Retrieval-Augmented Generation (RAG) for Phrasing Awkward Messages

Bella Perel (bellaperel@campus.technion.ac.il),
Tomer Shigani (tomershigani@campus.technion.ac.il),
Or Cohen (cohen.or@campus.technion.ac.il)

December 11, 2025

Abstract

This project introduces a novel model designed to phrase awkward messages in a manner that closely resembles human communication, surpassing the performance of leading models like GPT-4 and Claude. To accomplish this, we employed a Retrieval-Augmented Generation (RAG) framework, integrating datasets known for human-like, sensitive, and clear language, including the *Friends* TV show script, Reddit relationship advice, and the Google Emotion dataset. Various embedding models were examined to identify those that produced the best results. For evaluation, our model was compared to GPT-4O and Claude on three metrics—human-likeness, relevance, and coherence—using a set of 200 prompts. Results indicated that our RAG model significantly outperformed the alternatives in human-likeness ($p \approx 0$), while in relevance and coherence, the improvements over GPT-4 were not statistically significant ($p \succ 0.05$).

1 Introduction

In today's digital world, communication through text-based platforms—emails, messaging apps, and social media—has become central to both personal and professional interactions. With growing reliance on written communication, the need for AI-driven tools that can accurately and empathetically handle sensitive, nuanced exchanges is evident. Studies reveal that users increasingly depend on AI text generators, which have shown benefits like enhanced productivity and confidence; however, these tools often fall short in situations that require tact, emotional intelligence, and subtlety [Robinson(2024)], [Jackson(2024)]. While current large language models (LLMs) provide powerful assistance, they lack critical elements of human-like interaction. For instance, these models frequently produce responses that feel overly formulaic or lacking in emotional nuance, particularly in delicate or awkward communications. [Chien-Sheng Wu(2024)] shows these models struggle with conveying complex sentiments, often defaulting to generalized language that overlooks context-specific nuances essential for sensitive interactions. Additionally, in previous work that focuses on alignment training and context sensitivity in AI highlights the challenges LLMs face in producing varied, personalized responses that genuinely resonate with human users, rather than reinforcing predictable or biased patterns [Jackson(2024)]. Previous approaches have attempted to address these issues by focusing on model editing techniques, instruction-tuning, and dataset refinement. However, these efforts have proven limited, as models still exhibit weaknesses in adapting to the diverse linguistic cues involved in sensitive human interactions. The complex, multi-step editing processes introduced by researchers like [Chien-Sheng Wu(2024)] aim to better mimic human editing behavior, but they only modestly improve the emotional authenticity of generated text. Moreover, attempts to humanize AI outputs are often hindered by reward hacking, where models exploit evaluative flaws to produce artificially high-scoring responses that lack genuine empathy and coherence. This project seeks to address these limitations by implementing a Retrieval-Augmented Generation (RAG) framework. This approach integrates carefully selected datasets—*Friends* scripts, Reddit relationship advice, and the Google Emotion dataset—to more closely model the complexity of human emotional and social dynamics. By bridging these gaps, our RAG-based LLM aims to set a new standard for AI-driven communication tools, offering responses that align more naturally with human emotional sensitivity and context.

2 Methodology

The preprocessing methodology for this project involved preparing three key datasets to ensure consistent, high-quality inputs for the model.

1. **Friends Dataset**: To streamline the text and maintain its flow, all director's notes were removed. The cleaned text was then segmented by scene, with any remaining whitespace eliminated, and saved as a CSV file, with each row representing an individual scene. This process preserved the conversational structure while optimizing the dataset for training.

2. **Google Emotion Dataset**: The Google Emotion dataset, which includes multiple binary annotations for each emotion label from various annotators, necessitated further processing to derive cohesive labels for each text entry. A majority vote was calculated across annotators for each emotion, with the results rounded to generate binary labels. This method ensured that each text entry contained a singular, representative set of emotion labels. To enrich the semantic content of each text, the identified emotions were concatenated at the end of the text. This enhancement was designed to facilitate the Retrieval-Augmented Generation (RAG) search process, allowing for more appropriate reactions when tasked with expressing specific emotions. By integrating emotional context directly into the text, the model can better identify and generate responses that resonate with users' emotional needs.

3. **Reddit Relationship Advice Dataset**: This dataset included two comments per post, each labeled for sentiment. To increase the granularity of the data, the comments were exploded so each comment appeared in a separate row alongside its corresponding post text. Each post-comment pair was then concatenated, including the label, to create a "full text with label" field that integrates the advice context with sentiment annotation.

These preprocessing steps facilitated consistent formatting and enriched each dataset with relevant context, preparing them effectively for integration into the model pipeline.

3 Experiments

3.1 Experimental Setup

Application Structure

Initializing Cohere and Pinecone Clients

- **Cohere Client**: Used to generate responses based on the input prompt.
- **Pinecone Client**: Manages vector embeddings, storing and indexing them for fast similarity searches.

Loading Chosen Sentence Transformer Model

`load_model()` loads the `SentenceTransformer` model `"all-MiniLM-L6-v2"` for embedding user input into vector space.

Loading Precomputed Embeddings and Loading a Pinecone Index

Precomputed embeddings, generated using the `"all-MiniLM-L6-v2"` model, are loaded as context for Retrieval-Augmented Generation. The `create_pinecone_index()` function loads the pinecone index which is structured to store high-dimensional vectors and metadata, supporting efficient searches for similar content. The Pinecone index, named "mini-lm-6", is configured with cosine similarity in a 384-dimensional vector space, aligning with the embeddings generated by the all-MiniLM-L6-v2 model.

Upserting Vectors into Pinecone

`upsert_vectors()` stores vectors (embeddings) in the Pinecone index, preparing each entry as a tuple containing an ID, vector, and associated information from the dataset to provide context. Unique IDs are generated for each embedding, and the original text content is included as additional contextual data for retrieval.

Augmenting Prompt for Cohere

`augment_prompt()` encodes a user query using the loaded `all-MiniLM-L6-v2` model, transforming it into a vector representation for similarity search, retrieves the top 3 most relevant stored texts from the Pinecone index, and uses them as source context. This context builds an augmented prompt engineered to guide Cohere's language model to generate a friendly, empathetic, WhatsApp-style response tailored to the user's query. The prompt leverages our provided context (database of similar past messages) to enhance relevance by matching the style and tone found in these examples. The finalized prompt utilized to guide the model is presented below:

```
augmented_prompt = f"""You are tasked with composing a WhatsApp message that you would send
directly to the person, maintaining the tone, style, and level of empathy and directness used in the
provided source material.

    The response should simulate a real-time, casual WhatsApp message.
    Ensure the tone is empathetic and conversational, while remaining concise and clear.
    Use the human writing style from the source knowledge as a guide, but note that the
    source knowledge does not contain direct answers to the query.

    Important Guidelines:
    - The response must directly answer the query as if you are sending the message right now.
    - Maintain the casual tone, while ensuring the message is smooth and empathetic,
    like a typical WhatsApp conversation.
    - The source knowledge is provided solely to show the desired tone
    and writing style. It is not
    to be used as a source of answers or content for the response.
    - Stick strictly to the format of a direct message, avoiding extra
    advice or unwarranted sympathy.

    Example Query and Response Format:

    Query: \I agreed to be a bridesmaid, but now I can't commit.
    How can I let the bride know without causing drama?"

    Response: \Hey, I don't know if you've noticed, but I'm kind of freaking out about this whole
    bridesmaid thing. I'm so sorry, but I don't think I can do it anymore.
    I know how important your wedding is, and I don't want to let you down,
    but I'm just not in the right headspace. I hope you understand, and that we can still be cool."

    The source knowledge is as follows, and it is highly important to use the context
    and the human writing style in it to write the message:
    {source_knowledge}
    Query: {query}"""
```

3.2 Evaluation Metrics

To evaluate our Friends-RAG model, we compared its performance against two leading large language models (LLMs), GPT-4O and Claude. Notably, we opted not to utilize their APIs, as previous experiences indicated that the results obtained through them were suboptimal; instead, we directly interacted with their interfaces to achieve more reliable outcomes. Each of the three models was presented with 200 carefully crafted prompts that addressed various awkward situations, such as providing

unfavorable feedback to a vendor or informing a friend about unpleasant breath. We included both professional prompts (e.g., messages directed at a boss or colleague) and informal ones to assess the models' adaptability to different contexts.

Following the collection of results, we conducted a comprehensive evaluation of the models based on three primary aspects: human-likeness, relevance, and clarity. We established detailed annotation guidelines for the evaluation process (attached as part of the git repository), and each team member manually annotated a randomized selection of 200 examples, ensuring a diverse representation of categories and models. Initially, we considered evaluating three additional metrics. However, after testing these criteria, we found they were not consistently relevant across prompts. For example, while empathy is crucial in scenarios like breaking up with someone, it's far less applicable in cases such as filing a complaint about poor service. After careful consideration, we decided to focus exclusively on the three primary scores that demonstrated consistent applicability across all prompt types. Once the annotations were completed, we analyzed the results to identify any significant differences in the distribution of ratings among the models. We employed the Mann-Whitney test to assess the p-values, determining whether the RAG scores were significantly higher than those of the baseline models. Additionally, we calculated the confidence intervals for the differences between the models. figures and tables 1-3 present the results for all models and rates.

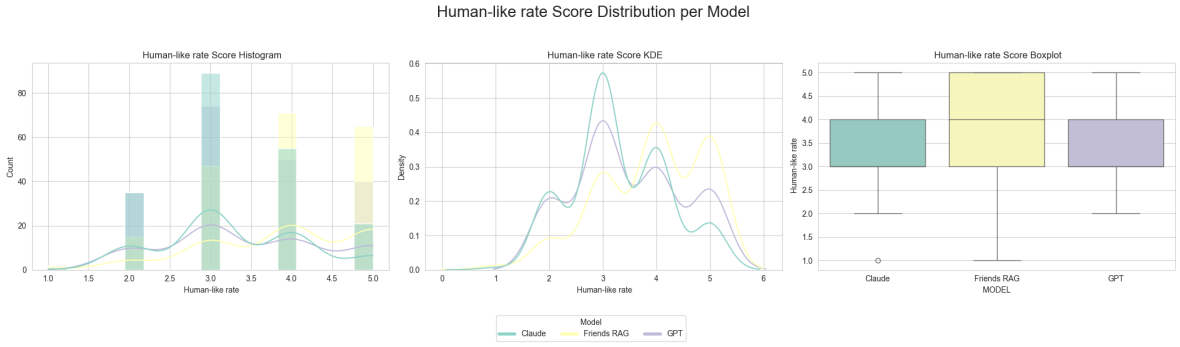

Figure 1: Human likeliness score distribution by model.

Model	P-value	Mean Difference	95% CI Lower Limit	95% CI Upper Limit
RAG Vs. GPT 4O	≈ 0	0.432	0.277	0.623
RAG Vs. Claude	≈ 0	0.61	0.417	0.796

Table 1: Human likeliness' statistical measurements comparing the performance of the Friends-RAG model against baseline models.

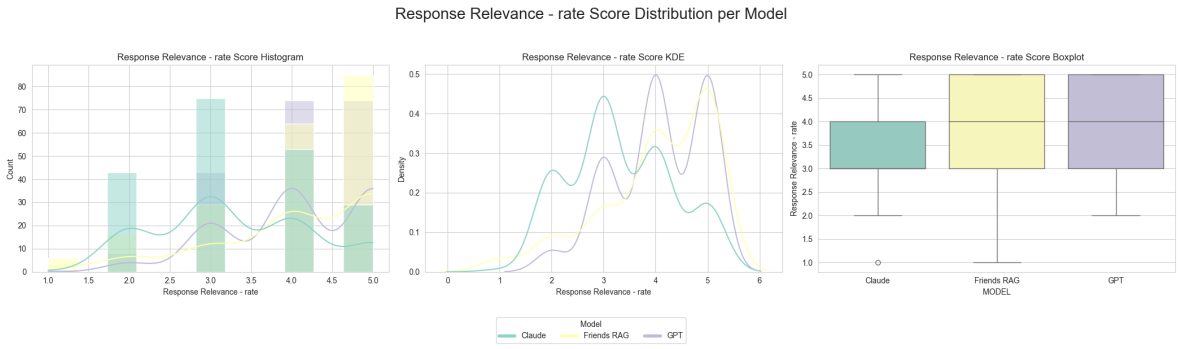

Figure 2: Relevance score distribution by model.

The findings indicate that the Friends-RAG model significantly outperformed both GPT-4O and Claude in terms of human-likeness, demonstrating its potential for effective interpersonal messaging in sensitive contexts. Additionally, the RAG model showed a notable advantage in generating contextually

Model	P-value	Mean Difference	95% CI Lower Limit	95% CI Upper Limit
RAG Vs. GPT 4O	0.34	-0.042	-0.23	0.15
RAG Vs. Claude	≈ 0	0.699	0.507	0.901

Table 2: Relevance's statistical measurements comparing the performance of the Friends-RAG model against baseline models.

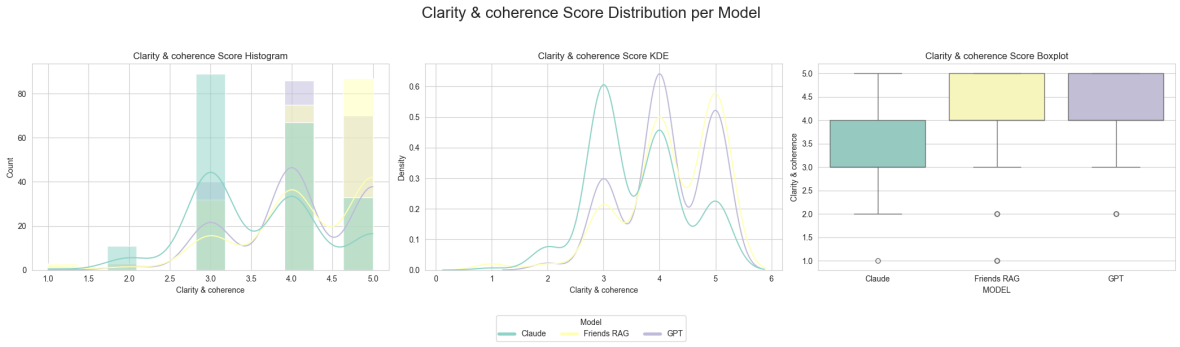

Figure 3: Clarity coherence score distribution by model.

appropriate responses compared to Claude, while no meaningful difference was observed between RAG and GPT-4O in this area. Furthermore, although the RAG model was more coherent and clear than Claude, it did not outperform GPT-4O in terms of coherence.
The distribution of human-likeness ratings from different annotators is illustrated in figure 4, in order to ensure that no individual annotator favors the RAG model over others, potentially introducing bias into the results.

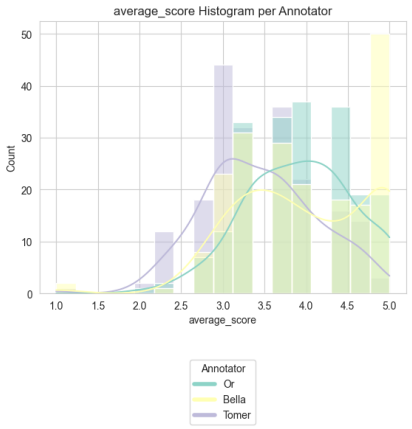

Figure 4: Average score histogram per annotator

The distribution of average scores among annotators reveals a notable variance in evaluation patterns. Specifically, Or and Bella's evaluations showed a p-value of 0.38, suggesting no significant bias between them. In contrast, comparisons involving Tomer with both Bella and Or yielded p-values near zero, indicating potential differences in their scoring patterns. Notably, Tomer's scores are generally lower than those of the other annotators, suggesting a stricter evaluation tendency. This analysis indicates that while Or and Bella share a similar scoring approach, Tomer's evaluations differ significantly.

4 Discussion

The results of this project indicate that the Friends-RAG model significantly outperforms Claude in generating human-like responses. However, its relevance scores did not significantly exceed GPT-4O's,

Model	P-value	Mean Difference	95% CI Lower Limit	95% CI Upper Limit
RAG Vs. GPT 4O	0.073	0.074	-0.091	0.23
RAG Vs. Claude	≈ 0	0.601	0.428	0.762

Table 3: Clarity's statistical measurements comparing the performance of the Friends-RAG model against baseline models.

which may be due to the model's tendency to hallucinate, or generate irrelevant content, rather than strictly follow the provided context. In terms of clarity and coherence, all models performed well, which likely limited the margin by which RAG could excel in this category. Importantly, the model's strong performance in human-likeness supports its suitability for our primary goal of enhancing conversational authenticity in challenging contexts.
To make the Friends-RAG model accessible, we also developed a user-friendly interface, enabling users to interact with the model directly and experience its capabilities in real-time.
One challenge in this evaluation was the manual annotation process, which introduced variability due to subjective judgments and potential biases among annotators. Even with the use of annotation guidelines, each annotator may interpret rating criteria differently, leading to inconsistencies that can affect the stability of the results. Additionally, the manual approach restricted the dataset size and made it harder to achieve consistent ratings across categories. These limitations may have influenced the model's evaluated performance and highlight areas for refinement in future assessments.

To mitigate these issues, a valuable improvement would be multi-tagging each example by multiple annotators and then calculating a majority vote. This approach would provide a more balanced perspective by reducing the impact of individual biases and outlier opinions. Majority voting helps stabilize scores, ensuring that ratings more accurately reflect consensus among annotators, which would provide a stronger foundation for evaluating model effectiveness.

Future work in the area could also focus on enhancing model robustness, interpretability, and relevance within sensitive communication contexts. Given the observed hallucination tendency of the RAG model, one avenue for improvement is developing mechanisms to minimize this issue, possibly through a filtering layer that identifies and reduces irrelevant or fabricated information. Fine-tuning with additional, high-quality data related to the task could also reduce hallucinations and improve contextual alignment. Another promising direction would be exploring adaptive retrieval strategies that allow the model to adjust its retrieval context based on prompt-specific requirements. For instance, adding context-sensitivity rules or dynamically adjusting retrieval sources based on prompt type could improve relevance scores, especially in complex or nuanced conversations.

5 Appendix

Git link

5.1 Datasets

Google Emotions
Friends Script
Reddit Relationship Advice

References

[Chien-Sheng Wu(2024)] Tuhin Chakrabarty Chien-Sheng Wu, Philippe Laban. 2024. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits.

[Jackson(2024)] Ivan Jackson. 2024. How to humanize ai text: A practical guide.

[Robinson(2024)] Kendra Robinson. 2024. Linguistic challenges in ai writing: Idioms, sarcasm, and context.