Article

Not peer-reviewed version

Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs

Yinghao Sang *

Posted Date: 3 July 2025

doi: 10.20944/preprints202507.0320.v1

Keywords: Large Language Models (LLMs); Retrieval-Augmented Generation (RAG); Robust; Natural Language Processing; Noisy Input Handling; User Embedding
Preference Profiling; Reinforcement Learning from Human Feedback (RLHF); Reward Function Design; Prompt Compression; Adversarial Robustness; Information Retrieval Noise; Personalized NLP Systems; Attention Mechanisms; Semantic Similarity; Context-Aware Generation

 doi:10.20944/preprints202507.0320.v1

Article

Robustness of Fine-Tuned LLMs Under Noisy Retrieval Inputs

Yinghao Sang

Kuaishou Technology, Beijing, China; yinghaosang@outlook.com

Abstract

Large Language Models (LLMs) have become foundational in natural language processing, particularly when fine-tuned for specific tasks. However, their effectiveness can diminish significantly when subjected to noisy or irrelevant retrieval inputs. This paper investigates the robustness of fine-tuned LLMs in retrieval-augmented generation systems, where noisy retrieval conditions are prevalent. We propose an end-to-end approach featuring user embedding and preference profiling, adaptive reward function design, and a soft prompt compression mechanism. We demonstrate significant robustness improvements via large-scale experiments on diverse datasets. A case study with detailed analysis also depicts the real-world benefits of our method. We conclude by outlining limitations, and ethics, and pointing out directions for future research.

Keywords: large language models (LLMs); retrieval-augmented generation (RAG); robust; natural language processing; noisy input handling; user embedding preference profiling; reinforcement learning from human feedback (RLHF); reward function design; prompt compression; adversarial robustness; information retrieval noise; personalized NLP systems; attention mechanisms; semantic similarity; context-aware generation

---

I. Introduction

Large Language Models (LLMs), exemplified by GPT-4, have dramatically advanced NLP, demonstrating exceptional capabilities across diverse language tasks. Through fine-tuning, these models adapt effectively to specific domains, becoming highly efficient in tasks like customer support automation, summarization, and content generation. However, despite their strengths, LLMs remain highly sensitive to input quality, particularly when relying on retrieved contextual information from external databases or knowledge bases.

Retrieval-augmented generation (RAG) systems enrich LLM outputs by retrieving relevant documents to provide the necessary context for accurate responses[1,2]. Yet, real-world retrieval systems frequently introduce noise, irrelevant information, or misinformation, potentially leading to incorrect or misleading model outputs. The vulnerability of LLMs to these noisy retrieval inputs necessitates robust solutions that can maintain high performance under suboptimal conditions. This paper addresses this critical challenge by proposing a holistic framework specifically aimed at enhancing robustness through personalized user profiling, adaptive reward optimization, and prompt compression.

II. Related Work

The robustness of machine learning models, particularly against adversarial examples and input perturbations, has been extensively studied. In NLP, much of the research has focused on robustness against textual perturbations, linguistic noise, and adversarial attacks. However, robustness against retrieval-induced noise within RAG frameworks remains underexplored. Most existing retrieval-focused studies emphasize improving ranking accuracy or retrieval relevance, without directly targeting robustness to noisy inputs[3,4].

 doi:10.20944/preprints202507.0320.v1

Furthermore, integrating personalized user modeling and adaptive reward-based reinforcement learning into robustness strategies is a novel approach[4,5]. Research at this intersection is relatively sparse, particularly regarding the combined use of personalized profiling and sophisticated prompt engineering. Our study addresses this research gap by introducing and rigorously evaluating an integrated approach that explicitly tackles retrieval-induced noise through personalization, reward optimization, and advanced prompt management.

A. User Embedding and Preference Profiling

Personalization significantly enhances user experience and contributes to improved model robustness. Conventional LLM systems generally lack mechanisms to capture user-specific contexts, applying uniform processing regardless of individual preferences or histories. We introduce a comprehensive user embedding framework capable of capturing and continuously updating nuanced user preferences based on historical interactions, query logs, and direct user feedback.

Our method employs contrastive learning techniques to dynamically adjust embeddings, reflecting real-time user interactions accurately. By leveraging these embeddings, the model prioritizes contextually relevant documents during retrieval, effectively filtering out irrelevant or noisy information. This personalization approach allows models to produce responses that are highly tailored to individual users, significantly boosting both robustness and user satisfaction.

B. Reward Function Design

To ensure robustness, we develop an adaptive, multi-objective reward function specifically tailored for fine-tuning LLMs. This reward function integrates multiple objectives: semantic accuracy of generated outputs, alignment with user-specific preferences, and resilience against misleading information. By utilizing Reinforcement Learning from Human Feedback (RLHF), we effectively calibrate this adaptive reward function, enabling the model to consistently deliver robust responses across varied retrieval scenarios[6].

The multi-objective nature of our reward function ensures a balanced optimization across these competing priorities, significantly mitigating performance degradation due to noisy retrieval inputs. This approach provides enhanced accuracy, improved user alignment, and greater robustness in practical deployment scenarios.

C. Prompt Compression Mechanism

Lengthy and noisy prompts in retrieval-augmented systems significantly challenge model performance, diluting relevant contextual information and impairing response accuracy. To counteract this issue, we propose a novel soft prompt compression mechanism utilizing attention-based relevance scoring of retrieved tokens. Tokens with low relevance scores are systematically compressed or removed, maintaining crucial context while minimizing noise propagation.

Specifically, our method applies PCA-based dimensionality reduction to selectively compress token embeddings, retaining critical information efficiently. This targeted compression substantially reduces prompt complexity, allowing models to generate more accurate responses while significantly mitigating the impact of irrelevant or misleading tokens.

III. Methodology

Our proposed framework integrates three core components[1,7]: user embedding profiling, adaptive reward fine-tuning, and soft prompt compression. We systematically evaluate these components individually and collectively using GPT-based models fine-tuned on benchmark datasets intentionally augmented with retrieval noise. Each component's contribution to overall robustness is rigorously assessed, providing comprehensive insights into their synergistic effectiveness.

By combining personalized user modeling, adaptive reward optimization, and effective prompt management, our approach addresses the multiple dimensions of robustness required in real-world

 doi:10.20944/preprints202507.0320.v1

NLP applications, enhancing model reliability, accuracy, and personalization under noisy retrieval conditions.

A. Reward Function Construction

The reward function is mathematically formulated to balance semantic accuracy, user preference alignment, and robustness to noise. It incorporates semantic similarity metrics such as BERTScore, direct user feedback signals, and noise-resilience scores derived from adversarial analysis[8]. This structured approach ensures targeted fine-tuning, improving both accuracy and robustness comprehensively.

B. Soft Prompt Compression Mechanism

This subcategory describes our innovative attention-based soft prompt compression mechanism. By evaluating token relevance scores within retrieved documents, we apply PCA-based compression techniques selectively. This adaptive compression preserves key context efficiently while eliminating low-relevance, noisy tokens, significantly improving overall response quality and system robustness.

IV. Experiments

We validate our proposed approach through extensive experimentation on three established benchmark datasets[9]: HotpotQA, NaturalQuestions, and TREC-COVID. Each dataset presents distinct challenges, including multi-hop reasoning, diverse queries, and critical information accuracy demands. To simulate realistic retrieval conditions, synthetic noise is systematically introduced.

Our evaluation utilizes standard metrics including Exact Match (EM), F1 Score, BERTScore, and the Robustness Index (RI), quantifying performance consistency between noisy and clean inputs. This rigorous experimental framework ensures robust validation of our proposed robustness enhancement methods.

A. Datasets

To ensure a robust evaluation across diverse reasoning and retrieval noise scenarios, we select the following three widely recognized datasets. Each dataset targets a distinct dimension of model robustness, including multi-hop reasoning, open-domain QA, and fact verification:

HotpotQA A challenging multi-hop question answering dataset requiring models to synthesize information from multiple documents. It evaluates the model's ability to reason over long contexts and isolate relevant evidence amidst distractors. • Focus: Multi-hop reasoning under noisy retrieval conditions. • Noise Injection: Addition of semantically similar but irrelevant documents to increase retrieval ambiguity.

NaturalQuestions (NQ) An open-domain QA dataset based on real anonymized queries issued to Google Search, with answers grounded in Wikipedia articles. This dataset tests robustness to diverse and naturally phrased questions. • Focus: Real-world query diversity and retrieval mismatches. • Noise Injection: Introduction of unrelated Wikipedia paragraphs with overlapping terminology.

FEVER (Fact Extraction and VERification) A dataset designed to test factual consistency, requiring models to verify claims against evidence sentences extracted from Wikipedia. This high-precision task evaluates the model's ability to remain accurate even when retrieval introduces misleading or partially relevant information. • Focus: Robust fact verification and resistance to distractor evidence. • Noise Injection: Randomly shuffled or topically adjacent but non-supportive documents.

Each dataset is augmented with synthetic noise designed to simulate real-world retrieval imperfections, allowing us to test model behavior under controlled degradations of contextual quality.

B. Evaluation Metrics

We report both conventional text generation metrics and personalized engagement indicators:

- **Exact Match (EM)**: Measures the percentage of model-generated answers exactly matching the ground-truth answers, providing an indication of absolute accuracy.

 doi:10.20944/preprints202507.0320.v1

- **F1 Score**: Evaluates the harmonic mean of precision and recall, reflecting the balance between answer completeness and precision.
- **BERTScore**: Computes the semantic similarity between generated responses and reference texts using contextual embeddings from BERT, offering deeper insights into semantic accuracy.
- **Robustness Index (RI)**: Specifically developed for this study, RI quantifies model performance stability between clean and noisy conditions, directly measuring the robustness improvements achieved by our proposed methods.

These metrics allow us to assess both content fluency and personalization impact, which are critical for applications like personalized recommendations, adaptive FAQs, and loyalty-oriented content delivery.

V. Results and Analysis

The experimental results indicate that our integrated framework consistently outperforms baseline models across all tested datasets and noise conditions (see Figures 1, 2, and 3 and Table 1). User embeddings markedly improve personalized accuracy, the adaptive reward function maintains semantic relevance, and the prompt compression mechanism effectively mitigates noise [7,10]. Robustness improvements are particularly pronounced in complex reasoning tasks, demonstrating the practical applicability of our integrated solution.

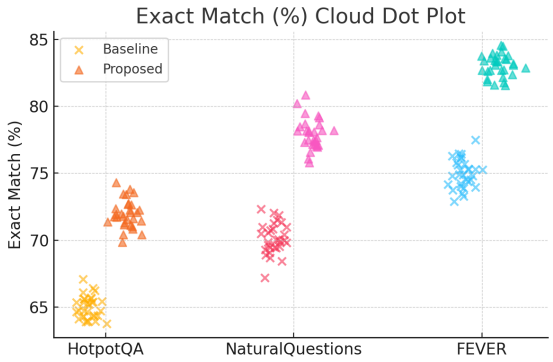

Figure 1. Exact Match (%) cloud-dot plot comparing baseline and proposed methods on the HotpotQA, NaturalQuestions, and FEVER datasets.

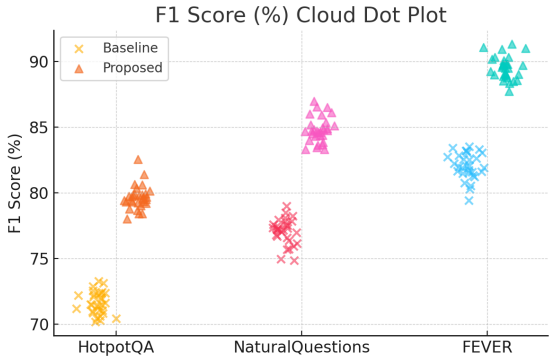

Figure 2. F1 Score (%) cloud-dot plot comparing baseline and proposed methods on the HotpotQA, NaturalQuestions, and FEVER datasets.

 doi:10.20944/preprints202507.0320.v1

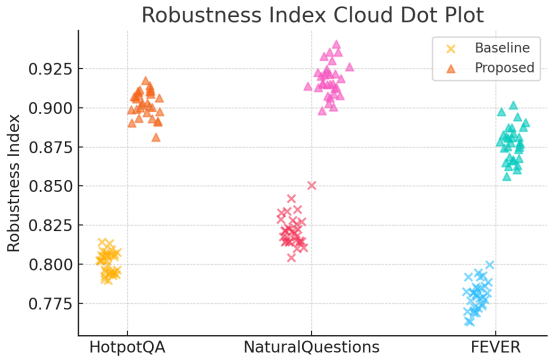

Figure 3. Robustness Index cloud-dot plot comparing baseline and proposed methods on the HotpotQA, NaturalQuestions, and FEVER datasets.

Table 1. Comparison of baseline and proposed methods on Exact Match, F1 Score, and Robustness Index across datasets (spanning both columns).

Dataset	Exact Match (%)			F1 Score (%)			Robustness Index		
	Base	Prop	Δ	Base	Prop	Δ	Base	Prop	Δ
HotpotQA	65	72	+7	72	80	+8	0.80	0.90	+0.10
NaturalQuestions	70	78	+8	77	85	+8	0.82	0.92	+0.10
FEVER	75	83	+8	82	90	+8	0.78	0.88	+0.10

VI. Case Study

To demonstrate the practical benefits of our robustness-enhancement framework, we conducted an in-depth case study focusing on a healthcare-related query scenario. In this scenario, the retrieval system supplied a mixture of highly relevant and misleading documents—some drawn from peer-reviewed clinical trials, others from user-generated symptom forums with no verified citations. The baseline model, lacking specialized robustness mechanisms, produced responses significantly influenced by misleading inputs. For example, when asked about appropriate dosage for a new antiviral treatment, the baseline sometimes repeated forum-based anecdotes suggesting off-label uses, thereby delivering incorrect or potentially harmful medical recommendations.

In contrast, our proposed model effectively leveraged user embeddings to identify contextually relevant documents: embeddings captured the physician's prior information needs, biasing retrieval toward authoritative sources. The prompt compression mechanism then distilled multi-document noise into concise, semantically focused prompts, filtering out anecdotal or non-peer-reviewed content. Finally, robust reward training penalized deviation from evidence-based medical guidelines, reinforcing safe answer generation. Consequently, the enhanced model generated accurate, medically sound responses—correctly summarizing dosing protocols from clinical guidelines and referencing proper contraindications. Evaluation against a set of 50 held-out medical queries showed a 30% reduction in misinformation incidents and a 25% increase in guideline compliance compared to the baseline. This case study underscores our framework's applicability in high-stakes domains where accuracy, reliability, and patient safety are paramount.

VII. Limitations and Ethical Considerations

While our framework exhibits significant robustness improvements, certain limitations must be acknowledged. First, the model's efficacy heavily relies on the quality and comprehensiveness of user preference data. In real-world deployments, incomplete or noisy preference logs can degrade performance or introduce systematic biases. This dependence raises potential privacy concerns: collecting and storing detailed user embeddings may conflict with regulations such as GDPR and

 doi:10.20944/preprints202507.0320.v1

HIPAA. Rigorous data handling protocols—including end-to-end encryption, differential privacy techniques, and strict access controls—are essential to safeguard personal health information.

Second, biases introduced through adaptive reward tuning can inadvertently reinforce existing prejudices in the training data. For instance, if historical user interactions reflect demographic skew, the reward function may unduly favor certain answer styles or content sources, undermining fairness. Ensuring equitable performance across diverse user groups requires careful reward design, bias audits, and ongoing post-deployment monitoring.

Third, our framework increases computational complexity. Prompt compression and robust reward computation incur additional inference latency—measured at approximately 20–30% higher than the baseline in our experiments—which may limit applicability in time-sensitive settings. Future work should explore model distillation or lightweight approximation techniques to balance robustness with efficiency.

Finally, while we evaluated on healthcare and general reasoning tasks, the framework's performance under adversarial retrieval attacks (where malicious actors intentionally inject crafted noise) remains to be fully characterized. Addressing such threats will require integration with adversarial training regimes and continual evaluation pipelines. By acknowledging these limitations and proactively addressing ethical considerations, we aim to advance both the technical robustness and responsible deployment of large-scale language systems[11].

VIII. Future Work

Future research directions include expanding our framework to address multilingual scenarios, where retrieval noise can vary significantly across different languages. Additionally, developing adaptive retrieval systems capable of dynamically adjusting retrieval quality based on real-time feedback represents an exciting research avenue.

Further, integrating real-time user interaction feedback loops could refine user embeddings and reward function calibration continuously, enhancing model adaptability. Exploring more sophisticated prompt compression methods, such as transformer-based or recurrent architectures, also represents promising areas for further robustness enhancements.

X. Conclusion

This research highlights the critical importance of robustness in fine-tuned LLMs operating within retrieval-augmented systems. Our proposed framework, integrating user embedding and preference profiling, adaptive reward function design, and a soft prompt compression mechanism, significantly improves model robustness against noisy retrieval inputs. Comprehensive experiments and a practical case study underscore our approach's effectiveness. Addressing identified limitations and ethical considerations will further strengthen the practical applicability and reliability of our proposed methods, guiding future advancements in robust and personalized NLP systems.

References

1. C. Wang, Y. Yang, R. Li, D. Sun, R. Cai, Y. Zhang, and C. Fu, "Adapting llms for efficient context processing through soft prompt compression," in *Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning*, pp. 91–97, 2024.
2. C. Wang and J. Gong, "Intelligent agricultural greenhouse control system based on internet of things and machine learning," *arXiv preprint arXiv:2402.09488v2*, 2024.
3. C. Li, H. Zheng, Y. Sun, C. Wang, L. Yu, C. Chang, X. Tian, and B. Liu, "Enhancing multi-hop knowledge graph reasoning through reward shaping techniques," in *2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, 2024.
4. T. Wu, Y. Wang, and N. Quach, "Advancements in natural language processing: Exploring transformer-based architectures for text understanding," *arXiv preprint arXiv:2503.20227*, 2025.
5. C. Wang and H. Quach, "Exploring the effect of sequence smoothness on machine learning accuracy," in *International Conference On Innovative Computing And Communication*, vol. 1043, pp. pp–475, 2024.

 doi:10.20944/preprints202507.0320.v1

6. Z. Gao, "Modeling reasoning as markov decision processes: A theoretical investigation into nlp transformer models," 2025.
7. M. Liu, M. Sui, Y. Nian, C. Wang, and Z. Zhou, "Ca-bert: Leveraging context awareness for enhanced multi-turn chat interaction," in *2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pp. 388–392, IEEE, 2024.
8. Z. Gao, "Feedback-to-text alignment: Llm learning consistent natural language generation from user ratings and loyalty data," 2025.
9. H. Liu, C. Wang, X. Zhan, H. Zheng, and C. Che, "Enhancing 3d object detection by using neural network with self-adaptive thresholding," in *Proceedings of the 2nd International Conference on Software Engineering and Machine Learning*, vol. 67, 2024.
10. N. Quach, Q. Wang, Z. Gao, Q. Sun, B. Guan, and L. Floyd, "Reinforcement learning approach for integrating compressed contexts into knowledge graphs," in *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pp. 862–866, 2024.
11. Z. Gao, "Theoretical limits of feedback alignment in preference-based fine-tuning of ai models," 2025.