AWS Machine Learning Engineer Nanodegree Capstone

Harry Potter English Language Assistant using RAG + Llama 3

Federico Puy

Definition

Project Overview

The domain of this project lies at the intersection of language learning, natural language processing (NLP), and popular literature. English language acquisition has always been a challenging task for non-native speakers, with various methods being employed to aid learners. One effective way to learn a language is through immersion in engaging and culturally significant texts. The "Harry Potter" series, written by J.K. Rowling, is one of the most beloved and widely-read book series in modern history. Its rich language and intricate storytelling make it an excellent resource for learning English.

Historically, language learning has evolved from rote memorization and grammar exercises to more immersive and interactive methods, such as using multimedia resources, games, and conversational agents. The rise of AI and NLP has opened new avenues for creating intelligent tutoring systems that can adapt to individual learners' needs. Research in language learning suggests that contextual learning, where learners are exposed to language use within a meaningful context, significantly improves retention and comprehension. Given the popularity of Harry Potter and the potential of AI-driven tools, creating an English language learning assistant based on the Harry Potter books represents a novel and effective approach to language acquisition.

Problem Statement

The problem this project aims to solve is the lack of engaging, context-rich tools for English language learners that can provide real-time feedback and personalized learning experiences. Traditional language learning tools often lack interactivity and contextual relevance, leading to reduced motivation and engagement. There is a need for a solution that combines the immersive experience of reading a popular book series like Harry Potter with the advanced capabilities of AI to create a dynamic learning environment.

Specifically, the challenge is to develop a chatbot capable of understanding and generating natural language responses, contextualizing its answers within the Harry Potter universe, and providing meaningful feedback to learners. The solution must be scalable, accessible, and effective in improving the learner's language skills.

Solution Statement

The proposed solution is an AI-powered English language learning assistant that utilizes RAG and the Harry Potter books to engage learners in meaningful conversation. The assistant will be deployed as a chatbot that can interact with users in real-time, answering their questions,

engaging in dialogue, and providing explanations based on the content of the Harry Potter series.

The assistant will be built using Meta's Llama 3 model, a state-of-the-art NLP model, and Chroma DB as the vector database to efficiently store and retrieve context-relevant passages from the books. The model will be hosted on AWS SageMaker, ensuring scalability and robustness, and deployed via AWS Lambda to provide seamless user interactions.

This solution is designed to be quantifiable, measurable, and replicable. Learner progress can be tracked through metrics such as the accuracy of responses, the complexity of language used by the learner over time, and user engagement metrics. The solution will be iteratively improved based on feedback and performance data, ensuring its effectiveness as a language learning tool.

Analysis

Datasets and Input

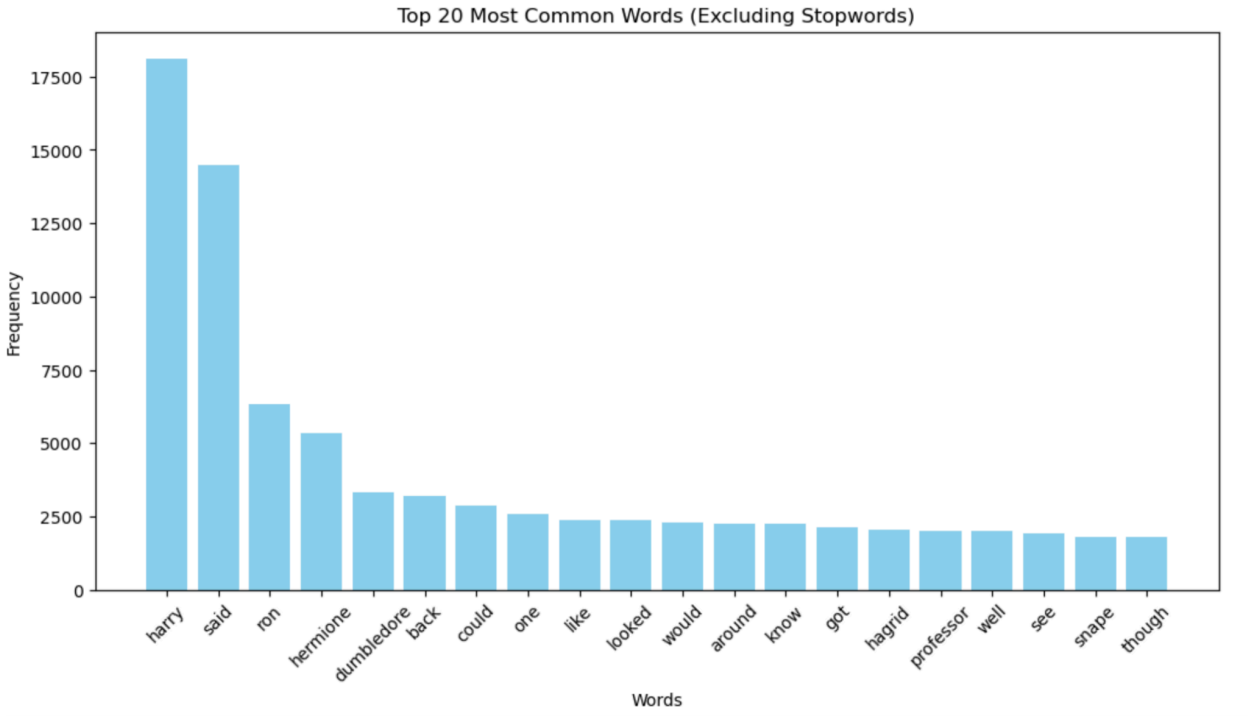
The primary dataset for this project consists of the full text of the Harry Potter books. These texts are structured into a DatasetDict format, which includes separate splits for different types of interactions (e.g., direct quotes, narrative descriptions, and dialogues). The dataset has been preprocessed to include tokens and vector representations suitable for embedding in Chroma DB.

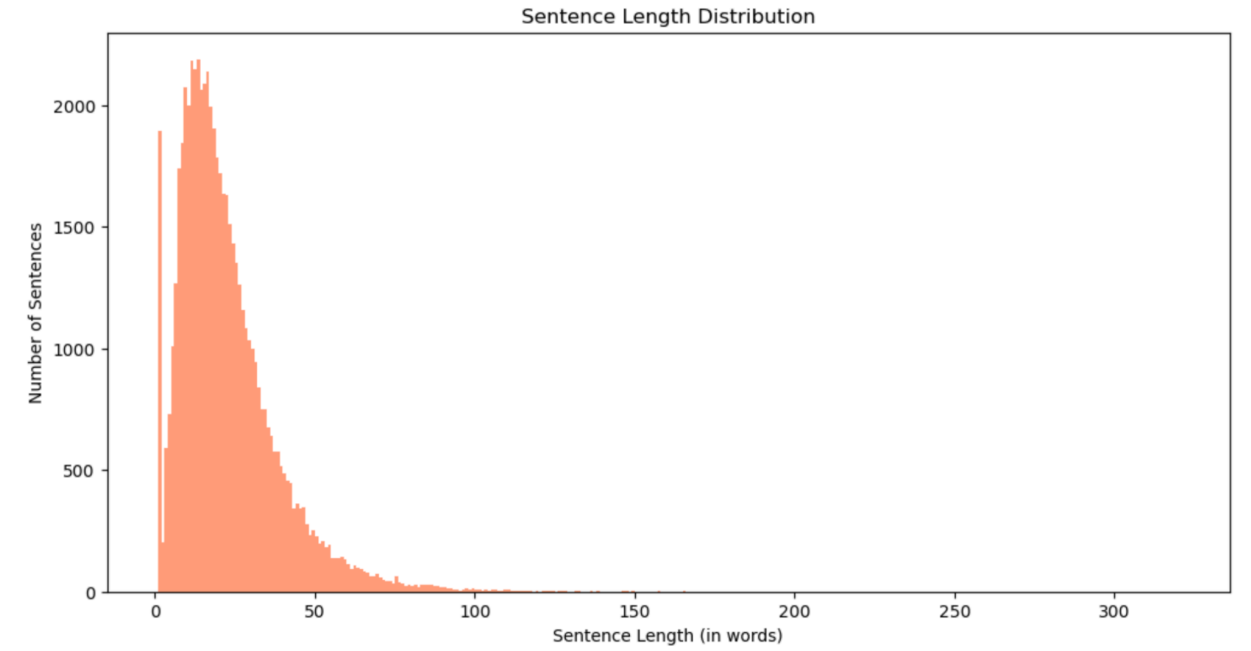
The dataset was obtained through legal means and formatted into a structure compatible with the Llama 3 model. Each book's content is indexed and vectorized to allow quick retrieval based on user queries. The use of this dataset is appropriate because it provides a rich and immersive context for learners to engage with, helping them to see language use in a narrative form, which is proven to be effective in language learning.

The files will be obtained from a Github repository and uploaded to a local S3 bucket. For the RAG implementation, we download the files to the local notebook instance.

Data Analysis

Total Number of Words: 1327660
Number of Unique Words: 24740
Total Number of Sentences: 60028
Average Sentence Length: 22.12 words





Benchmark Model

As a benchmark model, a base chatbot using the standard Llama 3 8b Instruct model will be implemented. This model will rely on keyword matching and predefined responses without any advanced contextual understanding. This benchmark will serve as a baseline to compare the

effectiveness of the RAG-based Llama 3 model. The comparison will focus on response accuracy, contextual relevance, and user satisfaction.

The benchmark model is crucial as it provides a clear, measurable point of comparison to evaluate the improvements offered by the advanced AI-driven solution.

Evaluation metrics

The performance of both the benchmark and the proposed solution will be evaluated using several metrics:

1. **Accuracy**: The correctness of the chatbot's responses is based on predefined ground truths.
2. **Contextual Relevance**: The extent to which the chatbot's responses are contextually appropriate, evaluated through user feedback and expert review.
3. **User Engagement**: Measured by interaction duration, the number of interactions per session, and user return rates.
4. **Learning Outcomes**: Assessed through pre- and post-interaction language assessments, tracking improvements in vocabulary, grammar, and comprehension.

These metrics will ensure that the chatbot is not only functional but also effective as a learning tool.

The tools used to evaluate these metrics will be ‘Human Evaluation’ and ‘BLEU’ score calculation. BLEU (Bilingual Evaluation Understudy) metric, a widely used method for evaluating the quality of machine-generated text. It works by comparing the overlap between n-grams (sequences of words) in the generated text and one or more reference texts created by humans. The more n-grams shared with the reference, the higher the BLEU score. However, BLEU emphasizes precision over recall, meaning it rewards models for producing correct sequences but doesn't penalize them for missing relevant content. While useful for assessing syntactic and lexical accuracy, BLEU doesn't fully capture aspects like fluency, creativity, or semantic equivalence, which limits its effectiveness in some contexts.

Presentation

Interactions between the user and the chatbot will be presented using a Gradio Chatbot interface. This provides an easy and intuitive way to interact with the Chatbot, just like most popular LLMs do.

Additionally, the endpoint will be accessible via a POST endpoint implemented in AWS Gateway, which will trigger an AWS Lambda function which accesses the endpoint. This is to allow other consumers to use the endpoint, such as mobile apps or other web interfaces.

Methodology

Data Preprocessing

The books text will be split into smaller chunks, which will then be stored as documents in the Chroma Database. To split them, we will use RecursiveCharacterTextSplitter from the popular LangChain library. The RecursiveCharacterTextSplitter takes a large text and splits it based on a specified chunk size. It does this by using a set of characters. The default characters provided to it are ["\n\n", "\n", " ", ""].
It takes in the large text and then tries to split it by the first character \n\n. If the first split by \n\n is still large then it moves to the next character which is \n and tries to split by it. If it is still larger than our specified chunk size it moves to the next character in the set until we get a split that is less than our specified chunk size.

This is a quick but effective solution. There are more efficient methods for splitting, such as Semantic chunking or LLM Chunking but they were not implemented at this point.

Implementation

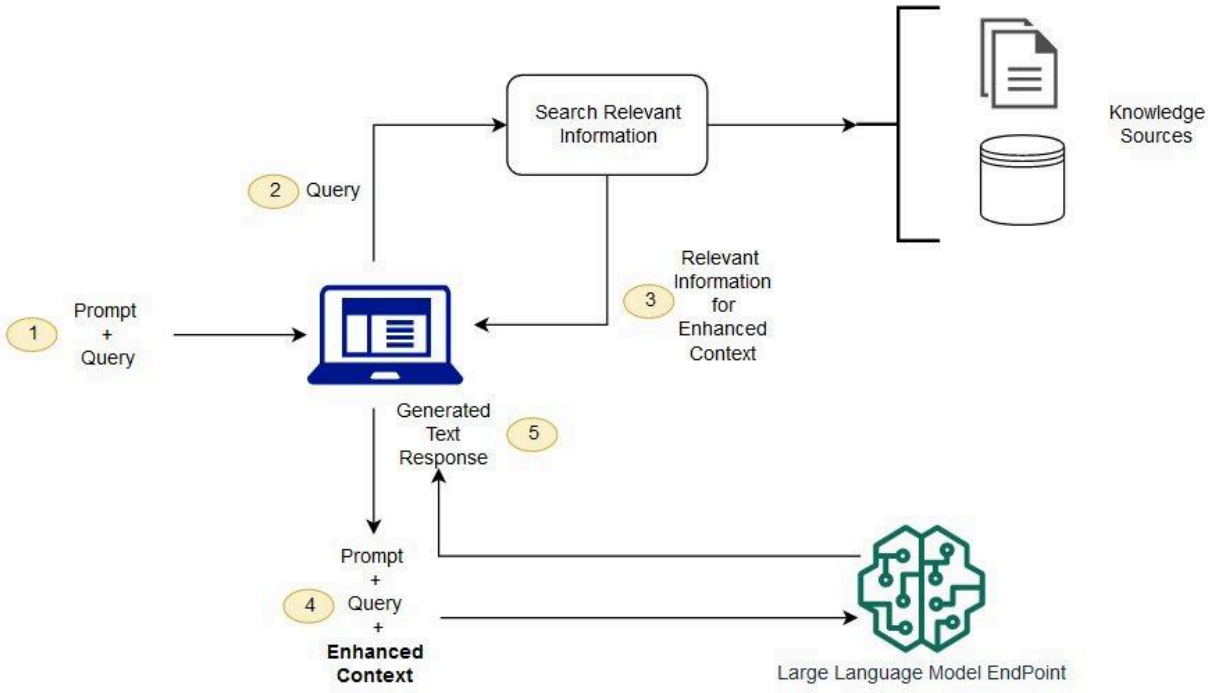
Once the text has been split, we create a new collection in ChromaDB, and the chunks are added to it. ChromaDB will create the embeddings for the chunks.

Once the database is populated, we can query the collection based on our desired prompt. By default, ChromaDB will use cosine distance to calculate the closest document to our prompt. This setting can be changed.

For our chatbot experience, we will allow the user to start the conversation. Immediately after, the chatbot will always ask the same question, “What is the thing you most like about Harry Potter”. The answer will be the prompt we use in our RAG system, and will be the topic to be used during the English Language Learning experience. The process is outlined below:

1) Chatbot asks ‘What is the thing you most like about Harry Potter’
2) User responds, answer is used to fetch the document with the smallest cosine distance to it. ChromaDB will embed the answer for us.
3) Chatbot will use the document as context to ask questions to the user.
4) After 10 questions, the Chatbot will stop the conversation.

Overview of RAG implementation:



[Extracted from AWS Documentation]

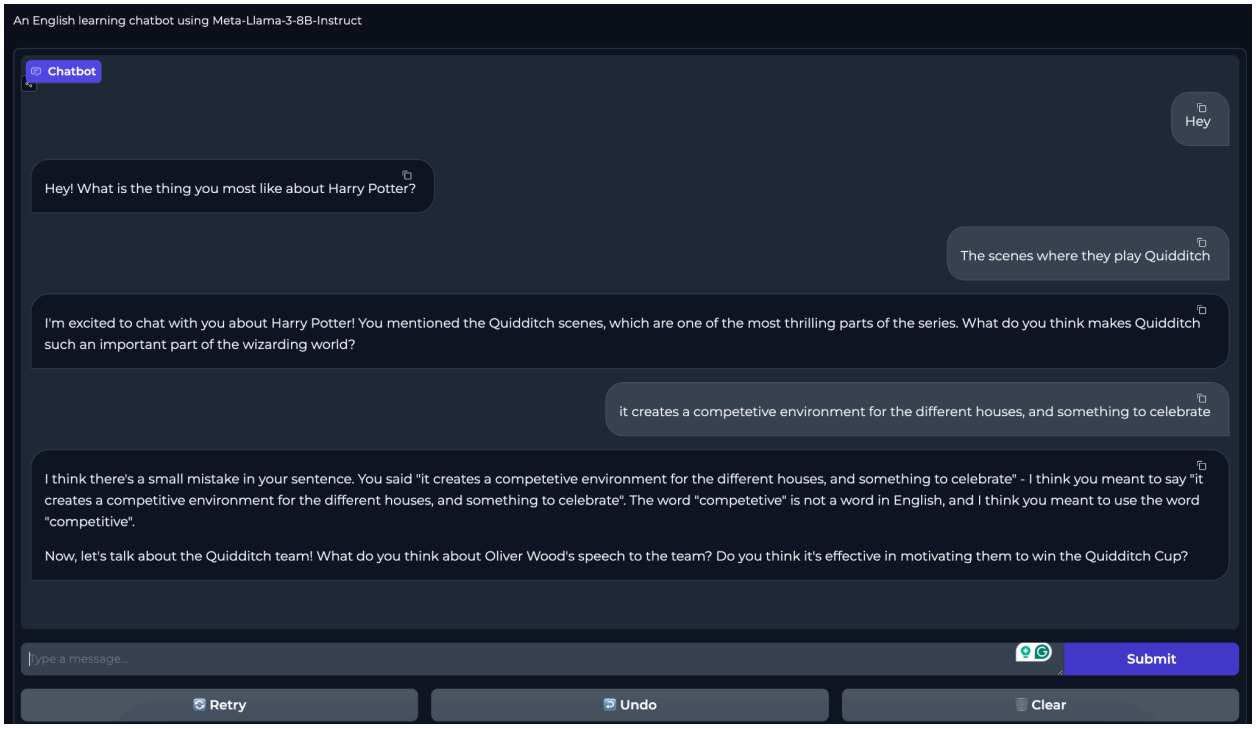
Refinement

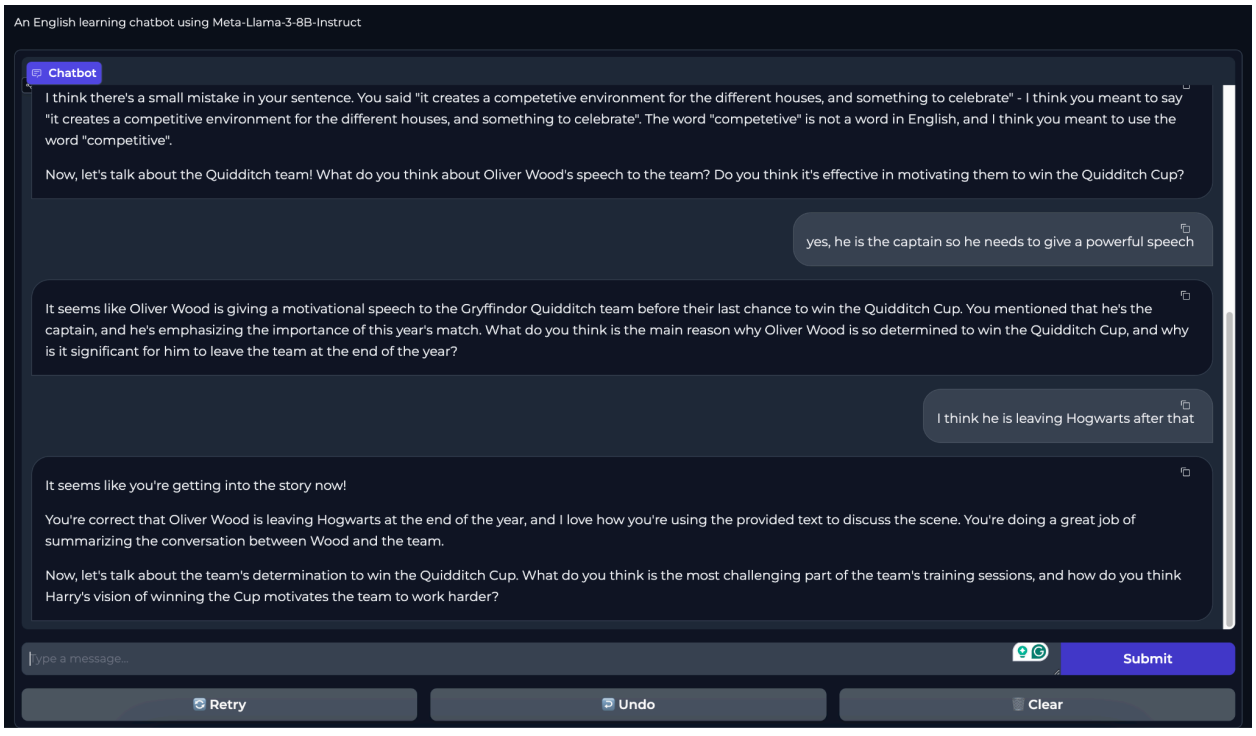
The initial solution proved to be effective in retrieving contextual information and presenting it to the user, but several iterations were required to obtain an engaging experience, where the chatbot would keep the conversation flowing. The key factor here was the system prompt. By leveraging **Prompt Engineering** techniques, I was able to craft a prompt that would result in a useful yet non-disruptive experience, that would help the conversation flow but still correct the user when making mistakes.

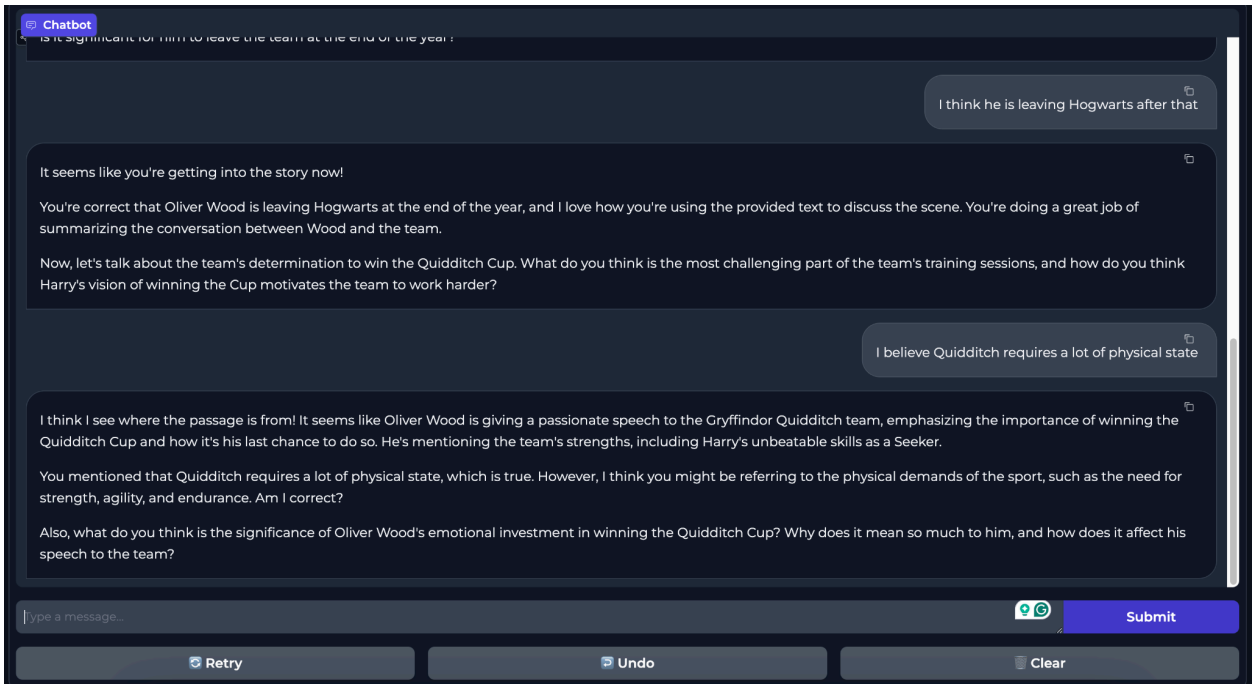
Results

Model Evaluation and Validation

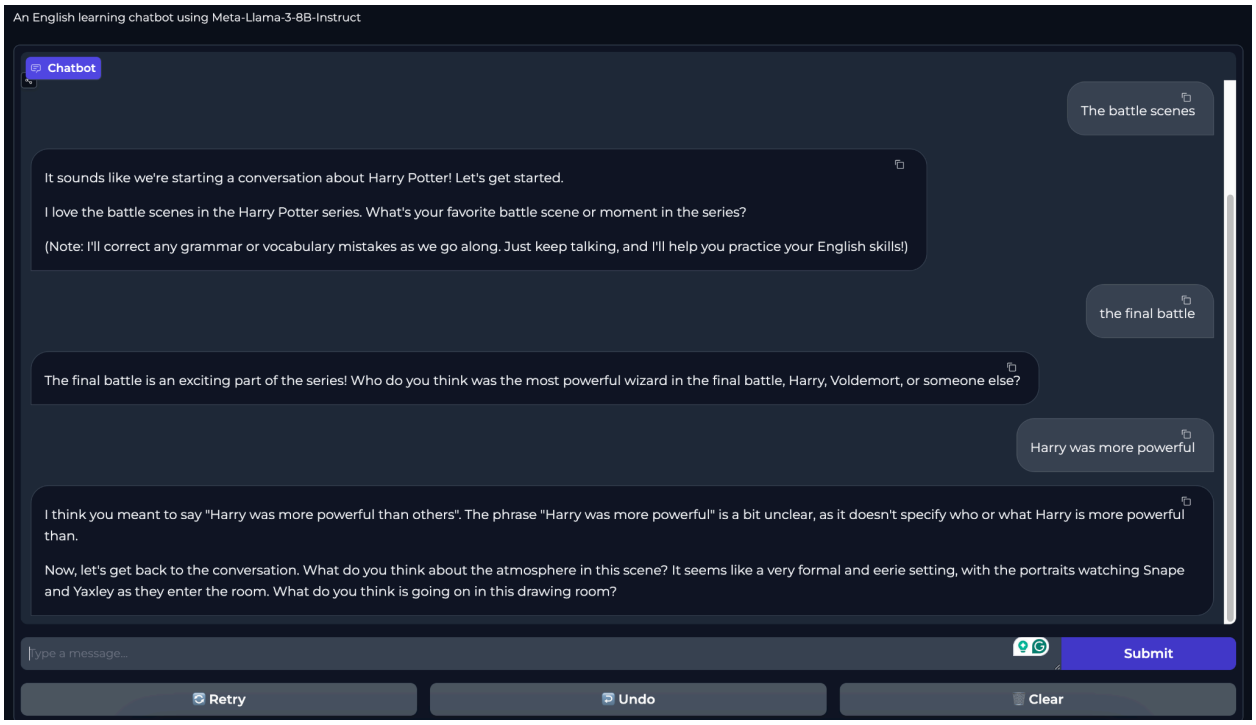
Example
Below is an example to showcase the interaction with the bot. The user mentions Quidditch, so the bot picks up a passage where a Quidditch speech is being given and bases the conversation around that. Grammar and vocabulary mistakes are corrected by the bot.


An English learning chatbot using Meta-Llama-3-8B-Instruct
Chatbot
Hey
Hey! What is the thing you most like about Harry Potter?
The scenes where they play Quidditch
I'm excited to chat with you about Harry Potter! You mentioned the Quidditch scenes, which are one of the most thrilling parts of the series. What do you think makes Quidditch such an important part of the wizarding world?
it creates a competetive environment for the different houses, and something to celebrate
I think there's a small mistake in your sentence. You said "it creates a competetive environment for the different houses, and something to celebrate" - I think you meant to say "it creates a competitive environment for the different houses, and something to celebrate". The word "competetive" is not a word in English, and I think you meant to use the word "competitive".
Now, let's talk about the Quidditch team! What do you think about Oliver Wood's speech to the team? Do you think it's effective in motivating them to win the Quidditch Cup?
Type a message...
Submit
Retry
Undo
Clear


An English learning chatbot using Meta-Llama-3-8B-Instruct
Chatbot
I think there's a small mistake in your sentence. You said "it creates a competetive environment for the different houses, and something to celebrate" - I think you meant to say "it creates a competitive environment for the different houses, and something to celebrate". The word "competetive" is not a word in English, and I think you meant to use the word "competitive".
Now, let's talk about the Quidditch team! What do you think about Oliver Wood's speech to the team? Do you think it's effective in motivating them to win the Quidditch Cup?
yes, he is the captain so he needs to give a powerful speech
It seems like Oliver Wood is giving a motivational speech to the Gryffindor Quidditch team before their last chance to win the Quidditch Cup. You mentioned that he's the captain, and he's emphasizing the importance of this year's match. What do you think is the main reason why Oliver Wood is so determined to win the Quidditch Cup, and why is it significant for him to leave the team at the end of the year?
I think he is leaving Hogwarts after that
It seems like you're getting into the story now!
You're correct that Oliver Wood is leaving Hogwarts at the end of the year, and I love how you're using the provided text to discuss the scene. You're doing a great job of summarizing the conversation between Wood and the team.
Now, let's talk about the team's determination to win the Quidditch Cup. What do you think is the most challenging part of the team's training sessions, and how do you think Harry's vision of winning the Cup motivates the team to work harder?
Type a message...
Submit
Retry
Undo
Clear



Another conversation where the bot corrects the grammar in the response.



Evaluating the chatbot was the hardest part of the project. Evaluating LLMs itself is complicated due to their vast scope and the multifaceted nature of their outputs. These models generate text

that can be assessed on various dimensions, including correctness, coherence, creativity, ethical considerations, and alignment with user intent. Additionally, their performance can vary greatly across different contexts and tasks, making it difficult to establish standardized metrics.

Initial evaluation was done manually, by engaging in multiple conversations with the chatbot and adjusting the different parameters, such as the chunk-splitting technique, document size, and system prompt. The initial implementation fetched the context from the RAG system on every interaction, but this proved to be overwhelming for the user, and it introduced additional latency to every single conversation which was unnecessary.

Further evaluation was done using the BLEU technique

A set of 10 prompts were chosen along with their respective 10 ‘ideal’ answers. These prompts were given to both the RAG system and the Base LLM, and their BLEU scores were calculated afterward.

Average BLEU score for RAG: 0.0960
Average BLEU score for LLM: 0.0912

Justification

The LLM + RAG system outperforms the base LLM, surely due to the added context. However, we need to note that the ‘ideal’ answers contain extra detail, so in this scenario, it makes sense for the LLM + RAG to have better performance.

The Base LLM model will surely outperform the RAG system if we were looking for more vague answers related to general Harry Potter topics.