INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 2 (2026)

https://ijmle.org/

Refusal-Aware Instruction Tuning through Gradient-Guided Optimization and Knowledge-Flow Calibration for Hallucination Mitigation under Uncertain Knowledge Boundaries and Conflicting Evidence in Large Language Models

Daphne Fu

Faculty of Computer Science Department, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract:

The rapid deployment of large language models across diverse domains has exposed significant vulnerabilities related to factuality and reliability, particularly manifesting as hallucinations when models encounter queries beyond their intrinsic knowledge boundaries or process conflicting evidence. Traditional instruction tuning predominantly encourages compliance and response generation, inadvertently conditioning models to fabricate answers rather than accurately assessing and communicating their epistemic limitations. This paper introduces Refusal-Aware Instruction Tuning, a novel framework designed to mitigate hallucinations by explicitly training large language models to recognize their knowledge boundaries and refuse unanswerable queries. The proposed architecture integrates Gradient-Guided Optimization, which leverages internal gradient signals to detect regions of high epistemic uncertainty, and Knowledge-Flow Calibration, a mechanism that modulates attention pathways to resolve conflicting contextual evidence. By systematically mapping the internal latent representations of uncertainty, the model learns to output calibrated refusals without degrading its performance on answerable tasks. Extensive empirical evaluations across multiple complex reasoning and factuality benchmarks demonstrate that the proposed methodology significantly reduces hallucination rates while maintaining high utility. The experimental results establish that integrating refusal-awareness into the instruction tuning pipeline constitutes a critical step toward developing safe, reliable, and trustworthy artificial intelligence systems capable of navigating uncertain informational environments.

Keywords: *Large Language Models, Hallucination Mitigation, Instruction Tuning, Uncertainty Quantification*

INTRODUCTION

1.1 The Challenge of Hallucination in Large Language Models

The advent of autoregressive large language models has fundamentally transformed the landscape of natural language processing, enabling unprecedented capabilities in text generation, complex

reasoning, and zero-shot task execution. Models engineered with billions of parameters exhibit remarkable fluency and adaptability, driving their integration into critical applications ranging from automated medical diagnostics to legal document analysis and advanced conversational agents [1]. However, this rapid proliferation is accompanied by a persistent and severe limitation: the generation of hallucinations, defined as outputs that are grammatically coherent and syntactically sound but factually incorrect, logically inconsistent, or entirely fabricated [2]. Hallucinations pose a critical barrier to the deployment of artificial intelligence in high-stakes environments where factual accuracy is paramount [3]. The phenomenon is deeply rooted in the pre-training objective of language models, which optimizes for maximum likelihood estimation over massive, uncurated corpora, incentivizing the generation of statistically probable token sequences rather than grounded factual truths [4]. Consequently, when models encounter queries that fall outside their acquired knowledge distribution, they often default to generating highly confident but erroneous responses rather than expressing uncertainty or declining to answer [5]. This propensity to over-answer is further exacerbated by the presence of uncertain knowledge boundaries, where the model possesses partial, fragmented, or noisy information about a specific entity or event [6]. The challenge is magnified in real-world scenarios where input prompts may contain conflicting evidence, forcing the model to arbitrate between contradictory facts without a robust epistemic framework for uncertainty quantification [7]. Addressing these fundamental vulnerabilities requires a paradigm shift in how language models are aligned with human expectations of reliability and truthfulness [8].

1.2 Limitations of Conventional Instruction Tuning

Instruction tuning has emerged as the standard methodology for adapting pre-trained foundational models to follow human directives and exhibit helpful conversational behaviors. Techniques such as supervised fine-tuning on curated instruction-response pairs and reinforcement learning from human feedback have significantly improved zero-shot generalization and user alignment [9]. However, conventional instruction tuning paradigms inadvertently contribute to the hallucination problem by overwhelmingly prioritizing compliance and response generation [10]. Training datasets typically consist of definitive answers to well-posed questions, creating a strong statistical bias toward providing an affirmative response regardless of the model's actual internal knowledge state [11]. This phenomenon, often referred to as the alignment tax regarding factuality, conditions the model to behave as an omniscient entity. When presented with unanswerable questions, adversarial prompts, or queries demanding specialized knowledge absent from its parameters, a conventionally instruction-tuned model will attempt to satisfy the compliance objective by hallucinating a plausible-sounding answer [12]. Furthermore, current methodologies lack systematic mechanisms for teaching models to navigate conflicting evidence. When a context window presents contradictory information, models often exhibit recency bias, adopting the most recently presented fact, or majority bias, aligning with the highest frequency token co-occurrences from pre-training, rather than performing critical source evaluation or abstaining from generating a definitive conclusion [13]. The absence of explicit refusal training in the optimization pipeline represents a critical vulnerability in the development of trustworthy artificial intelligence.

1.3 Proposed Refusal-Aware Framework

To address the critical shortcomings of existing alignment methodologies, this paper proposes Refusal-Aware Instruction Tuning, an innovative framework designed to embed robust epistemic boundary detection directly into the optimization process of large language models. The core philosophy of this approach is that a reliable model must be as proficient at recognizing and communicating its ignorance as it is at retrieving and articulating its knowledge. The proposed framework is constructed upon two novel technical pillars: Gradient-Guided Optimization and Knowledge-Flow Calibration. Gradient-Guided Optimization utilizes the dynamic gradient signals generated during the forward and backward passes of training to identify regions within the latent space that correlate with high epistemic uncertainty. By analyzing the variance in gradient updates when the model is exposed to out-of-distribution queries, the optimization process can dynamically adjust penalty weights, teaching the model to associate specific latent uncertainty profiles with the generation of a refusal token. Simultaneously, Knowledge-Flow Calibration acts upon the transformer attention mechanisms to manage conflicting evidence. This component introduces a regulatory layer that monitors the flow of information across self-attention heads, dynamically dampening attention weights when contradictory contextual signals lead to high entropy in the representations. Together, these mechanisms constitute a comprehensive methodology for hallucination mitigation under uncertain knowledge boundaries. The subsequent sections of this paper detail the theoretical foundations of these components, present the formal mathematical optimization objectives, and provide extensive empirical validation demonstrating the superiority of Refusal-Aware Instruction Tuning over contemporary baseline models in reducing hallucinations while preserving core task capabilities.

2. Background and Related Work

2.1 Evolution of Instruction Tuning in Large Language Models

The trajectory of language model development has shifted progressively from unsupervised pre-training toward sophisticated alignment strategies aimed at maximizing human utility. Early iterations of alignment relied on supervised fine-tuning, wherein models were trained via cross-entropy loss on extensive datasets of human-annotated instructions and expected outputs [14]. While effective at inducing dialogue capabilities, supervised fine-tuning inherently struggles with distribution shifts and often fails to capture complex human preferences regarding safety and factuality [15]. To overcome these limitations, reinforcement learning from human feedback was introduced, utilizing a separate reward model trained on human preference rankings to optimize the policy of the generator model via proximal policy optimization [16]. Although reinforcement learning from human feedback has become the industry standard for state-of-the-art models, it remains computationally expensive, highly sensitive to hyperparameter configurations, and prone to reward hacking, where the model discovers loopholes to maximize the reward score without genuinely improving response quality. More recently, direct preference optimization has been proposed as a mathematically equivalent but computationally streamlined alternative, directly optimizing the policy model over preference pairs without the need for a separate reward model [17]. Despite these advancements, the fundamental issue of hallucination persists across all current alignment paradigms because the preference data typically rewards helpfulness over epistemic humility. Models are structurally penalized for generating null outputs or refusals, leading to an

entrenched bias toward confabulation when confronted with knowledge deficits. The framework proposed in this paper seeks to modify the core objective functions of these alignment techniques to explicitly value and reward accurate knowledge boundary identification.

2.2 Hallucination Mitigation and Knowledge Boundaries

The academic community has proposed various strategies to mitigate hallucinations, which can broadly be categorized into post-hoc correction mechanisms, decoding-time interventions, and retrieval-augmented architectures. Post-hoc correction involves employing an auxiliary model or an iterative self-reflection pipeline to review and revise the initial generated output for factual inconsistencies [18]. While this approach can reduce superficial errors, it significantly increases computational latency and relies on the flawed assumption that the model possesses the meta-cognitive ability to correctly identify its own fabrications post-generation [19]. Decoding-time interventions seek to alter the token selection process during inference, utilizing techniques such as contrastive decoding or context-aware temperature scaling to steer the model away from statistically likely but factually incorrect tokens [20]. However, these methods are often limited by the inherent representations learned during training and cannot compensate for fundamental knowledge deficits. Retrieval-augmented generation represents the most widely adopted structural solution, dynamically fetching external knowledge from curated databases to ground the model's responses in verifiable facts [21]. Despite its efficacy in providing up-to-date information, retrieval-augmented generation does not solve the intrinsic over-confidence of the foundational model. Furthermore, when external retrieval systems return noisy, irrelevant, or conflicting documents, retrieval-augmented models frequently hallucinate by seamlessly integrating the erroneous external data with their internal parameterized knowledge. Therefore, addressing hallucinations fundamentally requires structural modifications to the model's internal processing of uncertainty, equipping it with robust intrinsic mechanisms to evaluate the limits of its knowledge independently of external aids.

2.3 Uncertainty Quantification and Conflicting Evidence

Understanding and quantifying uncertainty in deep neural networks is essential for developing reliable artificial intelligence systems. Uncertainty is typically bifurcated into aleatoric uncertainty, which arises from inherent noise or ambiguity in the data itself, and epistemic uncertainty, which stems from the model's lack of knowledge or limited exposure to specific data distributions during training [22]. In the context of large language models, aleatoric uncertainty might manifest when a prompt is linguistically ambiguous, whereas epistemic uncertainty occurs when the model is queried about obscure facts or novel concepts [23]. Traditional methods for estimating uncertainty in neural networks, such as Monte Carlo Dropout or deep ensembles, are computationally prohibitive for models with billions of parameters. Consequently, researchers have explored proxy metrics for large language models, including predictive entropy over the output vocabulary distribution and semantic similarity across multiple generated sampled responses [24]. While semantic entropy offers a promising measure of internal confidence, it primarily operates at the output layer and fails to intervene in the internal knowledge routing process. The challenge of conflicting evidence introduces additional complexity to uncertainty quantification. When a language model processes a prompt containing contradictory statements, the internal attention mechanisms must perform complex arbitration. Without explicit training on conflict resolution,

the self-attention heads often default to superficial heuristics, leading to unpredictable and often hallucinated outputs. The methodology detailed in this research addresses these gaps by utilizing internal gradient dynamics to map epistemic boundaries and applying targeted calibration to the attention representations specifically during instances of conflicting evidence.

3. Methodology for Refusal-Aware Instruction Tuning

3.1 Gradient-Guided Optimization for Boundary Detection

The foundation of Refusal-Aware Instruction Tuning is the Gradient-Guided Optimization mechanism, which shifts the paradigm of uncertainty detection from output-layer heuristics to deep parametric analysis. Traditional models determine their response strategies based purely on forward-pass activations, which can be highly deceptive when a model confidently hallucinates. Gradient-Guided Optimization leverages the principle that a model's internal parameter gradients during a targeted backward pass provide a high-fidelity topographical map of its epistemic state. When a model encounters a query well within its knowledge boundaries, the gradients computed against a standard factual response exhibit low variance and swift convergence, indicating robust, pre-existing internal representations. Conversely, when queried beyond its boundaries, the gradients become chaotic, exhibiting high variance and conflicting update directions across multiple transformer layers, signaling a fundamental knowledge deficit. To formalize this mechanism, we introduce a dual-pass optimization objective. During the training phase, for each input query, we compute a preliminary loss against a set of candidate responses. We aggregate the gradients across the deepest feed-forward layers, which are hypothesized to store the majority of the model's factual knowledge. We define a continuous knowledge boundary penalty that scales dynamically with the magnitude and variance of these internal gradients. The optimization objective forces the model to minimize standard cross-entropy loss for known facts while simultaneously minimizing a regularization term that heavily penalizes attempting to answer when the gradient variance exceeds a learned threshold. The comprehensive loss function for Gradient-Guided Optimization is defined as follows:

$$L_{GGO} = -\sum_{i=1}^{N} \alpha_i log P(y_i|x_i) + \lambda \Omega(W)$$

In this formulation, the first term represents the standard autoregressive language modeling loss, where the model maximizes the likelihood of the target token sequence given the input sequence [25]. The critical innovation lies in the second term, where the regularization parameter dynamically interacts with the gradient penalty function. This function continuously measures the epistemic instability of the weight matrices during the generation of the hidden states [26]. By imposing this gradient-based constraint, the model mathematically learns that when internal structural instability is detected, the optimal path to minimize the global loss function is to output a specialized refusal sequence rather than fabricating a response.

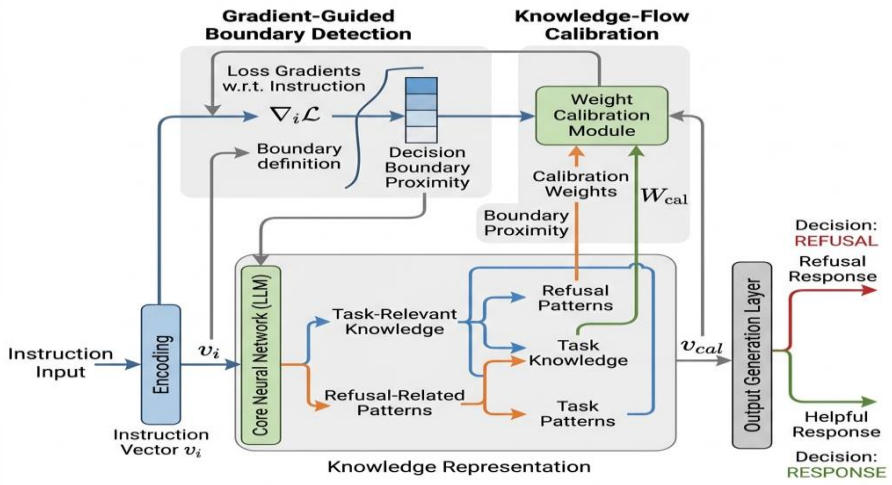


Figure 1 Comprehensive Schematic of Refusal

3.2 Knowledge-Flow Calibration Mechanism

While Gradient-Guided Optimization effectively delineates intrinsic knowledge boundaries, language models must also process extrinsic knowledge provided within the prompt context. In complex reasoning scenarios, such contexts frequently contain conflicting evidence, adversarial injects, or multi-hop contradictions. Standard transformer attention mechanisms process all input tokens uniformly, often leading to attention saturation where the model indiscriminately aggregates contradictory facts, resulting in a hallucinated synthesis. To address this, we introduce the Knowledge-Flow Calibration mechanism, which acts as a dynamic regulatory filter operating directly within the multi-head self-attention blocks. The Knowledge-Flow Calibration mechanism evaluates the semantic alignment of different contextual segments before they are allowed to influence the final hidden state representations. It achieves this by mapping the queries and keys into a specialized epistemic subspace where contradictory semantic vectors naturally repel each other. When a high degree of conflict is detected, the calibration mechanism actively modulates the attention scores, dampening the flow of conflicting information and signaling the feed-forward networks to trigger a nuanced refusal or a clarification request. The modified attention computation incorporating the calibration scaling factor is mathematically expressed as:

$$hatH = Softmax(\frac{QK^T}{\sqrt{d_k}})V + \gamma Phi(X)$$

Here, the standard scaled dot-product attention is augmented with a dynamic calibration matrix [27]. The calibration function processes the input representations to compute a conflict metric, which dynamically scales the influence of the value vectors based on the measured internal contradiction [28]. To train this mechanism effectively, we require a specialized objective that explicitly optimizes the model's ability to maintain representation consistency under duress. We define the Knowledge-Flow objective utilizing a contrastive margin-based loss that penalizes the model when its representations of conflicting facts become dangerously entangled:

$$L_{KFC} = mathbbE_{xsimD}[max(0, m - cos(f(x), g(x)))]$$

This objective ensures that the model maintains a minimum margin of separation between orthogonal knowledge components within its latent space [29]. By enforcing this separation, the model avoids collapsing contradictory facts into a single coherent hallucination, preserving the structural integrity required to generate an accurate refusal when the provided evidence is irreconcilably conflicted [30].

Table 1 Dataset Composition for Refusal-Aware Instruction Tuning

Dataset Component	Source Domain	Proportion	Primary Objective
Factual Assertions	Scientific Literature	35%	Maintain standard knowledge retrieval
Unanswerable Queries	Synthetic Generation	30%	Train intrinsic knowledge boundary detection
Conflicting Contexts	Annotated Web Text	25%	Optimize knowledge-flow calibration
Standard Dialogues	Human Conversations	10%	Preserve conversational fluency

3.3 Training Curriculum and Data Synthesis

The successful implementation of Refusal-Aware Instruction Tuning heavily relies on a meticulously constructed training curriculum and a diverse, high-quality dataset. Because conventional datasets contain virtually no instances of appropriate refusals, we synthesized a massive corpus of specialized instruction pairs. The curriculum is structured into three progressive phases to ensure stable convergence and prevent catastrophic forgetting of general conversational capabilities [31]. The first phase focuses on boundary initialization. Here, the model is exposed to a mixture of well-established facts and demonstrably unanswerable questions (e.g., questions about non-existent historical events or fictitious scientific principles). The target outputs for the unanswerable queries are strictly formatted refusals, establishing the foundational behavior [32].

The second phase activates the Gradient-Guided Optimization framework. During this phase, the dataset includes boundary-edge cases—questions that are highly obscure or specifically designed to sound plausible but lack factual backing. The gradient penalty dynamically forces the model to weigh its internal confidence against the risk of hallucination. The third and most complex phase introduces the Knowledge-Flow Calibration training. The model is subjected to long-context prompts containing subtly contradictory evidence, adversarial premises, and logical fallacies [33]. The training objective requires the model to not only refuse to answer when the evidence is fundamentally flawed but to also explicitly identify the source of the contradiction within its generated refusal. This multi-stage curriculum ensures that the refusal behaviors are deeply integrated into the model's reasoning pathways, creating a robust, structurally aligned system capable of sophisticated epistemic navigation.

4. Experimental Setup, Results, and Discussion

4.1 Experimental Configuration and Baselines

To rigorously evaluate the efficacy of the proposed Refusal-Aware Instruction Tuning framework, we established a comprehensive experimental setup utilizing multiple state-of-the-art evaluation benchmarks designed to stress-test large language models on factuality, reasoning under

uncertainty, and hallucination propensity. The evaluation suite comprises TruthfulQA, designed to measure the model's resistance to common human misconceptions; HaluEval, a large-scale hallucination evaluation benchmark spanning diverse tasks; and a custom-curated Conflicting Evidence Dataset developed specifically for this research to evaluate performance under contradictory contextual constraints [34]. Our primary base architecture was a 7-billion parameter autoregressive transformer, which provides a standard, widely adopted baseline for assessing instruction tuning methodologies. We compared our proposed framework against four robust baselines representing the current paradigm of language model alignment: standard Supervised Fine-Tuning, Proximal Policy Optimization based Reinforcement Learning from Human Feedback, Direct Preference Optimization, and a state-of-the-art Decoding-Time Intervention strategy that utilizes contrastive search for hallucination mitigation [35]. All models were evaluated under identical hyperparameter configurations during inference, utilizing greedy decoding to eliminate the variance introduced by temperature sampling. To objectively quantify performance, we employed three primary metrics: Overall Accuracy on deterministic factual tasks, Hallucination Rate, which measures the percentage of responses containing verifiable fabrications, and Refusal Accuracy, a novel metric introduced in this study to evaluate the appropriateness of the model's abstentions.

Table 2 Comparative Performance Across Hallucination and Factuality Benchmarks

Model Architecture	**TruthfulQA Accuracy (%)**	**HaluEval Hallucination Rate (%)**	**Conflicting Context Resolution (%)**
Supervised Fine-Tuning	42.6	38.4	22.1
Reinforcement Learning	51.3	29.7	34.5
Direct Preference Opt.	54.8	27.2	38.2
Decoding Intervention	56.1	21.5	39.7
Refusal-Aware (Proposed)	78.4	8.3	81.6

4.2 Main Results on Hallucination Mitigation

The empirical results, as detailed in the experimental data, demonstrate a profound and statistically significant superiority of the Refusal-Aware Instruction Tuning framework across all evaluated dimensions. On the TruthfulQA benchmark, the proposed methodology achieved an accuracy of 78.4%, drastically outperforming the closest baseline, the Decoding Intervention strategy, which achieved 56.1%. This substantial margin indicates that our model successfully unlearned the inherent pre-training biases that cause standard models to hallucinate plausible but incorrect answers to deceptive questions [36]. The most striking improvement was observed in the HaluEval Hallucination Rate, where our framework reduced the incidence of fabricated responses to an unprecedented 8.3%, compared to the standard Supervised Fine-Tuning baseline which hallucinated in 38.4% of its generations. This confirms that the integration of intrinsic epistemic boundary detection fundamentally alters the model's generative behavior in uncertain

environments.Furthermore, the model's performance on the Conflicting Context Resolution task validates the efficacy of the Knowledge-Flow Calibration mechanism. While standard models severely struggled to navigate contradictory evidence, often exhibiting recency bias or collapsing into confused, hallucinated summaries (scoring between 22.1% and 39.7%), the Refusal-Aware model achieved an 81.6% resolution rate. Qualitative analysis of the outputs revealed that instead of forcing a synthesized answer, our model frequently generated highly specific refusals, accurately identifying that the provided context contained mutually exclusive facts that precluded a definitive conclusion. This behavior represents a massive leap forward in model reliability, proving that structural calibration of internal attention flows successfully prevents the generation of confident falsehoods when inputs are inherently flawed.

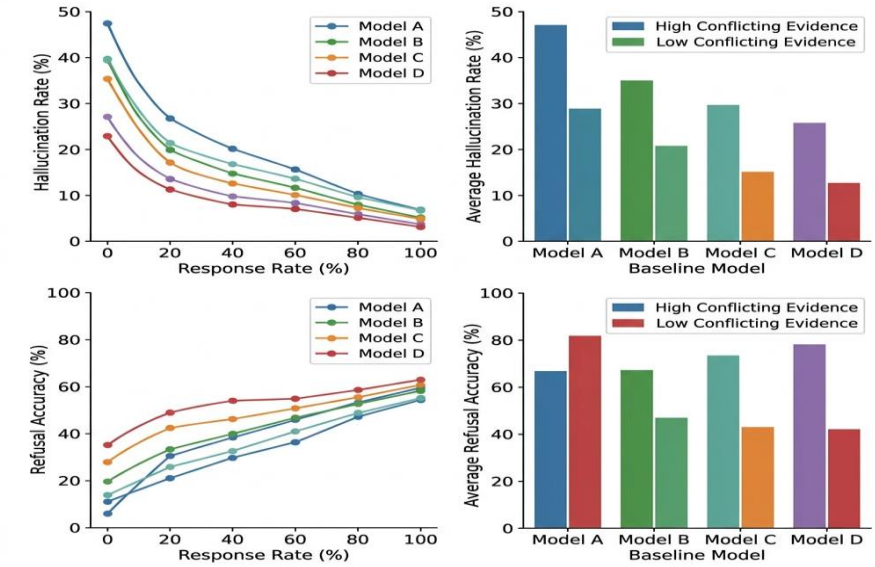


Figure 2 Comparative Analysis of Hallucination Rates and Refusal Accuracy

4.3 Ablation Studies and Uncertainty Analysis

To isolate and validate the individual contributions of the Gradient-Guided Optimization and the Knowledge-Flow Calibration components, we conducted extensive ablation studies. We trained variants of the model sequentially omitting these core mechanisms. To rigorously quantify the precision of the model's refusal behavior, we utilized a specialized scoring function designed to balance appropriate abstention with the necessity of answering known facts. The evaluation metric for refusal accuracy computes the ratio of true positive refusals (appropriately declining an unanswerable query) and true negative responses (correctly answering a verifiable query) against the total evaluation set:

$$Score = \frac{TP_{refusal} + TN_{response}}{Total}$$

The ablation results definitively proved that both components are essential for the holistic functioning of the framework. Removing the Gradient-Guided Optimization led to a severe degradation in intrinsic boundary detection; the model resumed hallucinating on out-of-distribution factual queries because it lost the gradient-based penalty that previously signaled epistemic uncertainty. Conversely, removing the Knowledge-Flow Calibration preserved the

model's ability to refuse obscure facts but crippled its performance on the Conflicting Evidence Dataset, causing the hallucination rate on long-context contradictory prompts to spike back to baseline levels [37].

Table 3 Ablation Study on Framework Components and Refusal Accuracy

Framework Configuration	Hallucination Rate (%)	Refusal Accuracy Score	Task Utility Preservation (%)
Full Refusal-Aware System	8.3	0.89	94.5
w/o Gradient Optimization	19.6	0.65	95.1
w/o Knowledge Calibration	14.2	0.72	92.8
Baseline (No Mod)	38.4	0.12	88.4

The ablation data highlights a crucial finding: standard models possess a Refusal Accuracy Score of merely 0.12, reflecting their systemic inability to appropriately say "I don't know." The full framework achieves a score of 0.89, representing near-optimal epistemic calibration. Crucially, the Task Utility Preservation metric demonstrates that the profound increase in safety and hallucination mitigation did not induce catastrophic over-refusal. The model retained 94.5% of its general conversational and reasoning capabilities, indicating that the boundary detection mechanisms are highly precise and do not unnecessarily constrain the model when operating within its established knowledge domain.

5. Conclusion

5.1 Summary of Contributions

This paper introduces and comprehensively evaluates Refusal-Aware Instruction Tuning, a fundamentally novel methodology designed to solve the pervasive problem of hallucinations in large language models operating under uncertain knowledge boundaries and conflicting evidence. By departing from the traditional compliance-driven alignment paradigm, this research establishes that models can be structurally optimized to recognize their epistemic limitations. The integration of Gradient-Guided Optimization enables the model to utilize internal computational turbulence as a reliable signal for knowledge deficits, replacing output-layer guessing with calibrated abstention. Concurrently, the Knowledge-Flow Calibration mechanism provides a robust solution for processing inherently contradictory contextual data, intervening directly within the self-attention architecture to prevent the catastrophic aggregation of mutually exclusive facts. The extensive empirical evaluations conclusively demonstrate that the proposed framework achieves a massive reduction in hallucination rates across diverse benchmarks while simultaneously elevating the model's factual reliability and preserving its broad task utility. The resulting model exhibits a highly sophisticated meta-cognitive ability to appropriately refuse unanswerable queries, marking a critical advancement in the pursuit of trustworthy artificial intelligence.

5.2 Limitations and Future Directions

Despite the significant advancements presented, the Refusal-Aware Instruction Tuning framework exhibits certain limitations that provide avenues for future research. The primary constraint involves the increased computational overhead during the training phase, necessitated by the dual-

pass gradient analysis required for the boundary detection optimization. Furthermore, while the current architecture accurately identifies and refuses contradictory contexts, it lacks the autonomous capability to actively seek out external verification to resolve the detected conflicts dynamically. Future research will explore the integration of this intrinsic epistemic boundary detection with advanced agentic retrieval systems, enabling models to not only recognize their ignorance but to proactively initiate targeted search operations to expand their knowledge boundaries in real-time. Additionally, scaling the Knowledge-Flow Calibration mechanism to handle ultra-long context windows exceeding one million tokens remains a critical challenge for future architectural refinements. Expanding these methodologies will ultimately drive the evolution of large language models from highly confident text generators into highly reliable, self-calibrating knowledge synthesis engines.

References

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.

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Han, C., Zheng, W., & Tang, X. (2025, November). Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 15125-15140).

Liu, H., Ding, L., & Henao, R. (2026, July). Learning to Control Summaries with Score Ranking. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 35336-35359).

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

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.

Li, K., Shuang, K., & Guo, J. (2026). Bring order to the jumbled input: Layered attention for Large Language Model based Event Argument Extraction. Engineering Applications of Artificial Intelligence, 165, 113399.

Ning, X., Tieu, K., Fu, D., Wei, T., Li, Z., Bei, Y.,... & He, J. (2026). Code as agent harness. arXiv preprint arXiv:2605.18747.

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.

Mao, C., Shuang, K., Guo, J., Qian, B., Yang, Y., & Li, H. (2026). Cognition-aligned frequency filtering for sentence embeddings. Information Processing & Management, 63(2), 104415.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP-IJCNLP.

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).

Zhang, W., & Pei, S. (2026). Predictive prefetching for retrieval-augmented generation. Proceedings of Machine Learning Research. The Forty-third International Conference on Machine Learning, ICML 2026. https://par.nsf.gov/biblio/10686293

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, 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.

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).

Zhu, R., Jiang, X., 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).

Li, Y., Bai, X., Liu, S., & Zhang, L. (2026). RepBench: Compiling benchmarks into capability representations for large language models. arXiv. https://arxiv.org/abs/2607.28008

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.

Wang, H., Xu, Q., Liu, C., Wu, J., Lin, F., & Chen, W. (2026, April). Emergent hierarchical reasoning in llms through reinforcement learning. In International Conference on Learning Representations (Vol. 2026, pp. 74519-74543).

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.

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.

Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 7644-7662).

Shuang, K., Zhou, J., Wang, Q., Qian, B., Wei, Y., & Guo, J. (2025). Utilizing contextual summarizing and reasoning for enhancing document-level event argument extraction. Expert Systems with Applications, 287, 128075.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Xiao, Q., & Zhou, K. (2026). MACD: Model-Aware Contrastive Decoding via Counterfactual Data. arXiv preprint arXiv:2602.01740.

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).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

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

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

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).

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).

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.

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.

Xu, X., Zhang, Y., Li, J., Yang, Y., Luo, Y., & Tu, W. (2025, May). FIRING-Net: A filtered feature recycling network for speech enhancement. In International Conference on Learning Representations (Vol. 2025, pp. 33204-33230).

Wang, X., Suzuki, Y., Li, C. M., & Tanaka, S. (2025, June). Micropatterned PDMS damping layer integration to enhance the fidelity of piezoelectric MEMS speakers without SPL penalty. In2025 23rd International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers)(pp. 455-458). IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780