Confidence-Aware Multi-Agent Orchestration for Evaluating Multimodal Rule Compliance

Jie Zhou
Amazon
Seattle, WA, USA
jiezhoua@amazon.com

Zhenyu Zhang
Amazon
Seattle, WA, USA
zhenyuzh@amazon.com

Xiaolong Kuang
Amazon
Seattle, WA, USA
xkkuang@amazon.com

Xinyang Shen
Amazon
Seattle, WA, USA
xinyans@amazon.com

ABSTRACT

Evaluating rule compliance in industry requires assessing products against complex regulatory standards using multimodal data sources—a task where both correctness and trustworthiness of automated judgments are critical. Existing approaches either rely on costly human audits, supervised classifiers that demand large-scale labeled training data, or monolithic multimodal models that apply uniform reasoning without mechanisms for quantifying uncertainty or resolving cross-modal conflicts. These limitations raise fundamental questions about how to evaluate and ensure trustworthiness in agentic systems operating over heterogeneous, noisy data.

We propose a confidence-aware multi-agent orchestration framework that addresses these evaluation challenges through three mechanisms: (1) a Rule Generation module that produces interpretable, auditable compliance specifications; (2) modality-specialized Base Agents with a perception-then-judgment protocol that separates observation from evaluation, enabling fine-grained assessment of evidence faithfulness; and (3) an Orchestrator Agent that performs confidence-aware triage and generates targeted follow-up queries to resolve inter-agent conflicts, providing an explicit audit trail of its reasoning process. Our framework demonstrates how multi-agent orchestration can be designed for trustworthiness by construction—through confidence-aware decision routing, evidence-grounded adjudication, and iterative uncertainty reduction. On a product industry guideline compliance task with 216 human-validated samples, the framework achieves 89.8% accuracy and 92.5% F1 score, outperforming the strongest baseline (77.3% accuracy, 83.5% F1) by 12.5 and 9.0 absolute points respectively. Detailed ablation analysis reveals that cross-modal conflict resolution contributes 13.9% absolute F1 improvement over naive multi-modal fusion, providing quantitative evidence for the value of confidence-aware orchestration in trustworthy agentic evaluation.

KDD'26, August 2026, Jeju, Korea

ACM ISBN 978-1-4503-XXXX-X/2018/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

agentic AI evaluation, multi-agent trustworthiness, cross-modal conflict resolution, confidence calibration, human-in-the-loop evaluation

ACM Reference Format:
Jie Zhou, Zhenyu Zhang, Xiaolong Kuang, and Xinyang Shen. 2026. Confidence-Aware Multi-Agent Orchestration for Evaluating Multimodal Rule Compliance. In *Proceedings of KDD Workshop on Evaluation and Trustworthiness of Agentic AI (KDD'26).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As agentic AI systems are increasingly used for autonomous decision-making, evaluating their trustworthiness becomes paramount—particularly when decisions involve multi-step reasoning over heterogeneous, noisy data sources [5, 20]. Rule-based compliance checking in industry exemplifies this challenge: large numbers of products must be evaluated against complex regulatory standards using multimodal evidence (text, images, video), where each modality is subject to noise, missing data, and partial observations [16]. The evaluation must not only be accurate but also *explainable*, *auditable*, and *calibrated in its uncertainty*—core requirements of trustworthy agentic AI.

Traditional compliance evaluation relies on supervised classifiers trained to identify rule violations [9, 19], requiring large labeled datasets and retraining when standards evolve. The LLM-as-a-Judge paradigm offers a scalable alternative for complex textual reasoning [20], and multimodal LLMs (MLLMs) extend this paradigm to diverse data types [5]. However, adopting these systems raises critical trustworthiness concerns. Existing MLLM-based evaluation methods apply a single generalized reasoning strategy across all modalities, overlooking modality-specific characteristics and providing limited mechanisms for uncertainty quantification or conflict resolution [1]. Multi-agent collaboration addresses task decomposition through specialized agents [7, 12], yet simple ensemble strategies (averaging, voting) treat all agent outputs as equally reliable regardless of modality-dependent signal strength, obscuring the reasoning pathway and undermining auditability.

A fundamental gap exists in how multi-agent systems handle *inter-agent disagreements*—a critical trustworthiness dimension that

determines whether the system can identify and resolve conflicting evidence rather than masking uncertainty through naive aggregation. When modality-specialized agents produce conflicting verdicts, the conflict itself carries rich diagnostic signal: it indicates regions of epistemic uncertainty where the system should invest additional reasoning before committing to a decision. Few existing works address inter-agent conflicts directly, and achieving coherent, trustworthy consensus among modality-specialized agents remains largely underexplored.

To address these limitations, we propose a confidence-aware multi-agent orchestration framework that makes trustworthiness a first-class design objective. Our framework coordinates modality-specialized agents through an intelligent orchestrator that: (i) performs confidence-aware triage to synthesize cross-modal outputs while accounting for modality-dependent reliability; (ii) generates targeted follow-up queries when disagreements or ambiguity arise, actively gathering complementary evidence to resolve conflicts; and (iii) maintains an explicit reasoning trace throughout the process, enabling post-hoc audit of every decision. Through iterative evidence-gathering, the system progressively reduces epistemic uncertainty, refining verdicts until consensus is achieved or termination criteria are met.

The main contributions of this paper are:

- We propose a multi-agent orchestration framework designed for trustworthy compliance evaluation, where confidence-aware decision routing, evidence-grounded adjudication, and iterative uncertainty reduction enable auditable and reliable autonomous judgments over noisy multimodal data.
- We introduce an active conflict resolution mechanism where the orchestrator generates targeted follow-up queries to elicit additional evidence from modality-specialized agents, rather than relying on naive ensemble strategies. To our knowledge, this is the first work to employ active information gathering for resolving cross-modal agent conflicts in rule-based evaluation.
- We present a systematic evaluation on 216 human-validated samples demonstrating that confidence-aware orchestration achieves 92.5% F1, and provide ablation analysis quantifying the contribution of each trustworthiness mechanism, showing that conflict resolution alone contributes 13.9% absolute F1 improvement over naive fusion.

2 RELATED WORK

2.1 LLM-as-Judge and Agentic Evaluation

The LLM-as-a-Judge paradigm leverages zero-shot reasoning for scalable evaluation without task-specific training [20], with recent applications in regulatory compliance [3, 13]. In multimodal settings, MLLM-as-a-Judge benchmarks [5, 15] reveal that while MLLMs align with human judgments in pairwise comparisons, they exhibit biases, overconfidence, and inconsistent scoring [18]—raising concerns about trustworthiness when these models serve as autonomous evaluators. Traditional compliance evaluation trains supervised classifiers to detect rule violations [2], requiring substantial labeled data and retraining when standards evolve.

2.2 Multi-Agent Systems and Trustworthiness

Multi-agent frameworks decompose tasks across modality-specialized agents [12], yet typically aggregate outputs through static strategies such as majority voting without mechanisms for conflict resolution or uncertainty quantification. Multi-agent debate (MAD) has been proposed to resolve disagreements through iterative debate [8, 11], but recent studies show that vanilla debate often underperforms simple ensembles [21], agents can be swayed by confident but incorrect reasoning [17], and homogeneous deliberation may degrade performance [4]. Adaptive approaches that selectively escalate hard cases [6] or terminate debate via stability detection [10] demonstrate that targeted intervention outperforms exhaustive debate.

Our work builds on these insights but addresses a distinct gap: *cross-modal conflict resolution as a trustworthiness mechanism*. Rather than engaging in open-ended debate, the orchestrator generates directed follow-up queries to elicit complementary evidence from modality-specialized agents, combining the efficiency of selective escalation with the reliability of evidence-grounded adjudication. This design yields both better accuracy and a complete reasoning audit trail, enabling post-hoc verification of every conflict resolution decision.

3 METHODS

3.1 Framework Overview

Figure 1 presents the overall architecture of the proposed framework. Given multimodal product data—text, images, video, and audio—the system evaluates compliance against rules specified in regulatory documents. The framework comprises three components designed to maximize both accuracy and trustworthiness:

- **Rule Generation:** Extracts structured, auditable compliance rules from regulatory documents, providing interpretable evaluation criteria.
- **Base Agents:** Modality-specialized agents that independently reason over their respective data types using a perception-then-judgment protocol, reporting evidence with confidence levels.
- **Orchestrator Agent:** Reconciles cross-modal outputs through confidence-aware triage, generating targeted follow-up queries when it detects uncertainty, ambiguity, or conflicts, and maintaining a complete reasoning trace for auditability.

3.2 Rule Generation

The Rule Generation module transforms regulatory documents into structured compliance rules through three stages: extraction, refinement, and routing.

Rule Extraction. An LLM extracts structured rules from guideline documents, each comprising a natural language description, severity level, source provenance, and a set of product attributes required for verification. Each rule retains traceability to its source regulation, enabling auditability.

Rule Refinement. Raw extracted rules often contain ambiguous descriptions or domain-specific terminology that leads to uncertain agent judgments. We refine rule representations using a small set

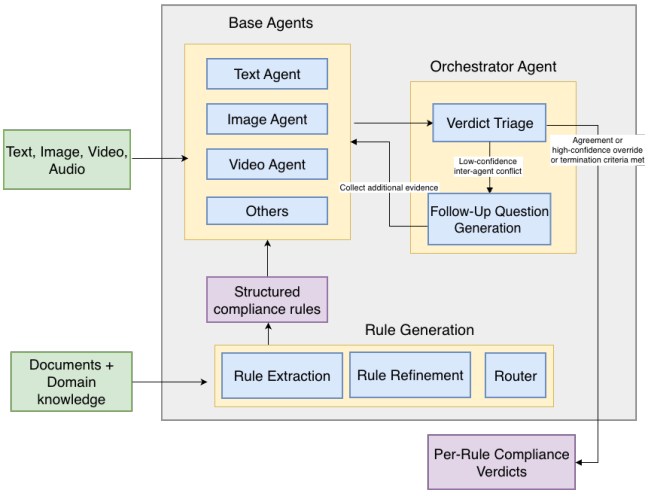

Figure 1: Framework architecture. The orchestrator synthesizes base agent outputs through confidence-aware triage and iterative follow-up queries, maintaining an audit trail of all conflict resolution decisions.

of labeled samples (5 positive and 5 negative per ambiguous rule). We analyze systematic error patterns in base agent outputs, then reformulate rule descriptions to include key attributes, explicit definitions, concrete positive and negative exemplars, and boundary conditions for ambiguous cases. Critically, this refinement modifies the rule specification rather than model weights, enabling rapid adaptation to regulatory changes without retraining.

Rule Router. Each rule is automatically assigned to base agents based on the attributes it requires and prior domain knowledge of modality dominance. Rules requiring structured metadata are routed to the text agent; those requiring visual inspection are routed to the vision agent; and those with ambiguous modality dependence are assigned to multiple agents for cross-modal validation. This deterministic routing enables efficient parallel execution while ensuring each rule is evaluated by the most appropriate expert.

3.3 Base Agents

The base agents function as an expert committee, with each agent contributing specialized insights. Rather than relying on a single MLLM that applies uniform reasoning across heterogeneous data, we assign dedicated foundation models to each modality, enabling reasoning strategies tailored to each data source.

The text agent evaluates compliance through logical reasoning over structured product metadata. For vision-based agents, motivated by findings that separating feature extraction from reasoning reduces hallucinations in visual understanding [14], we employ a two-step *perception-then-judgment* protocol:

(1) **Perception step:** The agent observes and describes relevant attributes from its input modality, producing factual observations without any compliance judgment.
(2) **Judgment step:** A separate reasoning process evaluates compliance based solely on the perception descriptions.

This separation is critical for trustworthiness: when observation and judgement are conflated, the model's expectation of the rule outcome can bias what it perceives. By enforcing a boundary between perception and conclusion, the protocol ensures that evidence remains grounded in actual observations rather than shaped by the desired judgment, yielding more faithful attribute extraction and enabling independent verification of the evidence chain.

Each base agent returns standardized output: a verdict (PASS/FAIL/UNCERTAIN), supporting evidence, a confidence level, and reasoning trace. This consistent format enables both uniform downstream processing and fine-grained evaluation of each agent's trustworthiness dimensions.

3.4 Orchestrator Agent

The Orchestrator Agent governs cross-modal evidence synthesis and conflict resolution through two mechanisms: verdict triage and follow-up question generation. Together, these progressively reduce ambiguity while maintaining a complete audit trail of all reasoning decisions.

Verdict Triage. Upon receiving verdicts from base agents, the orchestrator performs confidence-aware triage that integrates prior knowledge of modality dominance with each agent's output confidence:

- *Single agent triggered:* Verdict is returned directly with the agent's confidence assessment.
- *All agents agree:* Outputs are synthesized with aggregated confidence.
- *Agents conflict, one has high confidence in its dominant modality:* The orchestrator trusts the high-confidence agent operating in the modality known to carry dominant signals for the attribute being checked.
- *Agents conflict with low/uncertain confidence:* The case is escalated for iterative conflict resolution.

This mechanism avoids the pitfalls of majority voting, which treats all outputs as equally reliable regardless of modality-dependent signal strength. Importantly, the triage logic is transparent—the orchestrator records which branch was taken and why, enabling post-hoc audit of every decision.

Follow-Up Question Generation. When conflicts remain unresolved, the orchestrator initiates an active information-gathering process. Rather than re-querying agents with identical prompts or engaging in open-ended debate, it: (1) examines the conflicting verdicts, confidence levels, and supporting evidence; (2) uses an LLM to generate a focused clarification question targeting the specific point of disagreement; (3) determines which base agent is best positioned to resolve the ambiguity based on the nature of the disagreement; and (4) delegates that agent to re-evaluate the product data through the lens of the new question.

This process iterates until consensus is reached, the required confidence threshold is met, or a maximum iteration count is reached. Each iteration produces an updated verdict with revised confidence and additional evidence, which is fed back to the triage stage. Critically, this mechanism makes the system's uncertainty handling *visible*—each unresolved conflict, follow-up question, and resolution step is logged, providing a complete audit trail of how the system arrived at its final determination.

4 EXPERIMENTAL SETUP

4.1 Task and Data

We evaluate the framework on a product packaging compliance task. The goal is to determine whether products meet packaging eligibility criteria based on complex rules spanning package shape, structural integrity, and physical attributes. This task is representative of agentic evaluation challenges because: (1) it requires multi-step reasoning over heterogeneous evidence; (2) ground truth is expensive to obtain; (3) modalities provide complementary but potentially conflicting signals; and (4) decisions must be explainable to human operators.

We collect multimodal product data comprising textual catalog attributes and package images. Both modalities exhibit noise characteristics typical of real-world data: textual attributes may be incomplete or missing entirely, while package images often fail to capture all product faces and vary in visual quality due to differences in lighting and capture conditions. These data quality challenges make the task a rigorous testbed for evaluating how agentic systems handle uncertainty and incomplete information.

4.2 Evaluation Protocol

We curate a dataset of 216 products (100 eligible, 116 ineligible), with ground-truth labels validated by human domain experts. Performance is measured by accuracy, precision, recall, and F1 score.

The Rule Generation module processes compliance documents and extracts 38 structured rules across product categories. Each rule includes an ID, name, description, target attributes, and severity level ("must comply" vs. "recommended to comply"—failure on a must-comply rule renders the product ineligible). The router assigns each rule to appropriate base agents; when assignment is ambiguous, all agents are invoked.

4.3 Baselines

We compare against four baselines that represent different evaluation paradigms:

- **Text-only agent:** Evaluates compliance using only structured textual metadata, representing unimodal text-based evaluation.
- **Image-only agent:** Evaluates compliance using only product images with the perception-then-judgment protocol, representing unimodal visual evaluation.
- **Multimodal single agent (MLLM):** A monolithic multimodal LLM that processes both text and images simultaneously in a single inference pass, representing the standard MLLM-as-Judge approach.
- **Multi-agent + LLM fusion:** Employs the common "collect-then-reason" strategy where individual agent outputs are concatenated and a fusion LLM synthesizes a final decision, representing naive multi-agent aggregation without conflict resolution.

5 RESULTS AND ANALYSIS

5.1 Analysis of Trustworthiness Dimensions

Table 1 presents the performance comparison across all methods on the 216 human-validated samples.

Table 1: Performance comparison on 216 human-validated samples. Best results in bold.

Method	Acc.	Prec.	Rec.	F1
Text-only agent	0.741	0.760	0.897	0.823
Image-only agent	0.773	0.816	0.855	0.835
Multimodal single (MLLM)	0.611	0.796	0.566	0.661
Multi-agent + LLM fusion	0.759	0.812	0.834	0.823
Ours (full framework)	**0.898**	**0.913**	**0.938**	**0.925**

Cross-modal conflict resolution is critical. The most striking finding is that the multimodal single agent (MLLM) performs *worst* among all methods (F1 = 0.661) despite having access to both modalities. This demonstrates that naively combining multimodal evidence in a single inference pass introduces attention dilution—the model struggles to weigh conflicting signals without explicit reasoning structure. This result challenges the assumption that more data always yields better evaluation and highlights the trustworthiness risks of monolithic approaches that lack mechanisms for conflict identification and resolution.

Naive fusion does not capture cross-modal synergy. The multi-agent + LLM fusion baseline (F1 = 0.823) performs comparably to individual agents rather than exceeding them. Without the ability to gather new evidence or apply confidence-aware resolution logic, the fusion LLM often defaults to the more conservative verdict. This confirms that simply reasoning over existing agent outputs—the dominant paradigm in multi-agent systems—fails to exploit complementary modality information when agents disagree.

Confidence-aware orchestration resolves conflicts effectively. Our framework achieves F1 = 0.925, representing a 13.9% absolute improvement over the best non-orchestrated multi-agent approach (LLM fusion at 0.823) and a 26.4% improvement over the monolithic MLLM. The improvement is particularly pronounced in recall (0.938 vs. 0.834 for LLM fusion), indicating that the follow-up question mechanism successfully resolves cases where initial evidence was insufficient to confirm eligibility—precisely the cases where trustworthiness of the decision is most at risk.

Evidence of complementary modality contribution. Both unimodal agents achieve comparable F1 scores (text: 0.823, image: 0.835), suggesting that neither modality dominates overall. However, they exhibit different error profiles: the text agent achieves higher recall (0.897) but lower precision (0.760), while the image agent shows the opposite pattern (recall 0.855, precision 0.816). This complementarity is exactly what the orchestrator exploits through confidence-aware triage—delegating to the modality that is most reliable for each specific attribute being evaluated.

5.2 Qualitative Analysis: Conflict Resolution in Action

We illustrate the framework's trustworthiness mechanisms through a representative example. For a closure-seal rule check, both agents are triggered:

(1) *Text agent:* Returns UNCERTAIN (low confidence) because key attributes (seal type, closure mechanism) are missing from the structured metadata.
(2) *Image agent:* Returns PASS (medium confidence), observing a tongue-tab locking feature with no gaps at the top seam.

The orchestrator detects a conflict with insufficient confidence and generates a targeted follow-up question: "Are all open edges sealed or closed, not just the top? Please inspect all sides for gaps, unsealed flaps, or openings that could allow product exposure."

In round 2, the image agent returns UNCERTAIN, reasoning that it cannot confirm all edges are sealed since the bottom panel and front face are not visible in the available images. The final verdict for this rule is therefore UNCERTAIN—correctly reflecting the system's epistemic limitation rather than forcing an unreliable determination.

This example illustrates three key trustworthiness properties: (1) the system does not mask uncertainty behind a forced binary decision; (2) the follow-up question is targeted to the specific point of disagreement; and (3) the full reasoning trace is preserved for human audit.

5.3 Limitations and Failure Modes

Our analysis reveals several categories where the framework's confidence-aware mechanisms reach their limits:

Inherent data insufficiency. When critical physical attributes cannot be observed from any available modality (e.g., package weight, material thickness), the system correctly returns UNCERTAIN but cannot progress toward resolution regardless of iteration count. In our evaluation, approximately 10% of rule checks terminate as UNCERTAIN due to fundamental observability limitations.

Confidence calibration limitations. The current framework relies on LLM self-reported confidence, which may be poorly calibrated [5]. We observe cases where agents report high confidence on incorrect verdicts, bypassing the conflict resolution mechanism. Integrating learned calibration layers trained on historical agent outputs represents an important direction for improving trustworthiness.

Single-iteration resolution. Our current implementation performs at most one follow-up round. While this prevents unbounded computation, some complex cases involving interactions between multiple rules may benefit from additional iterations—a tradeoff between computational cost and evaluation thoroughness that merits further study.

6 DISCUSSION: IMPLICATIONS FOR AGENTIC AI EVALUATION

Our findings carry broader implications for the evaluation and trustworthiness of agentic AI systems:

Conflict patterns as trustworthiness signals. The orchestrator's confidence-aware triage reveals that inter-agent conflict patterns carry rich diagnostic signal about system reliability. The rate and nature of conflicts—which modalities disagree, on which rule types, and at what confidence levels—provide continuous monitoring signals that could detect distribution shift or capability degradation in agentic systems without requiring ground-truth labels.

Active uncertainty reduction vs. passive aggregation. Our results demonstrate that active information gathering (generating targeted follow-up queries) substantially outperforms passive aggregation strategies (voting, fusion) for trustworthy evaluation. This finding suggests that agentic evaluation systems should be designed with explicit mechanisms for *recognizing* what they don't know and *acting* to reduce that uncertainty, rather than forcing a determination from insufficient evidence.

Auditability through structured orchestration. The framework's layered architecture—where each decision point (triage branch, follow-up question, final verdict) is logged with its reasoning—provides a template for auditable agentic systems. Unlike end-to-end neural systems where the reasoning pathway is opaque, our orchestration structure makes each step independently verifiable, a critical requirement for regulatory compliance applications.

Scalable ground-truth construction. The agreement-based labeling strategy—where agent-agent consensus identifies easy cases while human evaluation is focused on genuinely ambiguous instances—offers a practical paradigm for evaluating agentic systems at scale without exhaustive manual annotation.

7 CONCLUSION

We presented a confidence-aware multi-agent orchestration framework that addresses key challenges in evaluating and ensuring trustworthiness of agentic AI systems operating over multimodal data. The framework introduces three mechanisms designed for trustworthiness: (1) perception-then-judgment separation that enables independent verification of evidence faithfulness; (2) confidence-aware triage that accounts for modality-dependent reliability rather than treating all agent outputs as equally trustworthy; and (3) active conflict resolution through targeted follow-up queries that makes the system's uncertainty handling visible and auditable.

Experiments on 216 human-validated samples demonstrate that the framework achieves 92.5% F1, substantially outperforming both monolithic multimodal evaluation (66.1% F1) and naive multi-agent fusion (82.3% F1). The 13.9% absolute F1 improvement from confidence-aware orchestration over naive fusion provides quantitative evidence that trustworthiness mechanisms—conflict detection, confidence-aware routing, and active evidence gathering—are not merely desirable properties but directly improve evaluation accuracy.

These results suggest that the path toward trustworthy agentic evaluation lies not in building more powerful individual models, but in designing orchestration architectures that make uncertainty explicit, conflicts resolvable, and reasoning auditable. Future work includes learned confidence calibration to replace self-reported uncertainty, multi-round iterative resolution for complex multi-rule interactions, and formalization of conflict-pattern monitoring as a continuous trustworthiness signal for agentic systems.

REFERENCES

[1] Tejas Anvekar, Krishna Singh Rajput, Chitta Baral, and Vivek Gupta. 2025. Rethinking Information Synthesis in Multimodal Question Answering A Multi-Agent Perspective. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (Eds.). The Asian Federation of

Natural Language Processing and The Association for Computational Linguistics, Mumbai, India, 3674–3686. https://doi.org/10.18653/v1/2025.ijcnlp-long.192

[2] Muhammad Azeem and Sallam Abualhaija. 2024. A Multi-solution Study on GDPR AI-enabled Completeness Checking of DPAs. *Empirical Software Engineering* 29 (06 2024). https://doi.org/10.1007/s10664-024-10491-3

[3] Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (March 2024), 2129–2145. https://doi.org/10.18653/v1/2024.eacl-long.130

[4] Blaz Bertalanic and Carolina Fortuna. 2026. The Cost of Consensus: Isolated Self-Correction Prevails Over Unguided Homogeneous Multi-Agent Debate. *arxiv* (04 2026). https://doi.org/10.48550/arXiv.2605.00914

[5] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark. *Proceedings of the 41st International Conference on Machine Learning* (2024), 34.

[6] Justin Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2025. MAgICoRe: Multi-Agent, Iterative, Coarse-to-Fine Refinement for Reasoning. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (11 2025), 32663–32686. https://doi.org/10.18653/v1/2025.emnlp-main.1660

[7] Anoop Cherian, River Doyle, Eyal Ben-Dov, Suhas Lohit, and Kuan-Chuan Peng. 2025. WISE: Weighted Iterative Society-of-Experts for Robust Multimodal Multi-Agent Debate. *arxiv* (12 2025).

[8] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning* 467 (2024), 31.

[9] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2020. Scalable Detection of Offensive and Non-compliant Content / Logo in Product Images. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020), 2236–2245.

[10] Tianyu Hu, Zhen Tan, Song Wang, Huaizhi Qu, and Tianlong Chen. 2026. Multi-Agent Debate for LLM Judges with Adaptive Stability Detection. *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2026). https://openreview.net/forum?id=Vusd1Hw2D9

[11] 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. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (11 2024), 17889–17904. https://doi.org/10.18653/v1/2024.emnlp-main.992

[12] Huawei Lin, Yunzhi Shi, Tong Geng, Weijie Zhao, Wei Wang, and Ravender Singh. 2025. Agent-Omni: Test-Time Multimodal Reasoning via Model Coordination for Understanding Anything. *arxiv* (11 2025).

[13] Bill Marino, Rosco Hunter, Zubair Jamali, Marinos Kalpakos, Mudra Kashyap, Isaiah Hinton, Alexa Hanson, Maahum Nazir, Christoph Schnabl, Felix Steffek, Hongkai Wen, and Nicholas Lane. 2025. AIReg-Bench: Benchmarking Language Models That Assess AI Regulation Compliance. (10 2025). https://doi.org/10.48550/arXiv.2510.01474

[14] Bowei Pu, Chuanbin Liu, Yifan Ge, Peicheng Zhou, Yiwei Sun, Zhiying Lu, Zhangchi Hu, and Hongtao Xie. 2026. Decoupling Perception from Reasoning for Hallucination-Resistant Video Understanding. *arXiv* (2026).

[15] Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. 2025. Judge Anything: MLLM as a Judge Across Any Modality. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (2025). https://api.semanticscholar.org/CorpusID:277271795

[16] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2026. Deep Multimodal Learning with Missing Modality: A Survey. *Transactions on Machine Learning Research* (2026). https://openreview.net/forum?id=tc7RFcx4hT

[17] Andrea Wynn, Harsh Satija, and Gillian K Hadfield. 2026. Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate. *ICML* abs/2509.05396 (2026). https://api.semanticscholar.org/CorpusID:281203300

[18] Ancheng Xu, Zhihao Yang, Jingpeng Li, Guanghu Yuan, Longze Chen, Liang Yan, Jiehui Zhou, Zhen Qin, Hengyun Chang, Hamid Alinejad-Rokny, Bo Zheng, and Min Yang. 2025. EVADE: Multimodal Benchmark for Evasive Content Detection in E-Commerce Applications. *arxiv* (05 2025). https://doi.org/10.48550/arXiv.2505.17654

[19] Kanwal Yousaf and Tabassam Nawaz. 2022. A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos. *IEEE Access* 10 (2022), 16283–16298. https://api.semanticscholar.org/CorpusID:246425347

[20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Curran Associates Inc.* (2023), 29.

[21] Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2026. Demystifying Multi-Agent Debate: The Role of Confidence and Diversity. *ArXiv* abs/2601.19921 (2026). https://api.semanticscholar.org/CorpusID:285102089