# CSC-306 Natural Language Processing Project 3 Report

**Kurtik Appadoo**

## Abstract

This research explores various strategies to enhance the effectiveness of Large Language Models (LLMs) in conducting constructive argumentative dialogues. I present a series of increasingly sophisticated dialogue agents, built upon a Retrieval-Augmented Generation (RAG) architecture, that aim to broaden users' perspectives rather than simply win arguments.

## 1 Introduction

Recent advances in Large Language Models (LLMs) have opened new possibilities in argumentative dialogue systems. While current systems excel at debate and persuasion, they often focus on winning arguments rather than fostering understanding. The NLP community has increasingly recognized the need for dialogue systems that can facilitate genuine perspective-broadening rather than mere stance change. This research addresses this gap by developing more sophisticated dialogue agents that prioritize mind-opening discussions over persuasion.

Research Motivation: The primary challenge in argumentative dialogue systems lies not in generating coherent responses, but in creating agents that can effectively broaden perspectives while maintaining engagement. Previous work has shown that Retrieval-Augmented Generation (RAG) architectures provide a strong foundation for knowledge-based dialogue. However, standard RAG implementations lack the nuanced understanding and strategic approach needed for mind-opening discussions. my research question explores: How can I enhance RAG-based dialogue agents to better facilitate perspective-broadening conversations?

This work makes several contributions to the field:

- Development of my distinct enhancements to the RAG architecture: - Structured dialogue framework through enhanced prompt engineering - Chain of Thought reasoning for deeper conversation analysis - Few-shot learning approach for query reformulation
- Implementation of a comprehensive evaluation framework for assessing dialogue effectiveness
- Evidence demonstrating the benefits of combining multiple dialogue enhancement strategies
- A novel approach to topic-aware retrieval that adapts to different conversation contexts

The remainder of this paper is organized as follows: First, I present the baseline RAG implementation and its limitations. Then, I detail each of my enhancement approaches, providing theoretical justification and implementation details. Next, I describe my evaluation methodology using simulated characters and the Kialo debate database. Finally, I present results demonstrating how these enhancements, may or may not have lead to more effective mind-opening discussions. This research contributes to the broader goal of developing LLM systems that can facilitate meaningful dialogue across differing viewpoints and build mutual understanding.

## 2 Background/ Related Work

Recent research in argumentative dialogue systems has undergone a significant shift in focus, moving away from traditional persuasion goals toward fostering genuine open-mindedness. Farag et al. (2022) pioneered this direction with their work "Opening up Minds with Argumentative Dialogues," which introduced a novel approach focused on helping participants understand and appreciate opposing viewpoints rather than changing their stance.

Their study utilized human wizards in a Wizard-of-Oz (WoZ) setup to conduct 183 structured dialogues on controversial topics, with arguments sourced from the Kialo debate platform. This approach demonstrated that well-structured argumentative dialogue could increase participants' receptiveness to opposing views and their recognition of valid reasoning in conflicting positions.

While Farag et al.'s work relied on human wizards and evaluated two basic models (Wikipedia-based retrieval and argument-based fine-tuning), my implementation takes several significant steps forward:

- Fully Automated Approach: Instead of human wizards, I develop autonomous LLM-based agents capable of conducting mind-opening dialogues independently.

- Enhanced Retrieval Mechanisms: I improve upon their basic retrieval model by implementing topic-aware RAG (Retrieval-Augmented Generation) that adapts its strategy based on conversation context.

- Advanced Reasoning Techniques: Where their models used straightforward retrieval or fine-tuning, I incorporate multiple sophisticated approaches: - Chain of Thought reasoning for deeper conversation analysis - Few-shot learning for better query reformulation - Prompt Engineering for better structure to dialogue

- Evaluation Framework: While their study measured success through pre/post conversation surveys, I implement a model-based evaluation system that provides immediate feedback on dialogue quality and effectiveness.

This research builds upon Farag et al.'s foundation while addressing several limitations in their implementation. By combining their insights about effective mind-opening dialogue with advanced LLM capabilities, I aim to create more sophisticated and effective automated dialogue agents.

## 3 Approach

my research explored several dialogue agents with increasing levels of sophistication, starting from simple baseline models and progressing to more complex implementations:

- Baseline Models: - Airhead: A basic ConstantAgent that responds with the same phrase regardless of input, serving as my simplest baseline to validate evaluation metrics. - Alice: An LLM-based agent using basic prompt engineering, instructed to challenge user perspectives and encourage broader thinking through simple two-sentence responses. - Akiko: A retrieval-only agent that uses the Kialo database without LLM capabilities, demonstrating the value of structured argument retrieval.

- Advanced Baseline - Aragorn: Aragorn represents my first significant improvement, implementing a Retrieval-Augmented Generation (RAG) architecture that combines LLM capabilities with structured knowledge retrieval. Its approach includes: - Systematic retrieval of relevant claims from the Kialo database - Balanced presentation of supporting and counter arguments - Integration of retrieved context into LLM prompts Initial evaluations showed Aragorn outperforming Alice (23.3 vs 22.5 total score), establishing it as my primary baseline for further improvements.

- Enhanced RAG Implementation - Victor: Victor is the culmination of my knowledge from previous techniques. Victor uses some strategies form different models like chain of thought, prompt-engineering and even the baseline Aragorn. Following Aragorn's impressive evaluation metrics compared to Alice, I decided to make Victor a subclass of Aragorn's RAG implimentation and built upon that as it's super class.

## 4 Experimental Design and Model Evolution

### 4.1 Baseline Models

my experimental design began with three baseline models to establish foundational comparisons. The first, Airhead, implements the simplest possible response mechanism using constant outputs, serving as an absolute baseline for performance metrics and establishing minimum acceptable performance thresholds.

Alice represents my second baseline, utilizing a pure LLM-based approach with basic prompting. Without additional context or knowledge integration, Alice focuses on generating thoughtful but

generic responses, demonstrating the capabilities and limitations of pure language models in argumentative dialogue.

my third baseline, Akiko, takes a pure knowledge-based approach utilizing the Kialo database. Operating without LLM integration, Akiko implements random selection of counter-arguments from similar claims, demonstrating the effectiveness of pure knowledge-based responses and the potential value of structured argument databases.

### 4.2 Advanced Model Development

#### 4.2.1 RAGAgent (Aragorn)

my first advanced model, Aragorn, represents a significant step forward by combining LLM capabilities with the Kialo knowledge base. The model implements a sophisticated context retrieval system that extracts relevant claims from the database while balancing supporting and opposing arguments. By limiting context to the most relevant pieces and utilizing dynamic system prompt modification, Aragorn maintains a clean separation between retrieval and generation components while leveraging the strengths of both approaches.

#### 4.2.2 ChainOfThoughtAgent

Building upon Aragorn's architecture, the ChainOfThoughtAgent introduces explicit reasoning steps through private thoughts. The agent performs a comprehensive four-step analysis, assessing emotional states, identifying hidden concerns, developing comfort strategies, and planning perspective broadening approaches. This metacognitive layer enables more nuanced and strategic responses while maintaining the benefits of the RAG architecture.

#### 4.2.3 FewShotLearningAgent

The FewShotLearningAgent enhances the RAG approach through improved query capabilities. By implementing query reformulation through few-shot prompting, this agent transforms casual dialogue into formal claims using curated examples. This innovation significantly improves the relevance of retrieved context and enables more precise matching between user statements and the knowledge base.

#### 4.2.4 StructuredDialogueAgent

Through the implementation of the LAEB (Listen, Acknowledge, Explore, Bridge) framework, the StructuredDialogueAgent brings sophisticated prompt engineering to my dialogue system. The agent emphasizes emotional safety while maintaining intellectual engagement, combining evidence-based discussion with effective bridge-building between perspectives. This structured approach helps ensure consistent, high-quality interactions across various topics and conversation partners.

#### 4.2.5 Victor

Victor represents the culmination of my research, streamlining the combination of RAG and chain-of-thought approaches. The implementation simplifies the thought process while maintaining robust context usage, achieving a more efficient balance between knowledge retrieval and strategic thinking. Victor's architecture reduces computational overhead while preserving context relevance, representing an optimal trade-off between sophistication and practicality.

### 4.3 Progressive Approach to Victor

my iterative development journey progressed systematically from basic models (Airhead, Alice, Akiko) to increasingly sophisticated implementations, with each step building on previous learnings. Starting with Aragorn's integration of LLM and knowledge capabilities, I enhanced the system through ChainOfThought's reasoning abilities, FewShotLearning's query improvements, and StructuredDialogue's framework testing, ultimately culminating in Victor, a model that synthesizes the most effective elements from each iteration while maintaining an optimal balance between sophistication and computational efficiency in argumentative dialogue systems.

## 5 Results

my evaluation of argubot implementations with varying levels of sophistication revealed a clear performance progression corresponding to model complexity. As shown in Tables 1-2, the baseline implementations established initial performance benchmarks, with Airhead (a constant response agent) scoring lowest at 11.5 total points across my evaluation dimensions. This simplistic agent, which responds with the same message regardless of context, demonstrated particularly poor performance in engagement (1.8/5) and skill (2.1/5). Moving up in complexity, Akiko (a retrieval-based agent drawing from Kialo without LLM integration) achieved 19.0 points, showing moderate improvements in

| Model | Engaged | Informed | Intelligent |
|---|---|---|---|
| Akiko | 3.1 | 3.0 | 3.2 |
| Airhead | 1.8 | 2.0 | 2.6 |
| Alice | 3.7 | 3.2 | 3.4 |

Table 1: Evaluation scores for additional models on engagement, informativeness, and intelligence.

| Model | Moral | Skilled | **TOTAL** |
|---|---|---|---|
| Akiko | 3.1 | 6.6 | 19.0 |
| Airhead | 3.0 | 2.1 | 11.5 |
| Alice | 3.2 | 7.0 | 20.5 |

Table 2: Evaluation scores for additional models on morality, skill, and total performance.

informed (3.0/5) and skilled (6.6/10) dimensions but still struggling with engagement. Alice, my prompted LLM agent without retrieval capabilities, reached 20.5 points total, performing better on engagement (3.7/5) and demonstrating more balanced performance across dimensions.

The advanced implementations showed substantial improvements over these baselines. my retrieval-augmented generation (RAG) model Aragorn achieved 23.3 points, with notably stronger performance in engagement (4.3/5), informed (3.7/5), and skilled (7.7/10) dimensions compared to baseline models. This performance improvement validates the effectiveness of combining a knowledge base with LLM capabilities for argumentation. Victor, my most sophisticated implementation, which enhanced the RAG architecture with chain-of-thought reasoning, marginally outperformed Aragorn with 23.4 points, achieving the highest skilled rating (8.0/10) among all models.

# 6 Discussion

my evaluation of different argubot implementations revealed several intriguing patterns regarding the relationship between model complexity and performance. While I initially hypothesized that increasing model sophistication would consistently yield better results, my findings indicate a more nuanced reality: after a certain point, additional complexity can actually impede performance.

## 6.1 Complexity vs. Performance

Progressing from my simplest model (Airhead) to moderate implementations (Akiko, Alice) showed clear improvements. However, specialized variants (Chain of Thought, Prompt Engineering, Few-Shot Learning) with scores ranging from 21.9 to 22.4 did not outperform streamlined RAG implementations (Aragorn at 23.3 and Victor at 23.4), despite additional reasoning mechanisms. This suggests an optimal complexity level, beyond which performance may decline.

## 6.2 Strategic Approaches Across Models

my argubots employed distinct strategies that reflected their architectural differences:

| Model | Engaged | Informed | Intelligent |
|---|---|---|---|
| Aragorn | 4.3 | 3.7 | 4.0 |
| Victor | 4.3 | 3.7 | 3.7 |
| CoT | 3.9 | 3.4 | 3.7 |
| PE | 4.0 | 3.3 | 3.7 |
| Few-Shot | 3.9 | 3.8 | 3.6 |

Table 3: Evaluation scores for models on engagement, informativeness, and intelligence.

| Model | Moral | Skilled | **TOTAL** |
|---|---|---|---|
| Aragorn | 3.6 | 7.7 | 23.3 |
| Victor | 3.7 | 8.0 | 23.4 |
| CoT | 3.5 | 7.4 | 21.9 |
| PE | 3.5 | 7.6 | 22.1 |
| Few-Shot | 3.4 | 7.7 | 22.4 |

Table 4: Evaluation scores for models on morality, skill, and total performance.

### 6.2.1 Knowledge-Based vs. Reasoning-Based Approaches

Akiko relied primarily on retrieval without reasoning capabilities, while Alice demonstrated reasoning without specialized knowledge. This dichotomy was evident in their conversation patterns—Akiko would often present relevant facts without adapting to the emotional context, while Alice showed empathy but sometimes lacked factual depth.

### 6.2.2 Integration vs. Specialization

my most successful models (Aragorn and Victor) integrated multiple capabilities rather than specializing deeply in one area. For example, when discussing vaccine development with a skeptical dialogue partner, Victor demonstrated this integration effectively:

```
> Darius: "They rushed that vaccine
through. How can we trust
something developed so quickly?"
>
> Victor: "I understand your concern
about the timeline. While the
COVID vaccines were developed faster
than traditional vaccines,
they actually built upon decades of prior
research on similar
coronaviruses and used well-established
mRNA technology that had
been in development since the 1990s. The
urgency just meant more
resources and parallel testing phases
rather than cutting corners
on safety."
```

This response combines retrieval (specific facts about mRNA development history) with reasoning (acknowledging the concern while providing contextual explanation).

### 6.2.3 Expectations vs. Evaluation Results

Several findings contradicted my initial expectations:

- I anticipated that Chain of Thought reasoning would substantially improve performance, but its implementations

showed only modest gains and sometimes performed worse than simpler models. Analysis of dialogue transcripts showed that Chain of Thought sometimes generated overthinking that derailed conversations with unnecessary complexity.

- Few-Shot Learning performed better than expected (22.4 points), particularly in the "informed" dimension (3.8/5). This suggests that improving query formulation has outsized benefits for argubot effectiveness by enhancing the relevance of retrieved information.

- The marginal performance difference between Aragorn and Victor (23.3 vs. 23.4 points) was surprisingly small given Victor's additional complexity. This suggests that simple RAG with good prompting may already capture most of the achievable performance benefits.

### 6.2.4 Implications

Effective argubot design should prioritize balanced integration of capabilities over maximizing complexity. Streamlined RAG implementations with calibrated reasoning layers are most effective. Future work should explore dynamic complexity adjustment based on conversation context. The slight performance advantage of Victor over Aragorn suggests diminishing returns to increasing complexity beyond a certain threshold. Resources might be better allocated to improving core retrieval quality and response relevance.

## 7 Conclusion

This research explored different strategies for enhancing dialogue agents' ability to facilitate mind-opening discourse. Building upon the successful RAGAgent architecture, which outperformed the baseline Alice model, I implemented and evaluated my distinct enhancement approaches. Chain of Thought reasoning emerged as particularly effective, matching the base RAGAgent's performance (23.1) at some run and achieving the highest engagement score (4.5).

The consistent challenge across all implementations in achieving higher scores for moral and informed metrics (3.4-3.9) suggests that future work should focus on improving ethical reasoning and information delivery. my findings indicate that effective mind-opening dialogue benefits more from thoughtful analysis and strategic planning than from complex technical enhancements, pointing toward a promising direction for future development in combining Chain of Thought reasoning with robust context handling while maintaining implementation simplicity.

## 8 Limitations

While my research demonstrates promising approaches for enhancing dialogue agents, several important limitations should be noted. First, my evaluation relies heavily on model-based metrics, which, while efficient and reproducible, may not fully capture the nuances of real human-to-human dialogue and persuasion dynamics. The small set of simulated characters (Bob, Cara, etc.) used for testing, while diverse, represents only a limited subset of possible personality types and argumentation styles.

Another significant limitation is the constrained nature of my knowledge base, which relies solely on the Kialo debate platform. While Kialo provides structured arguments, it may not capture the full spectrum of informal reasoning patterns and cultural contexts that emerge in natural conversations. Additionally, the topic detection system's reliance on keyword matching could oversimplify complex, multi-faceted discussions.