COSE474-2024F: Final Project Proposal
Optimizing Domain Knowledge Response Accuracy Through Contextual Memory in Conversational AI

2021170220 / Kang Seok Jeon

1. Introduction

The demand for conversational AI has surged in recent years, driven by the increasing need for AI-driven customer service across various industries. In 2024, the global conversational AI market reached approximately $1.386 billion, with a projected CAGR of 21.95%, indicating widespread adoption in sectors like healthcare, finance, and information technology.

Despite their strengths in general tasks, conversational AI models often struggle with domain-specific queries. Systems like ChatGPT tend to produce responses lacking the depth needed for specialized topics, as they rely on generalized datasets that miss domain-specific nuances.

This research explores strategies for enhancing response accuracy by incorporating domain-specific knowledge and structuring prompts effectively. The goal is to enable AI systems to deliver more reliable answers in professional settings.

2. Problem Definition & Challenges

This research addresses the challenge of conversational AI models delivering overly general answers to domain-specific questions. Models like ChatGPT often lack the specialized nuance needed for accurate responses due to their reliance on general training data.

The main focus is to explore how session-based interactions—through targeted prompts and pre-loaded context—can improve the precision of these models in handling specialized inquiries. This study aims to identify effective prompt strategies and types of information that can enhance domain-specific response accuracy.

3. Related Works

Research on enhancing domain-specific question answering has introduced methods like RAFT, Blended RAG, and DSQA-LLM. RAFT combines fine-tuning with retrieval to dynamically access domain-specific documents for accurate responses in specialized fields. Blended RAG uses semantic search and hybrid queries to improve document retrieval, while DSQA-LLM integrates structured retrieval for precise information. Together, these methods demonstrate how retrieval and fine-tuning can improve conversational AI accuracy in specialized domains.

4. Datasets

This study will use a mix of public benchmark datasets and custom data from domain-specific resources. For general evaluation, datasets like MedQA (healthcare), Stack Overflow (IT/software), and Legal Case Reports (law) will help assess accuracy across different fields. These datasets include real-world inquiries. Additionally, a small, curated dataset from domain-specific texts may be created to further validate the model's performance on specialized content. This selection is subject to change as the study progresses and more suitable datasets are identified.

5. State-of-the-Art Methods and Baselines

The research explores methods like RAFT and semantic search to enhance accuracy using domain-relevant documents. ChatGPT will serve as the baseline for domain-specific response accuracy. Comparisons will include RAFT, semantic search, and newly designed prompt and training strategies, evaluated on datasets such as MedQA and Stack Overflow to assess the impact of prompt structuring and context retention on response quality.

6. Schedule

- **Weeks 1-2**: Conduct literature review and prepare domain-specific datasets.
- **Week 3**: Design and test initial prompt strategies with pre-loaded contextual information.
- **Week 4**: Refine prompts based on initial results and conduct further tests.
- **Week 5**: Evaluate accuracy improvements with baseline comparisons.
- **Week 6**: Finalize analysis and complete the report.