CSMAD: Hallucination Detection via Multi-Agent Debate with NLI-Verified Contradictory Statements

Swapnil Gupta
Amazon
Gurugram, India
swapgupt@amazon.com

Akshay Verma
Amazon
Gurugram, India
aksve@amazon.com

Khushi Gupta
Amazon
Gurugram, India
ggkhushi@amazon.com

Prateek Sircar
Amazon
Gurugram, India
sircarp@amazon.com

Deepak Gupta
Amazon
Gurugram, India
dgupt@amazon.com

Abstract

Large Language Models (LLMs) are prone to hallucinations, producing fluent but factually incorrect statements. Recent multi-agent debate methods improve hallucination detection by jointly improving reasoning and decision-making. However, existing approaches either collaborate which amplifies shared overconfidence, or adopt adversarial preset stances, that can inject incorrect information complicating decision making. To address this, we propose **C**ontradictory **S**tatement **M**ulti-**A**gent **D**ebate (**CSMAD**), a multi-agent framework that creates structured disagreement by generating a contradictory claim for each input claim. CSMAD asks independent agents to evaluate the claim and the contradictory claim, which encourages different lines of reasoning without assigning preset stances. When the outcome is non-discriminative; both the contradictory statements are either accepted or rejected; the agents exchange rationales and update their judgments after considering opposing evidence. A final judge then decides the truth of the original claim, using both arguments as context. To make contradictory statement generation reliable, we add a Natural Language Inference (NLI) based verifier that checks whether the generated statement actually contradicts the original claim; if it does not, the system falls back to an explicit negation-based contradiction. Across public benchmarks for question answering and scientific claim verification, as well as a proprietary e-commerce claims dataset, we show that CSMAD consistently outperforms the strongest baseline for both large (Claude-3.5 Sonnet) and medium-sized (Qwen3-8B) language models, improving F1 by +2.3 and +4.1 points, respectively, while reducing LLM token cost by 28%.

CCS Concepts

• **Applied computing** → **Online shopping**.

Keywords

Hallucination Detection, Multi-agent Debate, LLM Reasoning

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808508

ACM Reference Format:
Swapnil Gupta, Akshay Verma, Khushi Gupta, Prateek Sircar, and Deepak Gupta. 2026. CSMAD: Hallucination Detection via Multi-Agent Debate with NLI-Verified Contradictory Statements. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3805712.3808508

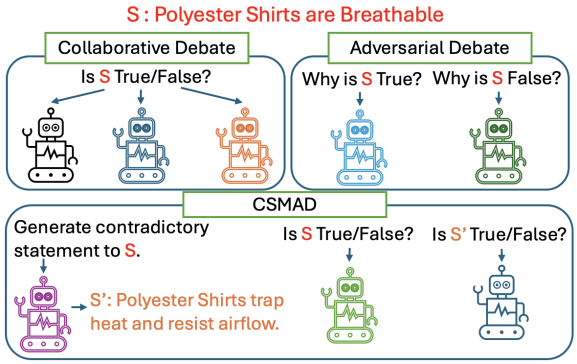

Figure 1: CSMAD vs. collaborative and adversarial multi-agent debate methods. CSMAD generates a contradictory claim for each input claim assign the pair to separate agents for evaluation.

1 Introduction

Large Language Models (LLMs) achieve strong results across many NLP tasks [2, 4, 11, 14], yet they are prone to generating fluent but factually incorrect statements [3, 24]. In e-commerce applications [8, 13, 16] even subtle inaccuracies can erode trust. For instance, describing a polyester shirt as "durable" and "wrinkle-resistant" is acceptable, but attributing "breathability" to polyester can be misleading.

Recent works leverage a model's own reasoning to assess factuality. Single-agent methods include chain-of-thought prompting, diverse sampling, and self-reflection loops [9, 12, 20, 21, 23] which aim to either deepen or broaden an LLM's reasoning. Multi-agent methods coordinate several LLM instances that interact via collaborative or adversarial debate [6, 7, 10, 15, 19]. However, single-agent methods can reinforce an LLM's initial answer, and diverse-sampling

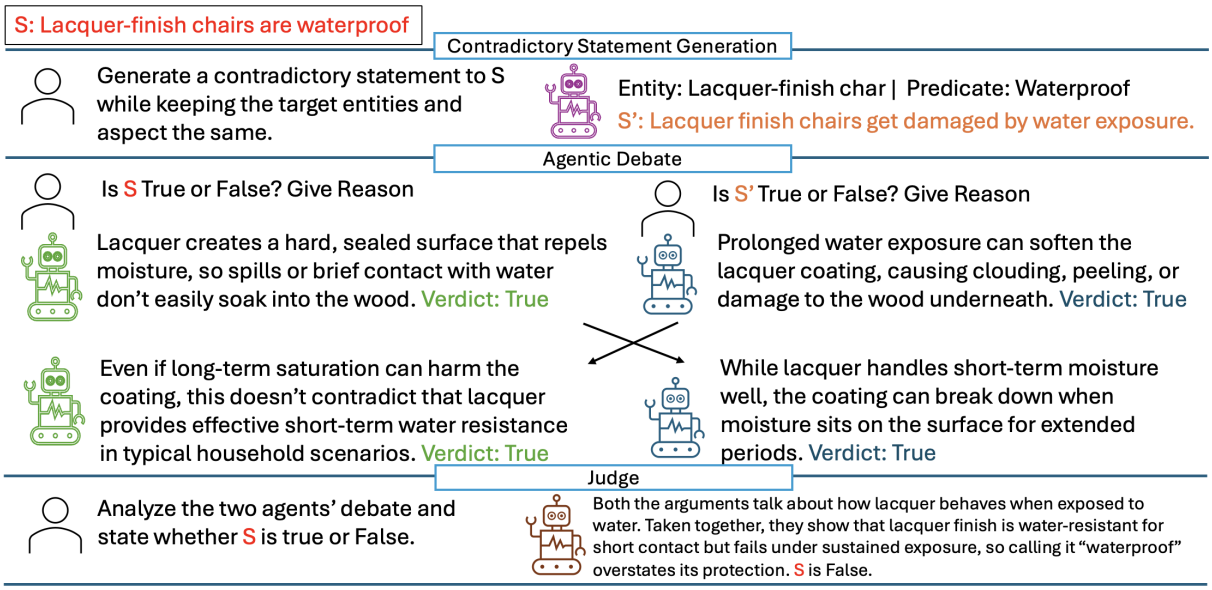

Figure 2: Illustration of the proposed CSMAD workflow. Shortened instructions are shown for clarity. See Section 2 for details.

or collaborative multi-agent setups may still converge on the same incorrect response [7]. Adversarial debates, in contrast, force preset opposing stances, which can introduce misleading arguments and complicate the final decision [19]. A key requirement is to promote divergent reasoning without forcing agents into preset roles. To achieve this, we compare the input claim against against its counter-claim. This reduces anchoring on a single narrative and encourages the LLM to search for alternative evidence and counterexamples.

We propose **CSMAD** (Contradictory Statement Multi-Agent Debate), a multi-agent framework for hallucination detection that debates an input claim against its LLM-generated contradictory claim. CSMAD initializes two agents that evaluate the claim and the contradictory claim respectively, and performs multi-round debate when the initial outcome is unclear. Unlike collaborative debates, CSMAD assigns different statements to different agents and unlike adversarial debates, no agent is forced to defend a preset stance (see Figure 1). A practical challenge is that contradictory claims generated by an LLM are not always reliable as they may be only weakly related to the input claim, or they may not truly contradict it. We therefore add a lightweight Natural Language Inference (NLI)-based verifier that checks whether the generated statement contradicts the original claim. When verification fails, the system falls back to an explicit negation-based contradictory claim, which provides a safe and unambiguous alternative. Finally, a judge aggregates the agent outputs and produces a veracity label for the original claim.

We make the following key contributions:

(1) We introduce CSMAD, a hallucination-detection framework that performs multi-agent debate using LLM-generated contradictory claims.
(2) We propose an NLI-based verification step for contradictory-claim generation, with a negation-based fallback, improving robustness to imperfect generations.
(3) We present extensive experiments on a proprietary e-commerce claims dataset (EcomClaims) and public benchmarks for claim verification and question answering tasks, comparing against strong single-agent and multi-agent baselines.

An A/B test on displaying LLM-generated educational content about product attribute values corrected via CSMAD at a popular e-commerce company achieved a 6 basis points lift in revenue.

2 Proposed Method

2.1 Notations and overview

Given a claim S, CSMAD first prompts an LLM to generate a *contradictory claim* S' (Section 2.2). CSMAD then initializes two identical agents to evaluate S and S', respectively. If the initial assessments are inconclusive i.e., both agents either accept or both reject their assigned claims, CSMAD runs a multi-round debate where the agents exchange their rationales (Section 2.3). Finally, a judge combines the two agent rationales and produces a decision for the original claim S. Figure 2 illustrates the full workflow.

2.2 Contradictory Statement Generation

Given an input claim S, we generate a contradictory claim S' by prompting an LLM. We constrain the generation to the claim context (e.g., preserving the main entities, the target attribute, and the scope). A simple option for contradiction generation is to output a direct negation, but pure negations can sound mechanical and often provide a weak counter-view for comparison. We therefore prefer a free-form, natural contradictory claim. For example, given "Chairs with cotton seats are durable," a suitable counterpart is "Chairs with cotton seats wear and tear fast." However, free-form contradictions are not always easy to generate reliably. For instance, in precise numeric claims such as "Battery capacity of iPhone 16 is 5000 mAh," contradictory statements can drift or become only loosely related. To

handle this, we use a two-step strategy with verification and fallback. In a single LLM call, we ask the model to produce two candidates: (i) a negation-based contradictory claim S_{neg}, and (ii) a free-form contradictory claim S_{free}. We then apply a lightweight off-the-shelf Natural Language Inference (NLI) classifier to check whether S_{free} contradicts S. If model outputs contradiction, we set $S' \leftarrow S_{\text{free}}$; otherwise $S' \leftarrow S_{\text{neg}}$. We used a DeBERTa-v3-large[1] based NLI model. The model is trained on a diverse mixture of large-scale NLI datasets (e.g., MNLI, FEVER-NLI, ANLI, and WANLI) spanning multiple domains, making it an effective *zero-shot* classifier.

2.3 Debate

We initialize two symmetric agents, A_S and $A_{S'}$, assigned to the claim S and S', respectively. First, each agent produces a judgment and a rationale. A_S outputs $(y_S^{(0)}, r_S^{(0)})$ and $A_{S'}$ outputs $(y_{S'}^{(0)}, r_{S'}^{(0)})$, where $y \in \{\text{TRUE}, \text{FALSE}\}$ and r is a natural-language explanation. If in the initial outcome $y_S^{(0)} == y_{S'}^{(0)}$ (both agents either accept or reject their statements), we start a debate for at most T rounds.

At round $t \in \{1, \ldots, T\}$, each agent is shown the other agent's statement, judgment, and rationale, and is asked to update its evaluation of its *own* statement. We implement this with a simple handoff in the prompt, e.g., for A_S:

> There is another agent who is evaluating the statement: $\{S'\}$. Its response is $\{y_{S'}^{(t)}\}$ for the reason: $\{r_{S'}^{(t)}\}$.
> Based on the other agent's opinion as additional information, update both your analysis and the final verdict of True or False for the input claim $\{S\}$.

The agent then returns an update

$$(y_S^{(t+1)}, r_S^{(t+1)}) = f_S\big(S, y_S^{(t)}, r_S^{(t)}, S', y_{S'}^{(t)}, r_{S'}^{(t)}\big),$$

and $A_{S'}$ updates symmetrically to $(y_{S'}^{(t+1)}, r_{S'}^{(t+1)})$. The debate stops early if the outcome becomes discriminative $(y_S^{(t+1)} \neq y_{S'}^{(t+1)})$. The final outputs $(y_S^{(\tau)}, r_S^{(\tau)})$ and $(y_{S'}^{(T)}, r_{S'}^{(T)})$ are then passed to a judge, which combines the two rationales to produce a final decision for the input claim S.

3 Experiments

We frame hallucination detection as a binary classification task where for an input statement the output is either True/False.

3.1 Datasets

We conduct experiments on two open-source benchmark dataset: SciFact [17] and BoolQ [5] and a proprietary e-commerce dataset which we have named as EcomClaims. BoolQ is a paragraph-based binary question answering dataset which instead of presenting a statement to be verified, contains questions to be answered as either true or false. To conform it in the hallucination detection task, we converted all the questions to assertive statements using Claude3.5-Sonnet [2]. For instance, a question "Is France the same timezone as the UK?" gets rewritten as "France is the same timezone as the UK." SciFact is a scientific claim verification dataset containing expert-written claims paired with evidence-containing abstracts where each claim is labeled as "SUPPORT/REFUTE". For abstract retrieval in SciFact, we adopt the same strategy as proposed in [1].

EcomClaims: We created EcomClaims, a dataset of LLM generated educational description about common e-commerce attribute values. For example, in Air Purifier category, HEPA is a common Filter Type. We split each description into granular claims using an LLM. An example claim is: "PVC paddleboards are lightweight." Three human annotators independently reviewed every claim by searching for evidence online and applying their judgment. Each claim was labeled TRUE, FALSE, or NA (if uncertain). We retained only those claims with agreement on TRUE or FALSE from at least two annotators. Following this process we created a set of 5481 manually labeled claims containing 1264 (23.1%) hallucinated claims. We used Claude3.5-Sonnet [2] for both content generation and claim segmentation.

[1] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

Table 1: Results (macro-F1, %) on BoolQ and SciFact.

Method	Claude3.5-Sonnet		Qwen3-8B	
	BoolQ	SciFact	BoolQ	SciFact
Self-reflection	75.2	68.5	63.2	53.8
Self-contrast	79.4	73.1	68	57.2
MAD-collaborative	80.5	72.7	71.3	59.4
MAD-adversarial	78.9	73.8	66.3	58.8
CFMAD	82.3	75.6	70.4	60.5
CSMAD (ours)	**84.3**	**77.6**	**75.2**	**64.7**

Table 2: Recall (R%) and Precision (P%) on EcomClaims dataset relative to Self-reflection.

Method	Claude3.5-Sonnet		Qwen3-8B	
	R%	P%	R%	P%
Self-contrast	+6.1	+1.6	+5.7	+2.1
MAD-collaborative	+6.3	+1.7	+7.1	+1.9
MAD-adversarial	+5.4	+1.3	+4.7	+1.9
CFMAD	+7.1	+2.6	+8.9	**+2.5**
CSMAD (ours)	**+12.7**	**+3.9**	**+15.3**	**+2.5**

3.2 Experimental Setup

Baselines. We experiment with an api-based LLM Claude3.5-Sonnet [2] and an open-source Qwen3-8B [22] LLM. We compare CSMAD with the following 5 baseline methods: Self-reflection [12], Self-contrast [23], MAD-collaborative [6], MAD-adversarial [10] and CFMAD [7] where we adopted the implementation settings from original papers. Following [6] we set the maximum number of debating rounds as 2 for CSMAD.

Evaluation Metrics. For **SciFact** and **BoolQ**, following prior work [18], we report *macro-F1*. For the real-world **EcomClaims** our primary focus is on detecting hallucinations and hence we report Precision (**P**) and Recall (**R**) on the FALSE class.

Table 3: Ablation on contradiction generation strategy using Claude3.5-Sonnet. We report macro-F1, (%).

Method	BoolQ	EcomClaims
CSMAD (Negation-only)	79.2	-
CSMAD (Free-form only)	82.5	+5.5
CSMAD (ours)	84.3	+8.5

4 Results

Performance comparison. We present our results on the public benchmarks BoolQ and SciFact in Table 1 and on the proprietary EcomClaims dataset in Table 2. For EcomClaims, we have shown incremental improvements to maintain anonymity as mandated by company policy. Across datasets and base LLMs, CSMAD consistently delivers the best results. We see especially large gains in recall on EcomClaims, where in absence of any reference database models rely on their parametric knowledge and reasoning. This suggests that pairing each claim with a well-formed contradictory claim imparts a broader and more effective reasoning for hallucination detection. Among baselines, Self-reflection consistently underperforms because it tends to reinforce the model's initial stance instead of introducing new lines of thought. Between the collaborative baselines Self-contrast and MAD-collaborative, the latter performs better, indicating that multi-agent debate can iteratively refine reasoning over rounds. CFMAD is the strongest baseline, underscoring the value of diverse thinking; however, MAD-adversarial often trails MAD-collaborative, showing that forcing incorrect abductions through preset stances can reduce reliability. We also observe that adversarial methods degrade more sharply when moving from Claude3.5-Sonnet to Qwen3-8B, suggesting that smaller models are more sensitive to forced counterfactual abductions.

Contradictory claim generation. We manually evaluate the quality of the LLM-generated contradictory claims on a stratified sample of N=300 from BoolQ, SciFact, and EcomClaims. Overall we observed that the free-form contradictory claim S_{free} is correct in **78.0%** of the cases, with higher correctness on qualitative attribute claims (**84.5%**) than on precise numeric claims (**63.2%**). After applying our NLI-based gate and negation fallback, the final selected contradictory claim S' is correct in **95.6%** of the cases. Across datasets, the fallback to S_{neg} rate ranges from **19–28%**, and is highest for numeric and multi-clause claims. Manual inspection shows that most rejected S_{free} candidates are either (i) *neutral* statements that do not contradict the claim, or (ii) cases with *entity/property drift*.

We next study the impact of contradiction generation on the performance of CSMAD. We compare three strategies: (1) Negation-only, which always uses S_{neg}; (2) Free-form only, which always uses S_{free} without verification; and (3) Free-form + NLI gate + negation fallback, which is our full approach. Table 3 reports the downstream hallucination detection performance. Our gated approach consistently performs best by retaining natural contradictions when reliable and falling back otherwise. Free-form contradictions consistently outperform Negations-only, demonstrating their importance in prompting divergent thinking in the multi-agent debate.

Computational Cost. Table 4 reports the average number of input and output tokens per sample on EcomClaims using Claude-3.5 Sonnet, comparing CSMAD against our baselines. Overall, CSMAD uses roughly one-third fewer tokens than other multi-agent debate methods. This reduction comes from fewer LLM operations (two fixed agents) as compared to collaborative methods and simpler debate setup than adversarial variants. In addition, CSMAD reaches an agent-level consensus without invoking the judge in 68% of samples, compared to 77% for MAD-collaborative and 52% for CFMAD, which lowers the average number of LLM calls per instance.

Table 4: Average number of input and output tokens per sample for Claude3.5-Sonnet in EcomClaims dataset.

Method	#Input tokens	#Output tokens
Self-reflection	321.3	203.8
Self-contrast	1344.8	1855.6
MAD-collaborative	1256.7	1947.2
MAD-adversarial	1520.1	2039.3
CFMAD	1219.9	1862.3
CSMAD (ours)	863.2	1337.8

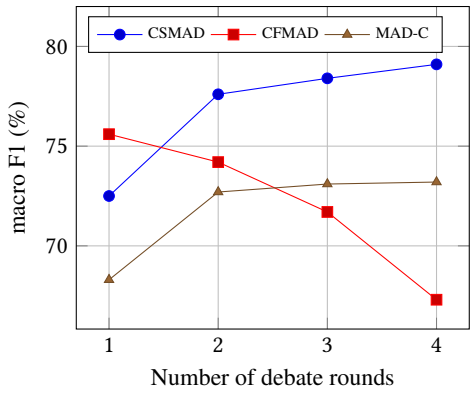

Figure 3: Comparison with changing number or debate rounds on SciFact dataset using Claude3.5-Sonnet. Here MAD-C implies MAD-collaborative.

Number of debate rounds. CFMAD reports declining performance as the number of rounds increases [7], whereas collaborative MAD benefits from longer debates [6]. We evaluate the impact of varying number of rounds on CSMAD, using SciFact with Claude3.5-Sonnet. As shown in Figure 3, CSMAD improves with more rounds, indicating a stable debate process that benefits from iterative refinement. Performance nearly saturates after two rounds by when 72% of examples reached consensus needing no further debate.

5 Conclusion

We introduced **CSMAD**, a high-efficiency multi-agent hallucination detection framework that debates a claim against a carefully constructed contradictory statement. By grounding the debate in contradictory claims, CSMAD promotes diverse reasoning without requiring preset stances. To make contradictory statement generation reliable, we propose a simple two-candidate strategy that produces a free-form contradiction alongside a negation-based fallback, and we

use an off-the-shelf NLI verifier to select the free-form contradiction when it truly contradicts the input claim. We performed experiments on BoolQ, SciFact, and our proprietary EcomClaims dataset, and demonstrated that CSMAD is effective for hallucination detection, consistently improving over strong baselines across both large and medium-sized LLMs while also reducing token cost.

References

[1] Carlos Alvarez, Maxwell Bennett, and Lucy Wang. 2024. Zero-shot Scientific Claim Verification Using LLMs and Citation Text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, Tirthankar Ghosal, Amanpreet Singh, Anita Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 269–276. https://aclanthology.org/2024.sdp-1.25/

[2] Anthropic. 2024. Claude 3.5 Sonnet. (2024). https://www.anthropic.com/news/claude-3-5-sonnet

[3] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 [cs.CL] https://arxiv.org/abs/2302.03494

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[5] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 2924–2936. doi:10.18653/v1/N19-1300

[6] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs.CL] https://arxiv.org/abs/2305.14325

[7] Yi Fang, Moxin Li, Wenjie Wang, Lin Hui, and Fuli Feng. 2025. Counterfactual Debating with Preset Stances for Hallucination Elimination of LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 10554–10568. https://aclanthology.org/2025.coling-main.703/

[8] Swapnil Gupta, Lucas Pereira Carlini, Prateek Sircar, and Deepak Gupta. 2025. Break-Ideate-Generate (BrIdGe): Moving beyond Translations for Localization using LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 627–637. doi:10.18653/v1/2025.naacl-industry.51

[9] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1827–1843. doi:10.18653/v1/2023.findings-emnlp.123

[10] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17889–17904. doi:10.18653/v1/2024.emnlp-main.992

[11] 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. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[12] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366 [cs.AI] https://arxiv.org/abs/2303.11366

[13] Prateek Sircar, Aniket Chakrabarti, Deepak Gupta, and Anirban Majumdar. 2022. Distantly Supervised Aspect Clustering And Naming For E-Commerce Reviews. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min (Eds.). Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 94–102. doi:10.18653/v1/2022.naacl-industry.12

[14] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[15] Akshay Verma, Swapnil Gupta, Deepak Gupta, Prateek Sircar, and Siddharth Pillai. 2026. SELENE: Selective and Evidence-Weighted LLM Debating for Efficient and Reliable Reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, Yevgen Matusevych, Gülşen Eryiğit, and Nikolaos Aletras (Eds.). Association for Computational Linguistics, Rabat, Morocco, 95–104. doi:10.18653/v1/2026.eacl-industry.7

[16] Akshay Verma, Swapnil Gupta, Siddharth Pillai, Prateek Sircar, and Deepak Gupta. 2026. ReflectiveRAG: Rethinking Adaptivity in Retrieval-Augmented Generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, Yevgen Matusevych, Gülşen Eryiğit, and Nikolaos Aletras (Eds.). Association for Computational Linguistics, Rabat, Morocco, 377–384. doi:10.18653/v1/2026.eacl-industry.27

[17] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7534–7550. doi:10.18653/v1/2020.emnlp-main.609

[18] Haoran Wang and Kai Shu. 2023. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6288–6304. doi:10.18653/v1/2023.findings-emnlp.416

[19] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6106–6131. doi:10.18653/v1/2024.acl-long.331

[20] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=1PL1NIMMrw

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

[22] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[23] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3602–3622. doi:10.18653/v1/2024.acl-long.197

[24] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219 [cs.CL] https://arxiv.org/abs/2309.01219

Authors Bio

Swapnil Gupta is an Applied Scientist at Amazon
Akshay Verma is a Senior Applied Scientist at Amazon
Khushi Gupta is an Applied Scientist at Amazon
Prateek Sircar is a Principal Applied Scientist at Amazon
Deepak Gupta is a Senior Applied Science Manager at Amazon