Kong SAR) is a significant importer of gold,
silverware, and jewelry in 2018, it is not the leading importing country -
Switzerland holds that position. The statement is therefore false.
Verdict: INCORRECT
Confidence: 0.8950844460263088
```

Illusion-based Image

This fact is referenced to the figure 4.

The fact to be verified is:

```
Is the top line shorter than the bottom line?
```

For this we pass the following system prompts – Role:

```
Fact Verifier
```

Instruction:

```
Verify from the image if the fact is correct or hallucinated
```

The extracted information from the image is:

```
Info extracted from image: image_0: After careful analysis and multiple
measurement passes, I can conclusively state that the two horizontal lines
in the image are exactly the same length. This is a classic example of the
Müller-Lyer illusion, where the arrowheads at the ends of the lines create a
false perception of different lengths. The top line with outward-facing
arrowheads appears longer, while the bottom line with inward-facing
arrowheads seems shorter. However, precise pixel measurements confirm that
the central horizontal segments of both lines are identical in length. This
optical illusion demonstrates how visual context and geometric elements can
significantly influence our perception, even when the objective
measurements show no difference.
```

The following is the debate structure:

```
Agent 1- agree: the statement in the question is "is the top line shorter than
the bottom line?" the information extracted from the image and document
confirms that the central horizontal segments of both lines are identical
in length, contradicting the perception of the müller-lyer illusion.
therefore, the top line is not shorter than the bottom line.

Agent 2- agree. the extracted image context clearly confirms that the
horizontal lines are identical in length, despite the visual illusion
created by the arrowhead orientation. precise pixel measurements validate
that the lines are objectively the same length.
```

The verdict provided by the judge after debate:

```
Hallucinated: True
Judgement Summary: Consensus has been reached. Both agents agree that the top
line is NOT shorter than the bottom line. The Müller-Lyer illusion creates a
visual perception of different line lengths, but precise measurements
confirm the lines are exactly the same length.
Verdict: The statement is FALSE. The lines are objectively equal in length.
Confidence: 0.6982270534792074
```