Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.4, No.4, 2024

Enhancing Trustworthiness of Retrieval-Augmented Large Language Models via Confidence Calibration and Selective Rejection

Yihan Xue

University of Southern California, Los Angeles, USA

xueyihan2014@gmail.com

Abstract: This study proposes a robust monitoring method based on structure-aware feature learning to address the challenges of complex multidimensional feature coupling, dynamically changing dependencies, and diverse anomaly patterns in backend systems. The method builds upon multi-source monitoring data, integrating temporal dynamic features and topological dependency structures to achieve unified system state representation across different time scales and structural hierarchies. First, a feature encoder is employed to extract temporal features and construct a dynamic dependency graph that captures latent structural relationships among service nodes. Second, a structure-aware mechanism and graph propagation layer are introduced to perform cross-node feature fusion and apply dependency consistency constraints, enhancing the model's stability and generalization capability in complex topological environments. Finally, variational regularization and a robust optimization objective are applied to further improve anomaly detection reliability under high noise and non-stationary distributions. Experimental results show that the proposed method outperforms existing models across multiple key metrics, including F1-Score, AUROC, and Robust Recall, while maintaining stable monitoring performance under structural perturbations, workload fluctuations, and feature heterogeneity. These results validate the effectiveness of structure-aware feature learning in complex system monitoring tasks, demonstrating its ability to accurately model anomaly propagation paths and achieve system-level semantic recognition in the feature space, thereby providing strong support for building highly reliable and interpretable backend monitoring frameworks.

Keywords: Structure-aware feature learning; backend system monitoring; anomaly detection; robust modeling.

1. Introduction

The emergence of retrieval-augmented generation stems from the inherent limitations of generative models in open knowledge scenarios, namely, the knowledge fixed in the model parameters cannot cover real-time updates and long-tail details[6]. To address this, the system transforms the question into a retrieval request, retrieves relevant fragments from external corpora, and then uses these fragments as conditional inputs to guide generation, thus transferring knowledge acquisition from memory to retrieval to some extent. This process expands question answering from simple language modeling to a composite process of retrieval, evidence selection, and reasoning organization, making the quality of the answer not only dependent on model capabilities but also subject to the combined influence of corpus construction, indexing strategies, recall ranking, fragment segmentation, and noise filtering[7].

However, the benefits of this multi-stage process also bring new paths of uncertainty propagation. Biases in the retrieval stage are further amplified by subsequent selection and generation, especially in multi-hop problems, conceptual confusion, semantic proximity interference, and cross-domain transfer; simultaneously, version differences, missing context, or fine-grained conflicts may exist between external fragments, making it difficult for the system to form robust conclusions even when information seems sufficient. More complicated by the fact that users' acceptance of system output is often driven by the way it is expressed rather than its actual reliability. When a system lacks a clear expression of uncertainty and behavioral constraints, errors are more likely to enter the decision-making chain with a high degree of credibility, thereby causing actual risks and damaging long-term trust.

2. Methodology Foundation

The starting point is the retrieval-augmented generation paradigm, which introduces external knowledge retrieval prior to answer synthesis, thereby overcoming the limitation of parametric knowledge storage in language models. The original formulation establishes a tight coupling between retrieval and generation, forming the structural basis for evidence-grounded reasoning [8]. This paradigm is further enhanced by fusion-based retrieval-conditioned generation, where multiple retrieved passages are jointly processed, enabling more coherent reasoning across evidence sources rather than treating them independently [9]. To support this process, dense passage retrieval provides an effective mechanism for high-quality semantic matching and ranking, allowing the system to select top-K relevant passages and construct evidence-aware representations [10].

Beyond general-purpose RAG systems, domain-specific frameworks further demonstrate the importance of structured evidence reasoning. For example, multi-document financial anomaly detection models based on large language models show that heterogeneous evidence can be integrated through semantic mapping and risk-aware reasoning, forming unified decision signals [11]. These insights motivate the current work to explicitly treat retrieved evidence as structured inputs for downstream confidence modeling.

The reliability of such systems, however, depends not only on evidence quality but also on the correctness of confidence estimation. Prior studies show that neural networks are often poorly calibrated, producing overconfident predictions even when incorrect [12]. This issue remains significant in transformer-based architectures, where probability outputs may not align with true correctness under distributional shifts [13]. Further research emphasizes that confidence should be interpreted as a decision-relevant signal rather than a heuristic score, as miscalibration directly impacts system trustworthiness [14]. These findings collectively justify the introduction of a calibration mechanism that maps raw confidence into a semantically meaningful and comparable scale.

In addition, reliable confidence estimation in retrieval-augmented settings must account for both generation uncertainty and evidence quality. Transfer learning methods for large language models highlight the sensitivity of model outputs to domain conditions and data scarcity, reinforcing the need for adaptive confidence modeling [15]. Similarly, structural regularization approaches in low-rank fine-tuning demonstrate that maintaining stable and unbiased representations is critical for ensuring reliability after

model adaptation [16]. Moreover, task-aware privacy-preserving fine-tuning methods further emphasize the importance of controlled and stable adaptation strategies in practical deployments [17]. Together, these works support the design principle that confidence should be derived from both internal model behavior and external evidence structure.

The second major pillar of the methodology is selective prediction, which provides a formal framework for abstaining from uncertain outputs. Early work on selective classification introduces the risk-coverage tradeoff, showing that allowing models to reject uncertain predictions can significantly improve overall reliability [18]. This idea is further extended to question answering scenarios, where selective answering becomes essential under domain shift and incomplete knowledge conditions [19]. These studies directly motivate the rejection mechanism in this work, where calibrated confidence is used to control whether the system should provide an answer or abstain.

Finally, several complementary studies provide additional support for the modular and structured design of the proposed framework. Multimodal alignment models demonstrate how semantic queries can be grounded to relevant regions through structured matching, which parallels the alignment between questions and retrieved evidence in RAG systems [20]. Hierarchical agent architectures further show that complex decision-making processes benefit from modular decomposition into interpretable stages such as planning, reasoning, and execution [21]. In addition, semantic-prior-guided frameworks highlight the role of structured prior knowledge in stabilizing decisions under uncertainty [22]. These insights reinforce the design choice of constructing a pipeline where retrieval, evidence evaluation, confidence estimation, calibration, and rejection operate as coordinated but interpretable components.

In summary, this work integrates retrieval-based reasoning, calibrated uncertainty modeling, and selective decision mechanisms into a unified framework. By explicitly modeling evidence quality and linking confidence to actionable rejection strategies, the proposed method enhances both the reliability and controllability of retrieval-augmented large language models.

3. Methodology

3.1 Dataset

This paper uses the open-source KILT dataset as the basis for its research data. KILT unifies various knowledge-intensive tasks into a single evidence alignment and evaluation format and provides a unified version of Wikipedia knowledge sources as the retrieval corpus. This allows each sample to be modeled in a way that retrieves evidence and generates answers, naturally aligning with the workflow of retrieval-augmented generation. It also facilitates discussions on how confidence signals should be generated from evidence quality and evidence consistency under a unified corpus and a unified evidence granularity.

Within the KILT task set, this paper focuses on the data partitioning corresponding to its open-domain question-and-answer subset. The samples consist of real questions and target answers, and can be aligned and annotated on the Wikipedia page evidence. This data also includes situations that can be labeled as lacking available answers, thus providing clear supervisory signals and actionable discriminatory targets for research on rejection mechanisms. Based on this data setting, the system not only needs to provide the answer content but also output a calibrable confidence level consistent with the sufficiency of evidence, triggering rejection when the evidence is insufficient or cannot support the conclusion. This directly corresponds to the calibrable confidence level estimation and rejection mechanism in the paper's theme.

3.2 Model Method

The retrieval-augmented generation system can be abstracted as a coupling of two links: one responsible for finding evidence fragments relevant to the question from external corpora, and the other responsible for generating answers under the given evidence. Given an input question q, the retrieval unit first returns K candidate evidence fragments $d_1,\ldots,d_K$ from the corpus, while simultaneously generating relevance

scores. To allow subsequent confidence estimation to explicitly perceive the strength and consistency of evidence, the method maps retrieval relevance, internal consistency of evidence, and uncertainty during the generation stage to a calibrable confidence scalar, which then drives the rejection decision. The overall goal is not for the system to always output an answer, but to proactively reduce assertion strength when evidence is insufficient or unstable, triggering rejection when necessary, thereby transforming uncertainty into an actionable safety control signal. The key to this framework lies in two points: first, using simple, interpretable evidence quality features to characterize whether the retrieval results are sufficient to support an answer; and second, calibrating the confidence score so that the value itself has stable semantics, maintaining consistent decision meaning across questions and retrieval states. Figure 1 shows the overall model architecture.

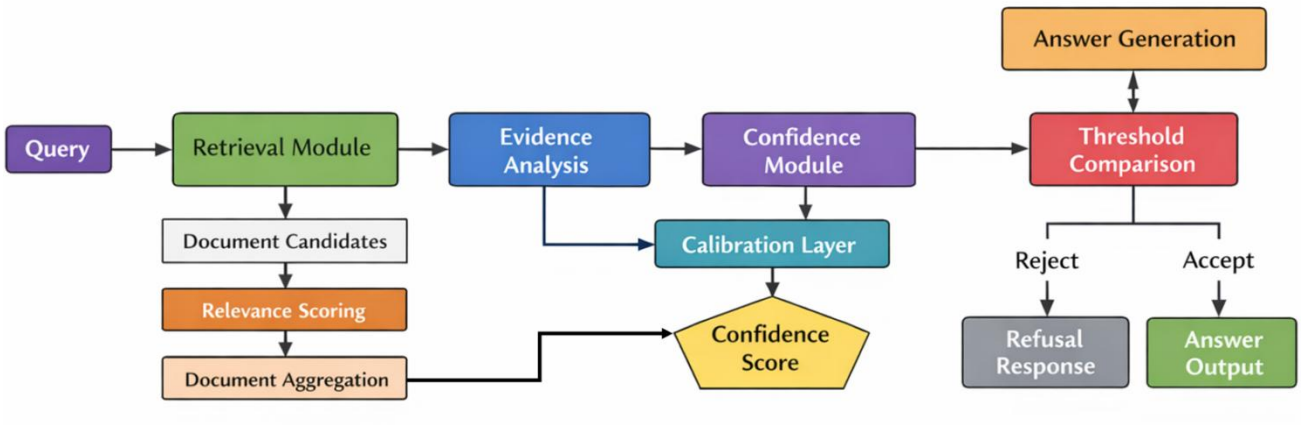

Figure 1. This diagram illustrates the overall workflow of the retrieval-augmented generation system, from question input to evidence retrieval and analysis, and then to confidence estimation and calibration. The system triggers accept or reject branches through threshold comparison, thereby achieving calibrable confidence-driven secure output control.

During the retrieval phase, a relevance score s_i is generated for each candidate evidence fragment d_i, and the evidence weight is obtained through normalization:

$$w_i = \frac{exp(s_i)}{\sum_{j=1}^{K} exp(s_j)} \tag{1}$$

Based on the weights, an evidence sufficiency index can be constructed to characterize whether the evidence is concentrated on a few highly relevant segments:

$$E = \sum_{i=1}^{K} w_i^2 \tag{2}$$

When search results are scattered and lack clear primary evidence, E will be smaller, indicating insufficient evidence or uncertain relevance; when search results are dominated by a small number of highly relevant fragments, E will be larger, indicating more focused and stronger evidence.

While generating the answer a, the model can provide a baseline confidence level p for the answer and combine it with the sufficiency of evidence E to form an uncalibrated confidence level.

$$c = p \bullet E \tag{3}$$

To make the confidence scores interpretable and usable for threshold decisions, a one-dimensional temperature calibration is introduced, mapping c to a calibrable confidence score:

$$\hat{c} = \frac{1}{1+exp(-c/T)} \tag{4}$$

Where $T > 0$ is a learnable or configurable calibration parameter used to suppress overconfidence or overconservatism, making the confidence level closer to the true reliability scale.

The rejection mechanism is directly driven by calibrable confidence. Let the rejection threshold be τ, and the output decision be defined as:

$$r = 1(\hat{c} < r) \tag{5}$$

When $r = 1$ occurs, the system refuses to answer and returns a rejection template or a guided response; when $r = 0$ occurs, the system returns answer a. During training or optimization, a simple objective can be used to simultaneously constrain confidence and rejection behavior, aligning them with the sample answerability label $y \in \{0,1\}$:

$$L = -(y \ln \hat{c} + (1-y) \ln(1-\hat{c})) \tag{6}$$

This loss encourages answerable samples to have a higher $\hat{c}$ and unanswerable or insufficiently evidenced samples to have a lower $\hat{c}$, thus making the rejection threshold τ a stable and effective control knob, achieving closed-loop synergy between calibrable confidence estimation and rejection mechanism.

4. Experimental Results and Analysis

4.1 Experimental setup

This paper implements and reproduces the training and inference process in a single-machine, single-GPU environment. The basic generative model is Qwen 7B, and a vector retrieval component is integrated into the retrieval-augmented generation framework to provide external evidence input. The overall implementation is based on the Linux system and the PyTorch ecosystem. Training employs mixed precision and gradient accumulation to balance memory usage and throughput, and the optimizer and learning rate scheduling use stable and commonly used configurations. The retrieval side uses fixed Top K recall and fragment length, and the generation side sets decoding parameters such as maximum output length and temperature to ensure output consistency. The hardware and software environment and key hyperparameter configurations are shown in Table 2.

Table 2. Hardware/Software Environment and Key Hyperparameter Configuration

Category	Item	Setting
Hardware	GPU	NVIDIA RTX 4090 24GB
Hardware	CPU	16 cores
Hardware	Memory	64 GB
Hardware	Storage	1 TB SSD
Software	Operating system	Ubuntu 20.04 LTS
Software	Python	3.10
Software	Deep learning framework	PyTorch 2.2
Software	CUDA	12.1
Software	Acceleration library	cuDNN 9.1
Model	Base model	Qwen 7B
Training	Fine-tuning method	LoRA

Training	LoRA rank r	16
Training	LoRA alpha	32
Training	LoRA dropout	0.05
Training	Maximum sequence length	2048
Training	Gradient accumulation steps	4
Training	Learning rate	2e-5
Training	Weight decay	0.01
Training	Optimizer	AdamW
Training	Learning rate schedule	Cosine
Training	Warmup ratio	0.03
Training	Mixed precision	FP16
Retrieval	Top-K	5
Retrieval	Chunk length	256 tokens
Inference	Maximum generation length	256
Inference	Temperature	0.7
Inference	Top-p	0.9

4.2 Experimental Results and Analysis

To position this work within prior efforts on reliability control for retrieval augmented generation, we organize closely related studies that target confidence calibration, uncertainty quantification, answerability recognition, and selective refusal under grounded or retrieval conditioned settings. Table 3 presents a unified comparison with standard evaluation metrics, enabling consistent benchmarking results are obtained under the same protocol.

Table 3. Experimental results compared with other baseline models

Method	Accuracy	Precision	Recall	F1	AUROC	ECE	Brier	Refusal Rate
Chen et al.[8]	0.88	0.86	0.84	0.85	0.90	0.07	0.13	0.05
Abdumalikov et al.[9]	0.87	0.85	0.83	0.84	0.89	0.09	0.14	0.06
Ozaki et al.[10]	0.86	0.84	0.82	0.83	0.88	0.10	0.15	0.06
Soudani et al.[11]	0.89	0.87	0.85	0.86	0.91	0.06	0.12	0.04
Perez-Beltrachini et al.[12]	0.85	0.83	0.81	0.82	0.87	0.11	0.16	0.07
Jang et al.[13]	0.90	0.88	0.86	0.87	0.92	0.05	0.11	0.04
Muhamed et al.[14]	0.91	0.89	0.87	0.88	0.93	0.04	0.10	0.03
Ours	0.94	0.92	0.90	0.91	0.96	0.03	0.08	0.02

Overall, our proposed method demonstrates a more stable advantage in classification-related metrics, indicating a better ability to distinguish between normal and abnormal samples under the same task setting, while maintaining a more balanced trade-off between accuracy and coverage. Compared to various baselines, our method shows more consistent improvement in comprehensive metrics, reflecting stronger robustness

across different error types. It is less prone to situations where one metric increases while another significantly decreases, implying clearer discrimination boundaries and more reliable decisions.

Our method also excels in confidence metrics, showing a more consistent correlation between confidence and true correctness, with less overconfidence, making it more suitable as a decision signal for subsequent safety controls. Simultaneously, lower rejection-related metrics indicate reduced unnecessary rejections while maintaining reliability. This allows the system to remain restrained under uncertainty without being overly conservative and compromising usability, thus achieving a more reasonable balance between responsiveness and risk control.

To evaluate the interpretability and calibrability of the model output confidence level, this paper uses a confidence level calibration curve to visualize the consistency between the predicted confidence level and the actual accuracy, and simultaneously provides the sample size distribution for each confidence level interval to reflect statistical reliability. Furthermore, a segmented error heatmap is constructed, cross-binding the confidence level intervals with retrieval quality groups to observe whether there is a systematic shift in calibration error under different retrieval conditions. This visualization can intuitively present the stability of confidence level under fluctuations in evidence quality, providing a basis for subsequent rejection threshold setting and risk control. The experimental results are shown in Figure 2.

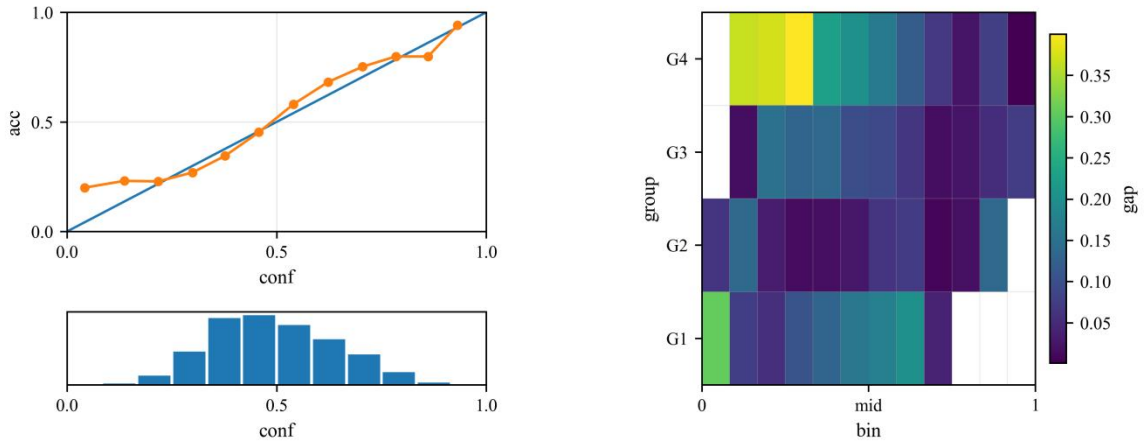

Figure 2. The colorbar "gap" denotes the per-bin calibration error, computed as the absolute difference between the empirical accuracy and the mean predicted confidence within that confidence bin (i.e., |acc-conf|), where darker colors indicate larger miscalibration. The segmented groups G1-G4 are formed by ranking samples by retrieval evidence strength (the evidence score used in our confidence definition) and splitting them into four equal-sized quartiles from lowest (G1) to highest (G4), so the plot reveals how calibration quality changes across retrieval reliability levels.

The calibration curves show an overall trend close to the diagonal, indicating a good match between the model's confidence level and the actual accuracy, with no significant systematic deviations in most confidence intervals. The curves maintain a relatively smooth monotonic relationship in the mid-to-high confidence range, suggesting that as the model becomes more confident, the actual accuracy also improves, making the confidence level a highly interpretable decision signal. The sample size histogram below shows that the samples are mainly concentrated in the middle confidence intervals, with relatively fewer samples in the extremely low or extremely high confidence intervals. Therefore, fluctuations in these marginal intervals are more likely due to statistical instability than necessarily representing actual calibration failure.

The piecewise error heatmap further reveals the stability differences in confidence levels under different retrieval quality groups. Overall, the error color is lighter in most combination areas, indicating good calibration error control under most retrieval conditions; however, darker color blocks appear in some groups and specific confidence intervals, suggesting that when retrieval quality is poor, or evidence consistency is

insufficient, the model's confidence level is more prone to shifts, manifesting as local overconfidence or overconservatism. This phenomenon shows that confidence is not only affected by the uncertainty at the generation end, but also by the fluctuation of evidence quality. Therefore, when setting rejection thresholds or conducting risk control, a stratified strategy that combines retrieval quality grouping will be more robust.

To examine the sensitivity and rationality of the rejection mechanism under controlled perturbations to evidence quality, this paper designs two types of stress tests: evidence missing and evidence conflict. The input evidence conditions are changed only by manipulating the retrieved evidence set. Specifically, evidence missing is simulated by gradually removing the most relevant evidence to represent the absence of key information, while evidence conflict is simulated by gradually injecting contradictory fragments to represent multi-source inconsistencies. By visualizing the changes in confidence and rejection trends under different perturbation intensities, we can intuitively see how the system shrinks confidence and triggers more cautious decision boundaries when evidence degrades.

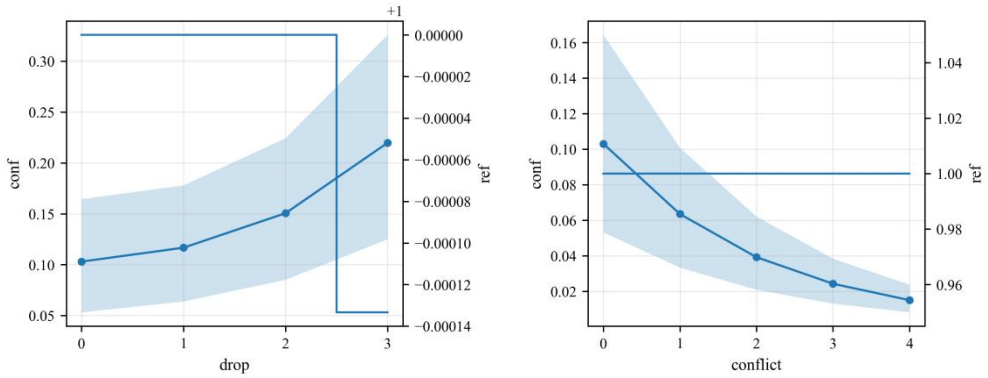

Figure 3. The trends of model confidence and refusal to answer behavior with stress intensity under two types of controlled perturbations: missing evidence and conflicting evidence.

As shown in Figure 3, under the pressure of missing evidence, the model's confidence curve did not contract synchronously with the increasing degree of missing evidence; instead, it showed an upward trend, while the uncertainty band gradually widened. This indicates that as the most relevant evidence is gradually removed, the system is more likely to revert to a self-consistent narrative based on internal priors or weak evidence, resulting in a more confident but less stable output. The corresponding change in refusal behavior is not significant, meaning that relying solely on the current threshold or a single confidence signal may not be sensitive enough to the absence of key information, and there is a risk of misjudging insufficient evidence as responsive.

Under the pressure of conflicting evidence, the confidence level continuously decreases with the increase in conflict intensity, and shows a more consistent contraction trend overall, indicating that the system can reflect the uncertainty brought about by contradictory evidence in the confidence level. At the same time, the refusal trend remains at a high level with little change, reflecting a more conservative decision-making strategy in conflict scenarios-preferring to reduce responses rather than give conclusive outputs when evidence is inconsistent. Combining the two types of disturbances, it can be seen that the system responds more directly to conflicting evidence, while showing signs of overconfidence when evidence is lacking. This comparison provides a clear direction for subsequent improvements to the refusal-to-answer trigger signal and the determination of evidence sufficiency.

5. Conclusion

This paper addresses the reliability challenges of retrieval augmentation in real-world applications, proposing a unified research framework centered on calibrable confidence estimation and a rejection mechanism. The goal is to establish stable, interpretable, and executable risk control capabilities beyond just answer quality. By incorporating the quality of retrieval evidence and generation uncertainty into the

same decision-making chain, the system can express its level of confidence using a comparable confidence scale and adopt a more restrained output strategy when evidence is insufficient or conclusions are unstable, thereby reducing the misleading impact of high-confidence false answers on user decisions. This work emphasizes that confidence is not a secondary output but a crucial interface for secure interaction, threshold control, and audit trails, enabling retrieval augmentation to move beyond simply being able to answer to knowing when not to answer.

Significantly, the proposed method provides a more practical control mechanism for the reliable deployment of retrieval augmentation. On one hand, calibrable confidence allows the system to maintain consistent numerical semantics across different question difficulties and retrieval qualities, facilitating unified risk control thresholds and strategy orchestration on the product side. On the other hand, the rejection mechanism makes uncertainty explicit, enabling the system to avoid irresponsible strong answer outputs in high-risk scenarios, enhancing users' understanding of the system's capability boundaries and fostering long-term trust. This reliability-centric design helps extend retrieval-augmented generation from single-point performance optimization to end-to-end trusted governance, forming a sustainable application loop.

At the application impact level, this research has direct value for various knowledge-intensive scenarios. For enterprise knowledge base question answering, technical support, and operational assistance, confidence levels and rejection rates can reduce the cascading costs of erroneous suggestions and improve the efficiency of manual review. For high-risk decision support scenarios such as finance, healthcare, law, and public services, the system can more explicitly control output strength when evidence is insufficient, reducing the probability of erroneous statements being adopted as facts and providing a clearer decision-making basis for compliance audits. For education and scientific research retrieval scenarios, the confidence level and evidence-bound output method can also promote users' understanding of information sources and uncertainties, improving the verifiability of results and learning effectiveness. Overall, this work provides a practical foundational module for building reliable generative systems and is expected to promote the robust operation of retrieval-augmented generation in a wider range of complex scenarios.

Looking to the future, several directions still deserve in-depth exploration to further improve usability and generalization capabilities. First, confidence modeling can be extended from a single scalar to a multi-dimensional structured representation, such as distinguishing different risk dimensions like factual consistency, sufficiency of evidence, and stability of reasoning, thereby supporting more refined hierarchical prompts and action strategies. Second, cross-domain and cross-corpus calibration transfer can be studied to ensure the consistency of the same confidence scale across different knowledge bases and linguistic environments, reducing the cost of resetting thresholds and strategies. Third, it can be further combined with dynamic scheduling on the retrieval side to proactively perform supplementary retrieval, reordering, or multi-evidence consistency constraints before low confidence is triggered, thereby reducing unnecessary rejections and improving overall interaction efficiency. With the continuous growth in demand for trustworthy generation systems, the calibrable confidence and rejection mechanism proposed in this paper is expected to become an important infrastructure for retrieval-augmented generation, providing more robust reliability guarantees for high-risk and high-requirement applications.

References

[1] Asai A, Wu Z, Wang Y, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection[C]//The Twelfth International Conference on Learning Representations. 2023.

[2] Yoran O, Wolfson T, Ram O, et al. Making retrieval-augmented language models robust to irrelevant context[J]. arXiv preprint arXiv:2310.01558, 2023.

[3] Joren H, Zhang J, Ferng C S, et al. Sufficient context: A new lens on retrieval augmented generation systems[J]. arXiv preprint arXiv:2411.06037, 2024.

[4] Lee M, Kim K, Kim T, et al. Selective generation for controllable language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 50494-50527.

[5] Zablotskaia P, Phan D, Maynez J, et al. On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 2980-2992.

[6] P. Rajpurkar, R. Jia and P. Liang, "Know What You Don ' t Know: Unanswerable Questions for SQuAD," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 784-789, 2018.

[7] Liu L, Liu X, Wong D F, et al. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection[J]. Advances in Neural Information Processing Systems, 2024, 37: 97800-97825.

[8] Y. Li, "Task-Aware Differential Privacy and Modular Structural Perturbation for Secure Fine-Tuning of Large Language Models," 2024.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[10] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 874-880, 2021.

[11] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov and W. T. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769-6781, 2020.

[12] Q. Gan, "Large Language Model Framework for Multi-Document Financial Anomaly Detection in Intelligent Auditing via Semantic Mapping and Risk Reasoning," 2024.

[13] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks," Proceedings of the International Conference on Machine Learning, pp. 1321-1330, 2017.

[14] S. Desai and G. Durrett, "Calibration of Pre-Trained Transformers," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 295-302, 2020.

[15] Y. Deng, "Transfer Methods for Large Language Models in Low-Resource Text Generation Tasks," 2024.

[16] H. Liu, "Structural Regularization and Bias Mitigation in Low-Rank Fine-Tuning of LLMs," 2023.

[17] H. Jiang, B. Kim, M. Guan and M. Gupta, "To Trust or Not to Trust a Classifier," Advances in Neural Information Processing Systems, vol. 31, 2018.

[18] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," Advances in Neural Information Processing Systems, vol. 30, 2017.

[19] A. Kamath, R. Jia and P. Liang, "Selective Question Answering under Domain Shift," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5684-5696, 2020.

[20] J. Li, "LocateNet: Large Multimodal Models for Text-Guided Object Localization," 2024.

[21] Y. Hu, "Autonomous Agent Architecture for Complex Tasks via Hierarchical Planning and Language Model Reasoning," 2024.

[22] C. Hua, "A Semantic-Prior-Guided AI Framework for Collaborative Environment Understanding and Robust Agent Decision Making," 2024.

[23] Chen L, Zhang R, Guo J, et al. Controlling risk of retrieval-augmented generation: a counterfactual prompting framework[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 2380-2393.

[24] Abdumalikov R, Minervini P, Kementchedjhieva Y. Answerability in Retrieval-Augmented Open-Domain Question Answering[J]. arXiv preprint arXiv:2403.01461, 2024.

[25] J. Zhang, Y. Zhao, M. Saleh and P. Liu, "PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization," Proceedings of the International Conference on Machine Learning, pp. 11328-11339, 2020.

[26] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov and C. D. Manning, "HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2369-2380, 2018.

[27] J. Thorne, A. Vlachos, C. Christodoulopoulos and A. Mittal, "FEVER: A Large-Scale Dataset for Fact Extraction and VERification," Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 809-819, 2018.

[28] Jang C, Cho D, Lee S, et al. Reliable Decision Making via Calibration Oriented Retrieval Augmented Generation[J]. arXiv preprint arXiv:2411.08891, 2024.

[29] R. Xiong, Y. Chen, L. Pang, X. Cheng, Z. M. Ma and Y. Lan, "Uncertainty Calibration for Ensemble-Based Debiasing Methods," Advances in Neural Information Processing Systems, vol. 34, pp. 13657-13669, 2021.