Research Proposal: A Reinforcement Learning Framework for Verifiable, Citation-Driven Responses in LLM Customer Service

Abstract: To mitigate hallucinations in enterprise LLM customer service, this proposal designs a reinforcement learning (RL) framework that optimizes for verifiable answering. The core idea is to learn a behavioral policy that appropriately cites evidence, asks clarifying questions when retrieval is incomplete, and refuses when answers cannot be supported. This is achieved by formalizing verifiability as a multi-component reward signal and training with preference-based optimization. The expected outcome is a more reliable, auditable, and user-trustworthy system suitable for high-stakes domains.

Keywords: hallucination reduction, reinforcement learning, verifiability, citation accuracy, customer service systems

1. Research Questions and Objectives

The core research question is: How can reinforcement learning be used to train LLMs in customer service to dynamically and appropriately generate verifiable responses—by accurately citing existing evidence, proactively seeking clarification, or refusing—thereby minimizing hallucinations? The primary objectives are to: (1) formally define verifiability and hallucination in this interactive setting; (2) design an RL policy that balances answer faithfulness, utility, and operational efficiency; and (3) develop and evaluate interactive strategies for handling evidence uncertainty.

2. Literature Review

Retrieval-augmented generation (RAG) improves response grounding in external knowledge but fails to inherently prevent evidence misuse or citation inconsistency,

with extensions like Self-RAG introducing self-reflection mechanisms to enhance factuality [1]. Reinforcement learning from human feedback (RLHF) aligns models with human preferences but prioritizes fluency and helpfulness over explicit verifiability, while related work on safe RLHF explores refusal mechanisms for high-risk scenarios [2]. A comprehensive survey of hallucination in LLMs notes that prior mitigation efforts often focus on post-hoc detection or static benchmarks rather than proactive, policy-based prevention in dynamic dialogues [3]. This leaves a critical gap: the lack of integrated RL frameworks that learn behavioral policies governing citation, clarification, and refusal within coherent, multi-turn customer service interactions — especially under realistic constraints of evidence quality and query ambiguity.

3. Methodology

Hallucination is operationally defined as any generated claim that contradicts provided evidence, lacks a supporting citation, or references a source incorrectly [3]. Key measurable variables include retrieval quality, citation accuracy, answer faithfulness (via human evaluation), clarification/refusal rate, and user satisfaction.

The proposed RL framework treats the dialogue as a sequential decision-making process. The state encompasses the user query, retrieved evidence with confidence scores, and dialogue history. The action space allows the model to: generate an answer with citations, ask a specific clarifying question, or politely refuse. The reward function is a weighted sum of components reflecting citation faithfulness, user satisfaction (from feedback), and negative costs for escalations or excessive handling time. A reward model trained on human preferences for cautious, evidence-backed responses will provide this signal. The policy will be trained using Proximal Policy Optimization (PPO), initialized from a supervised fine-tuned baseline.

For data, public customer service dialogues (e.g., Banking77) will be augmented

with synthetic evidence passages and annotated for verifiability. Ambiguous and low-coverage query scenarios will be deliberately introduced to train and test the interactive policy.

4. Expected Outcomes and Significance

We expect the resulting model to demonstrate a significant reduction in hallucination rates compared to standard RAG and RLHF baselines, while maintaining comparable user satisfaction. The model should show increased and appropriate use of clarification and refusal actions in uncertain contexts, leading to more traceable and auditable interactions.

The research holds both theoretical and practical significance. Theoretically, it advances methods for embedding hard safety and verifiability constraints into RL-based dialogue policy learning. Practically, it provides a blueprint for developing more reliable and compliant LLM assistants for high-risk enterprise domains, directly enhancing operational trust and reducing regulatory risk.

5. References

[1] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[2] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[3] Zhang, Y., Li, P., Hu, L., Peng, X., Liu, G., & Zhou, J. T. (2023). A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922.