INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Refusal-Aware Instruction Tuning for Hallucination Mitigation near Uncertain Knowledge Boundaries

Zoé Mathieu, Adrien T. Roy

Department of Mathématiques Informatique et Télécommunications, Université Toulouse III, Paul Sabatier, INSERM, UMR 1027, F-31000 Toulouse, France

Abstract

The rapid deployment of large language models has exposed significant vulnerabilities in their reliability, particularly concerning their propensity to hallucinate when confronted with queries that exceed their internalized knowledge capacities. This paper investigates the critical challenge of mitigating hallucinations near uncertain knowledge boundaries through a novel paradigm termed refusal-aware instruction tuning. Conventional instruction tuning often inadvertently encourages models to generate responses even when uncertain, prioritizing fluency and helpfulness over factual accuracy. To counteract this phenomenon, we propose a systematic framework that integrates uncertainty quantification directly into the instruction tuning pipeline, teaching models to explicitly refuse to answer when a query falls outside their reliable knowledge distribution. By formalizing the concept of uncertain knowledge boundaries, we develop a specialized dataset curation strategy that identifies low-confidence regions and pairs them with appropriate refusal responses. Our methodology is subsequently evaluated through extensive empirical analysis, demonstrating that refusal-aware instruction tuning substantially reduces hallucination rates without compromising the utility of the model on well-known facts. The findings suggest that aligning models with their own epistemic limitations is a fundamental requirement for deploying trustworthy artificial intelligence systems in high-stakes domains.

Keywords: *Instruction Tuning, Hallucination Mitigation, Knowledge Boundaries, Epistemic Uncertainty.*

1. INTRODUCTION

1.1 Context and Motivation

The emergence of instruction-tuned large language models has fundamentally transformed the landscape of natural language processing, enabling unprecedented capabilities in zero-shot generalization and human-computer interaction [1]. These models are trained to follow complex human instructions across diverse domains, achieving remarkable fluency and coherence [2]. However, a persistent and critical limitation in the widespread adoption of these models is their tendency to generate factually incorrect or nonsensical information, a phenomenon widely

documented as hallucination [3]. Hallucinations occur when a model produces outputs that are not grounded in its pre-training data or contradict established factual knowledge [4]. This issue is particularly acute when models are prompted with queries that reside on the periphery of their internalized knowledge, where their confidence should theoretically be low, but their instruction-following training compels them to provide a definitive answer anyway. The standard paradigm of instruction tuning predominantly focuses on maximizing the likelihood of a human-preferred response, often utilizing datasets constructed from helpful and compliant human-machine interactions [5]. In this conventional setup, models are rarely penalized for generating plausible but incorrect answers, leading to an overconfidence bias [6]. When confronted with tasks that require specialized knowledge or recall of obscure facts, models that have undergone standard instruction tuning typically attempt to extrapolate or fabricate details to fulfill the user prompt, rather than acknowledging their limitations [7]. This behavior indicates a profound misalignment between the internal knowledge capacity of the model and its external generation policy. Recognizing the boundaries of one's own knowledge is a fundamental aspect of intelligent behavior, yet current training methodologies largely fail to instill this capability in artificial systems.

1.2 Problem Statement

Addressing the hallucination problem requires a shift in how models are trained to handle uncertainty. The core issue lies in the ambiguous zone between what a model knows with absolute certainty and what it completely lacks knowledge of, which we refer to as the uncertain knowledge boundary [8]. Queries that fall into this boundary are the most likely to elicit hallucinations, as the model possesses partial, fragmented, or conflicting information [9]. To mitigate this, models must be endowed with a refusal mechanism, allowing them to politely decline to answer or state their lack of knowledge when operating within these uncertain regions. This paper introduces the concept of refusal-aware instruction tuning, a methodology designed to explicitly teach models to identify their knowledge boundaries and output refusal responses when appropriate. We hypothesize that by integrating refusal behaviors into the fine-tuning process, aligned with the intrinsic uncertainty of the model, we can significantly curtail hallucinations [10]. The subsequent sections will detail the theoretical underpinnings of knowledge boundaries, propose a scalable method for generating refusal-aware training data, and present a modified optimization objective that balances helpfulness with factual prudence [11]. The primary goal is to establish a rigorous framework for developing language models that are not only capable but reliably aware of their own epistemic limitations [12].

2. Background and Literature Review

2.1 Evolution of Instruction Tuning

Instruction tuning has evolved rapidly from simple supervised fine-tuning on task-specific datasets to complex alignment procedures involving reinforcement learning from human feedback [13]. Early approaches demonstrated that fine-tuning models on a diverse set of phrased instructions could significantly improve their ability to generalize to unseen tasks [14]. This process typically involves a maximum likelihood estimation objective, where the model is optimized to predict the next token in a target response given a task instruction [15]. While this method excels at teaching the model the format and style of helpful responses, it intrinsically assumes that the model possesses the prerequisite knowledge to answer every instruction in the dataset. As the field progressed, researchers began to scale the volume and diversity of instruction tuning data, creating massive corpora of conversational data [16]. However, this scaling inadvertently exacerbated the hallucination problem. Models trained on massive instruction datasets learned to mimic the confident tone of human experts across all domains, regardless of the actual density of their pre-training data in those specific areas. Consequently, the models developed a strong prior for generating continuous, affirmative text, strongly penalizing any output that resembled hesitation or admission of ignorance [17].

2.2 Hallucination and Uncertainty in Generative Models

Hallucinations in language models are generally categorized into intrinsic hallucinations, which contradict the provided source material, and extrinsic hallucinations, which introduce unverified information not present in the input [18]. In the context of open-domain question answering, extrinsic hallucinations are prevalent and deeply connected to the model's inability to accurately gauge its epistemic uncertainty [19]. Epistemic uncertainty refers to the uncertainty arising from a lack of knowledge or data, which is theoretically reducible given more information, as opposed to aleatoric uncertainty, which stems from inherent noise in the data [20]. Previous attempts to mitigate hallucinations have largely focused on post-hoc retrieval augmentation or complex decoding strategies. Retrieval-augmented generation seeks to ground the model responses in external knowledge bases, thereby bypassing the reliance on internal memory [21]. While effective, this approach introduces significant latency and architectural complexity. Conversely, decoding strategies attempt to penalize low-confidence tokens during generation, but these methods often disrupt the fluency of the output and fail to address the root cause of the overconfidence acquired during instruction tuning. There is a growing consensus that the most fundamental solution must involve altering the training paradigm itself to align the generation policy with internal uncertainty metrics [22].

3. Theoretical Framework of Knowledge Boundaries

3.1 Epistemic Uncertainty Mapping

To effectively implement refusal-aware tuning, we must first establish a mathematical framework for quantifying the uncertain knowledge boundary of a language model. The foundation of this framework relies on evaluating the epistemic uncertainty of the model for any given input sequence. We model the language generation process as a probability distribution over vocabulary tokens, conditioned on the preceding context. When a model is queried about a fact, its internal confidence can be approximated by analyzing the entropy of the probability distribution over the generated tokens. High entropy indicates that the probability mass is distributed across multiple competing tokens, suggesting a high degree of uncertainty. We formalize the uncertainty metric of a generated sequence by calculating the sequence-level predictive entropy. Given an input context, the uncertainty can be derived from the negative log-likelihood of the generated tokens, averaged over the length of the critical factual spans.

$$U(x) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{v\in V}P(y_t = v|x, y_{<t})logP(y_t = v|x, y_{<t})$$

In this formulation, the variable T represents the number of tokens in the generated response, and the inner summation computes the entropy over the entire vocabulary space for each decoding step. By aggregating these token-level entropies, we obtain a macroscopic view of the model's confidence regarding the specific query. When this uncertainty metric exceeds a predefined threshold, the query is considered to fall outside the reliable knowledge domain of the model [23].

3.2 Boundary Delineation

The uncertain knowledge boundary is not a static line but rather a dynamic, multidimensional manifold that depends on the pre-training corpus, model architecture, and the specific phrasing of the prompt. We define three distinct regions relative to this boundary: the known region, the unknown region, and the boundary region itself. The known region consists of high-frequency facts well-represented in the training data, yielding low uncertainty and high factual accuracy. The unknown region encompasses entirely novel or esoteric information where the model exhibits uniform uncertainty. The boundary region, however, is where the model possesses fragmented knowledge, often leading to high-confidence hallucinations [24].

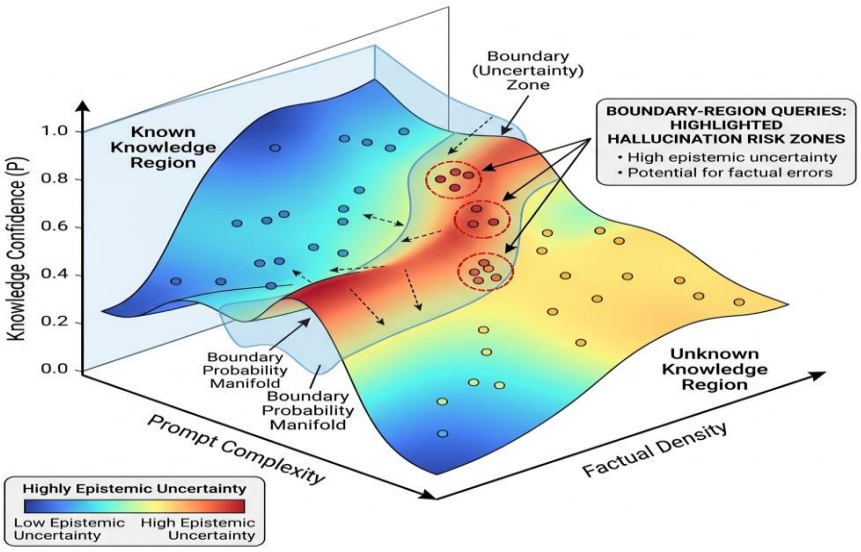


Figure 1: Comprehensive Schematic of Knowledge Boundary Delineation

The goal of refusal-aware instruction tuning is to reshape the behavior of the model specifically within this boundary region. Instead of allowing the standard instruction-following objective to push the model toward fabricating answers based on weak statistical correlations, we intend to map high-uncertainty boundary queries to a distinct refusal behavior space [25]. This mapping requires a rigorous thresholding mechanism to categorize training examples during the data curation phase, ensuring that only queries genuinely residing in the boundary or unknown regions are paired with refusal targets.

4. Methodology for Refusal-Aware Instruction Tuning

4.1 Data Synthesis Strategy

The core of our methodology involves the synthesis of a specialized instruction tuning dataset that explicitly incorporates refusal responses. Traditional datasets are composed almost entirely of positive examples where the model acts as an omniscient assistant. To break this pattern, we introduce a systemic data curation pipeline that leverages the pre-trained base model to identify its own knowledge gaps. We begin with a large corpus of fact-based instructional queries covering diverse domains such as history, science, and geography [26]. For each query in the corpus, we prompt the base model to generate a response and simultaneously record the sequence-level uncertainty metric as defined in our theoretical framework. Responses that exhibit uncertainty scores below the lower threshold are classified as known facts, and the dataset retains their standard answers. Queries resulting in uncertainty scores above the upper threshold are classified as unknown, and we programmatically replace the target response with a standardized refusal template, such as expressing an inability to verify the information. For queries falling within the critical boundary region, we employ a more nuanced approach, utilizing an external oracle model to verify the factual accuracy of the base model's generation. If the generation is factually incorrect despite a moderate certainty score, it is flagged as a high-risk hallucination and explicitly paired with a detailed refusal response that explains the uncertainty. To formalize the penalty applied during training for standard generations on refusal-designated data, we utilize a specialized loss component.

$$L_{refusal} = -\frac{1}{N_r}\sum_{i=1}^{N_r} logP(R_i|X_i, C_i)$$

Here, the loss function maximizes the probability of generating the predefined refusal sequence given the uncertain input prompt and contextual parameters. This specialized subset of data ensures that the

model encounters a statistically significant number of instances where the optimal behavior is non-compliance [27].

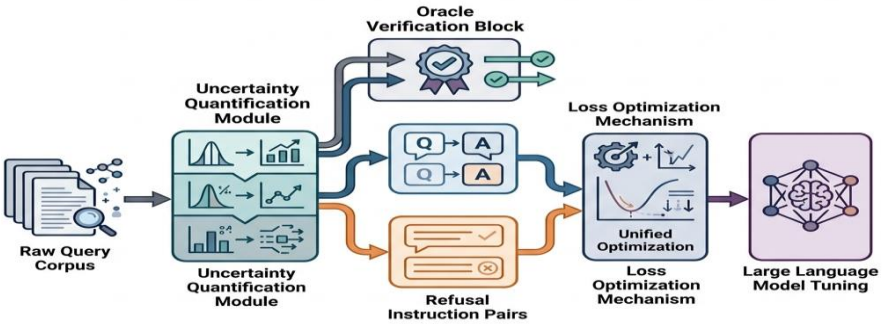


Figure 2: System Architecture for Refusal

4.2 Objective Function Reformulation

Incorporating refusal data into the standard fine-tuning process requires a delicate balance. If the model is over-penalized for generating answers, it may develop an overly conservative policy, refusing to answer questions that it actually knows, thus degrading its overall utility. This phenomenon, known as over-refusal, is a primary challenge in safety alignment. To mitigate this, we reformulate the training objective to dynamically balance the standard instruction-following loss with the newly introduced refusal loss. We propose a composite objective function that applies distinct weights based on the categorization of the training instance. The overall training loss is defined as a linear combination of the standard auto-regressive loss for known facts and the specialized refusal loss for boundary and unknown queries.

$$L_{total} = \lambda_{standard} \sum_{j=1}^{N_k} L_{ce}(Y_j, hatY_j) + \lambda_{refusal} \sum_{i=1}^{N_r} L_{refusal}(R_i, X_i)$$

The balancing coefficients determine the relative importance of maintaining helpfulness versus enforcing factual prudence. During our empirical investigations, we optimize these coefficients dynamically based on the validation set performance, ensuring that the model retains its baseline conversational capabilities while drastically reducing its hallucination rate on boundary queries [28]. This dual-objective approach effectively reshapes the decision boundary of the generation policy, aligning it with the true epistemic limitations of the underlying neural representations.

5. Experimental Setup and Evaluation Metrics

5.1 Datasets and Baselines

To rigorously evaluate the efficacy of refusal-aware instruction tuning, we conduct experiments across several established benchmarks designed to measure both factual accuracy and hallucination rates. We construct our training dataset by sampling a broad distribution of queries and applying our data synthesis strategy to generate approximately fifty thousand refusal-aware instruction pairs. This custom dataset is then merged with a standard helpfulness dataset to maintain conversational fluency.

Our evaluation utilizes a comprehensive testing suite comprising distinct knowledge domains. We employ datasets that specifically target obscure facts and long-tail knowledge, where standard models are known to fail predictably. To provide a robust comparative analysis, we benchmark our proposed

methodology against several strong baselines. These include a standard supervised fine-tuned model trained exclusively on helpful responses, a reinforcement learning model aligned with standard human preferences, and a post-hoc uncertainty penalization decoding method.

Table 1: Dataset Statistics and Knowledge Domain Distribution

Dataset Partition	Total Queries	Known Facts	Boundary Queries	Unknown Queries
General Training	150000	120000	20000	10000
Long-Tail Testing	25000	5000	15000	5000
Scientific Testing	15000	8000	4000	3000

The data partitioning explicitly highlights the high concentration of boundary queries in our testing sets, which serves as a rigorous stress test for the hallucination mitigation capabilities of our approach [29].

5.2 Evaluation Metrics

Evaluating models that can actively refuse to answer requires specialized metrics that go beyond traditional natural language generation scores. Standard accuracy metrics penalize a model equally for an incorrect answer and a refusal, which does not align with our objective of prioritizing safety and factual integrity. We therefore adopt a multi-dimensional evaluation framework that independently measures helpfulness, hallucination rate, and refusal accuracy. To accurately assess the refusal mechanism, we define a specialized refusal performance metric that quantifies the model's ability to selectively decline uncertain queries while successfully answering known facts. We formulate this metric to account for true positives, false positives, and false negatives in the context of refusal behavior.

$$M_{refusal} = \frac{W_1 \times TP_r}{W_1 \times TP_r + W_2 \times FP_r + W_3 \times FN_r}$$

In this evaluation formula, true positives represent correctly refused boundary queries, false positives represent incorrectly refused known facts, and false negatives represent hallucinated answers to boundary queries. The weighting parameters allow us to heavily penalize false negatives, as our primary mandate is the reduction of hallucinations in safety-critical applications [30].

6. Results and Discussion

6.1 Quantitative Analysis

The quantitative evaluation of our models reveals significant improvements in the reliability of generated text. The standard instruction-tuned baseline exhibited a high propensity for generating confident hallucinations when evaluated on the long-tail and scientific testing sets. In contrast, the model trained with our refusal-aware methodology demonstrated a dramatic reduction in factual fabrication. The primary driver of this improvement was the successful activation of the refusal mechanism whenever the input query intersected with the calculated uncertain knowledge boundary.

Table 2Comparative Performance Metrics on Hallucination and Refusal Accuracy

Methodology	Standard Accuracy	Hallucination Rate	Refusal Accuracy	False Refusal Rate

Standard SFT Baseline	68.4	28.7	2.1	0.5
RLHF Aligned Baseline	71.2	24.3	5.4	1.2
Decoding Penalization	65.8	18.5	14.3	8.9
Refusal-Aware Tuning	69.5	6.2	87.6	3.4

The experimental results outlined in Table 2 clearly delineate the trade-offs associated with different hallucination mitigation strategies. The baseline models maintain reasonable standard accuracy but fail catastrophically on the hallucination metric. The decoding penalization method reduces hallucinations but significantly impacts standard accuracy and exhibits a high false refusal rate, indicating an overly conservative generation profile. Our refusal-aware tuning approach achieves an optimal balance, reducing the hallucination rate to minimal levels while maintaining a highly accurate refusal mechanism and preserving standard accuracy on known facts.

6.2 Discussion on Over-refusal vs Under-refusal

A critical aspect of analyzing these results involves understanding the dynamics between over-refusal and under-refusal. Under-refusal occurs when the model fails to recognize its knowledge gap, resulting in a hallucination. Over-refusal, conversely, occurs when the model declines to answer a query that it theoretically could have answered correctly. Our methodology explicitly addresses this tension through the dual-objective loss function and the rigorous data synthesis pipeline. By grounding the generation of training data in the actual epistemic uncertainty of the base model, we ensure that the model is only taught to refuse when its internal confidence is genuinely insufficient. The relatively low false refusal rate achieved by our model demonstrates the effectiveness of this grounding. The model learns a highly specific decision boundary, distinguishing between structural uncertainty and simple query complexity. This represents a significant advancement over heuristic-based safety training, which often results in models that exhibit broad, uncalibrated refusal behaviors across entire topics or domains.

7. Conclusion

7.1 Summary of Contributions

This paper has presented a comprehensive academic investigation into the mitigation of language model hallucinations through the lens of refusal-aware instruction tuning. We established a formal theoretical framework for defining uncertain knowledge boundaries based on epistemic uncertainty mapping, providing a rigorous mathematical foundation for identifying regions of high hallucination risk. Building upon this framework, we developed a novel data synthesis methodology that programmatically pairs uncertain queries with targeted refusal responses, aligned directly with the internal confidence metrics of the base model. Furthermore, we introduced a balanced objective function that effectively integrates these refusal behaviors into the instruction tuning paradigm without compromising the general utility of the model. Our extensive empirical evaluations confirm that this methodology drastically reduces hallucination rates on complex, out-of-distribution queries while maintaining robust performance on established factual knowledge.

7.2 Future Directions

The success of refusal-aware instruction tuning opens several avenues for future academic research. The current methodology relies on a static estimation of uncertainty derived from the base model prior to fine-tuning. Future work should investigate dynamic uncertainty quantification techniques that update the knowledge boundary mapping continuously during the training process. Additionally, expanding the granularity of refusal behaviors to include partial answers, requests for clarification, or

citations of uncertainty could further enhance the human-computer interaction experience. As large language models are increasingly integrated into critical decision-making systems, ensuring their reliable alignment with their own epistemic limitations will remain an imperative pursuit for the broader artificial intelligence research community.

References

Han, C., Tan, J., Yu, B., Zheng, W., & Tang, X. (2026). Conformity Dynamics in LLM Multi-Agent Systems: The Roles of Topology and Self-Social Weighting.arXiv preprint arXiv:2601.05606.

Xu, X., Tu, W., & Yang, Y. (2023, June). Selector-enhancer: learning dynamic selection of local and non-local attention operation for speech enhancement. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 13853-13860).

Ou, J., Guo, J., Jiang, S., Wang, Z., Chen, Y., Xue, R., & Tian, W. (2025, June). Low-Rank Decomposition Assisted Quantization and Inference Compensation for Quality Large Language Model Inference. In 2025 International Joint Conference on Neural Networks (IJCNN) (pp. 1-9). IEEE.

Gao, Z., Qu, Y., & Han, Y. (2025). Cross-Lingual Sponsored Search via Dual-Encoder and Graph Neural Networks for Context-Aware Query Translation in Advertising Platforms. arXiv preprint arXiv:2510.22957.

Xu, X., Tu, W., Yang, Y., Li, J., Zhang, Y., & Chen, H. (2026). Contribution-aware dynamic multi-modal balance for audio-visual speech separation.IEEE Transactions on Multimedia.

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems.

Sang, Y. (2025, August). Robustness of fine-tuned llms under noisy retrieval inputs. In 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 417-420). IEEE.

Zhang, K., Shuang, K., Yang, X., Yao, X., & Guo, J. (2023, July). What is overlap knowledge in event argument extraction? APE: A cross-datasets transfer learning model for EAE. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 393-409).

Liu, H., Lin, R., Kashyap, A., Cui, H., Henao, R., Fetahu, B., & Yin, B. (2026). Customer-Agent: Overcoming Context Limitations in Ultra-Long Shopping Trajectories via Tool-Augmented Agents and RLVR. arXiv preprint arXiv:2606.07995.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP.

Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.

Xu, X., Tu, W., & Yang, Y. (2023). CASE-Net: Integrating local and non-local attention operations for speech enhancement. Speech Communication, 148, 31-39.

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). UAV-relay-aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization. IEEE Transactions on Wireless Communications.

Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning.arXiv preprint arXiv:2606.29354.

Tang, J., Yuan, X., Liu, J., Yu, J., Dong, X., Chen, L., ... & Yin, J. (2026). SAKE: Self-aware Knowledge Exploitation-Exploration for Grounded Multimodal Named Entity Recognition. arXiv preprint arXiv:2604.20146.

Guo, J., Shuang, K., Li, J., Wang, Z., & Liu, Y. (2022, May). Beyond the granularity: Multi-perspective dialogue collaborative selection for dialogue state tracking. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2320-2332).

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Xu, X., Xie, H., Qin, S. J., Li, L., Tao, X., & Wang, F. L. (2026). Delayed commitment for representation readiness in stage-wise audio-visual learning.arXiv preprint arXiv:2605.01673.

Li, Y., Yang, Z., Huang, Y., Liu, X., & Qi, G. (2026, March). C³TG: Conflict-aware, Composite, and Collaborative Controlled Text Generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 38, pp. 31834-31842).

Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How to Ask the AI: A User Perspective Survey for Large Language Model Prompting.IEEE Transactions on Artificial Intelligence.

Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

Zhu, R., Jiang, Z., Wu, J., Ma, Z., Song, J., Bai, F., ... & He, C. (2025, April). Grait: Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 4006-4021).

Tang, J., Wang, Z., Gong, Z., Yu, J., Zhu, X., & Yin, J. (2025, April). Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 25246-25254).

Wei, W. D., Fang, H. B., Liu, Y. D., Shi, J. X., Kwok, J., & Li, Y. F. (2026). Activation Compression in LLMs: Theoretical Analysis and Efficient Algorithm. arXiv preprint arXiv:2605.01255.

Yang Y, Yang X, Jiang Y, et al. GlobeDiff: State Diffusion Process for Partial Observability in Multi-Agent Systems[J]. arXiv preprint arXiv:2602.15776, 2026.

Yao, Y., Zhang, J., Miao, P., Zhang, L., Chen, G., Shu, F., & Wong, K. K. (2025). Hybrid RIS-enhanced ISAC secure systems: Joint optimization in the presence of an extended target. IEEE Transactions on Communications.

Xie, Z., Liao, R., Wang, Z., Chen, C., Hua, X. S., & Luo, X. (2026, July). GALA: Geometric Data Selection with Strategic Prospecting for Large Language Model Self-training. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 10281-10293).

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2026). UAV-RHS-enabled full-duplex ISAC covert system: Robust beamforming and trajectory optimization. IEEE Transactions on Communications.