# Using LangChain and Large Language Models to Predict Plot Points in Novels: Agatha Christie

Ryan Peruski
*Department of EECS*
*University of Tennessee, Knoxville*
yhg461@vols.utk.edu

Nolan Coffey
*Department of EECS*
*University of Tennessee, Knoxville*
ncoffey3@vols.utk.edu

Triton Eden
*Department of EECS*
*University of Tennessee, Knoxville*
teden@vols.utk.edu

William Duff
*Department of EECS*
*University of Tennessee, Knoxville*
wduff@vols.utk.edu

***Abstract*—This paper investigates the application of Natural Language Processing (NLP) techniques, specifically the use of LangChain with Large Language Models (LLMs) utilizing Retrieval Augmented Generation (RAG), to analyze and predict key plot points in Agatha Christie's novels. We also investigated the use of Coreference Resolution using spaCy's Coreferee library. The dataset comprises several Agatha Christie novels sourced from Project Gutenberg. Our analysis compares the performance of the use of LLMs with RAG and how accurately it can distinguish significant plot elements, such as the protagonist, antagonist, victim, climax chapter, and murder weapon. This work illustrates both the potential and limitations of LLM RAG methods in literary analysis.**

***Index Terms*—NLP, Natural Language Processing, Agatha Christie, Mystery Novels, Word Embeddings, LangChain, LLM, OpenAI, RAG**

## I. Introduction

State of the art Large Language Models (LLMs) have opened new possibilities in literary analysis, enabling researchers to delve into narratives with unprecedented precision. While traditional methods rely on close reading to interpret themes, characters, and language, advancements in Natural Language Processing (NLP) provide a means to identify structural patterns in text computationally. This report examines how Retrieval Augmented Generation (RAG), a technique within the LangChain framework, can be applied to the works of Agatha Christie, a celebrated figure in the mystery genre. By focusing on critical plot components such as protagonists, antagonists, victims, climax chapters based on when the antagonist is revealed, and murder weapons, we explore whether computational models can capture and predict the distinctive elements of Christie's storytelling. Using a dataset of her novels sourced from Project Gutenberg, this study highlights both the promise and challenges of integrating RAG with the LangChain framework for analysis of literary texts.

## II. Dataset

### A. Overview

For this study, we selected our dataset from Project Gutenberg [5], a well-established digital library offering free access to public domain literary works. Project Gutenberg provides reliable, standardized text formats suitable for Natural Language Processing analysis, making it an ideal source for literary datasets. By using this open access resource, we ensured consistent formatting across texts, facilitating a more straightforward pre-processing pipeline and maintaining a high level of text quality.

Our dataset includes three notable Agatha Christie novels: The Mysterious Affair at Styles, The Murder of Roger Ackroyd, and The Murder on the Links. These works were chosen for their availability on Project Gutenberg and their consistent use of Hercule Poirot as the central detective, allowing us to examine a recurring character across multiple narratives. By focusing on novels with the same protagonist, we can more effectively analyze character relationships and thematic patterns unique to Poirot's investigative style, making these selections well suited for studying structural patterns in Christie's writing. Together, these novels provide a focused yet representative sample of Christie's writing style and plot construction, forming a strong foundation for our LangChain-based reasoning system.

### B. Data Processing

To begin the data cleaning process, we applied the spaCy pipeline to perform standard pre-processing techniques, including tokenization, normalization, and removing stop words, on the raw text file for each book. A key enhancement to this workflow was the use of multi-word expressions to ensure consistency in character references. Characters are often referred to by various names, such as their full names, titles, or shortened forms. To address this, we created a single token for each character across the novels. For instance, 'Alfred Inglethorp' and 'Mr. Inglethorp' were both tokenized to the same value. This adjustment improves the accuracy of word

embeddings and ensures consistency throughout our analysis. Importantly, this step requires compiling a list of characters from each novel obtained using character wikis to accurately map these references.

Another critical step, which we introduced as an independent variable was coreference resolution. This process assigns the same token to all references to a character within a text. For example, in the sentence ”John went to the store where he purchased an apple,” both ’John’ and ’he’ would be tokenized to the same value. Using spaCy’s Coreferee Library, we implemented this step and ran tests with and without coreference resolution to assess its impact on the accuracy of our results.

## III. Methodology

### A. Methods

After retrieving the selected novels and completing the pre-processing steps, we manually extracted key plot components from each text to establish ground truths for evaluating prediction accuracy. Using LangChain’s Retrieval-Augmented Generation (RAG) system, we queried the texts with straightforward prompts such as “Who is the murderer?” and “Who is the protagonist?” The model’s predictions were then compared to the ground truths, and an accuracy score was calculated based on how many of the five plot components matched correctly. This scoring system provided a straightforward metric to assess the model’s effectiveness in identifying critical narrative elements.

### B. Code Libraries

The implementation leveraged several key libraries to facilitate the analysis. LangChain was used to incorporate RAG, enabling us to input the preprocessed novels into an LLM for querying. For simplicity and consistency, we utilized OpenAI’s GPT-4o-mini model via LangChain to generate predictions. The spaCy pipeline was employed for text tokenization and coreference resolution, ensuring accurate and consistent representation of character references throughout the analysis. Facebook AI Similarity Search (FAISS) was utilized to efficiently retrieve relevant context for the RAG system by indexing the text and performing similarity searches using Euclidean distance. This approach allowed the model to focus on the most relevant sections of the text, enhancing its ability to answer narrative based queries with greater accuracy.

## IV. Results

### A. Metrics

The performance of the model varied across the three novels tested. For *The Mysterious Affair at Styles*, the model successfully predicted all five plot components. In *The Murder of Roger Ackroyd*, three out of five components were correctly predicted, with the antagonist and climax chapter identified incorrectly. For *The Murder on the Links*, the model achieved a score of two out of five, with the victim, antagonist, and climax chapter being predicted incorrectly.

The addition of coreference resolution to the preprocessing phase had little to no impact on the model’s performance, suggesting that this step may not have been a significant factor in improving prediction accuracy for this task. We believe that the LLMs used were advanced enough to link these expressions to the correct character without the text being explicitly resolved. In terms of the model’s prediction capabilities, it performed well in identifying the protagonist, victim, and murder weapon across the novels. However, the predictions for the climax chapter were less accurate. This may be attributed to the fact that the timing and definition of the climax in a novel can be subjective and open to interpretation, which likely contributed to the model’s difficulty in consistently predicting this component.

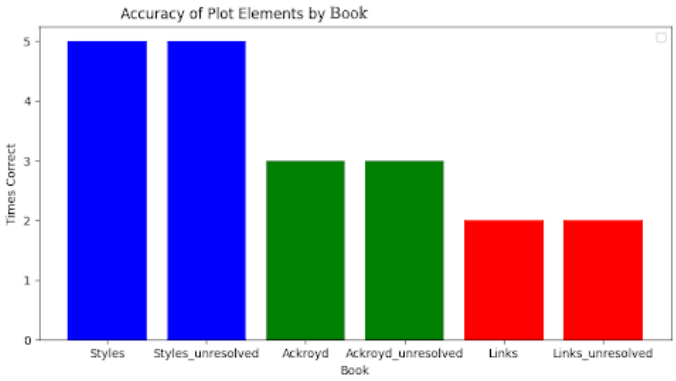


Fig. 1. Accuracy of Plot Elements by Book [4]

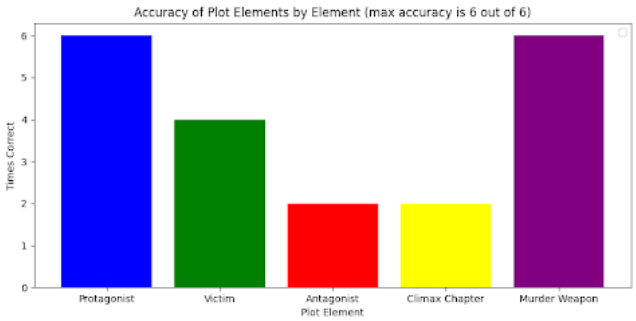


Fig. 2. Accuracy of Plot Elements by Element [4]

TABLE I
Book Analysis Results

| Book | Protagonist | Victim | Antagonist | Climax Chapter | Murder Weapon | Total Correct |
|---|---|---|---|---|---|---|
| **Styles** | Correct | Correct | Correct | Correct | Correct | 5/5 |
| **Ackroyd** | Correct | Correct | X | X | Correct | 3/5 |
| **Links** | Correct | X | X | X | Correct | 2/5 |

### B. Long Response Tests

Another test we performed was using long queries that asked for multiple plot components at the same time. We did this using the query ”Who is the victim of this novel, how

were they killed, who killed them, and why did they do it?". The results for this test varied based on the novel.

- *The Mysterious Affair at Styles*: The victim of the novel is Emily Inglethorp, who was killed by strychnine poisoning. The murderer is Alfred Inglethorp, who, along with Evelyn Howard, conspired to kill her to gain her money. Their motive was to induce the wealthy old lady to leave her money to them through a cleverly conceived crime.
- *The Murder of Roger Ackroyd*: The victim of the novel "The Murder of Roger Ackroyd" is Roger Ackroyd, who was killed with a dagger. The context does not provide enough information about who specifically killed him or the motive behind the murder.
- *The Murder on the Links*: The context does not provide enough information.

Unlike the previous test, there was a noticeable difference in the addition of coreference resolution. Five trials were run using both resolved and unresolved texts to determine the average accuracy of predictions, with resolved texts producing better accuracy for longer queries.

- *The Mysterious Affair at Styles*:
  - Resolved: Avg. 100% Accuracy
  - Unresolved: Avg. 87.5% Accuracy - Missing Inheritance Motive
- *The Murder of Roger Ackroyd*:
  - Resolved: Avg. 50% Accuracy - Missing killer, motive
  - Unresolved: Avg. 20% Accuracy - Missing weapon, killer, motive
- *The Murder on the Links*:
  - Resolved: Avg. 0% Accuracy
  - Unresolved: Avg. 0% Accuracy

These results showcase how coreference resolution can enhance predictive analysis for longer prompts. We believe that these results were achieved due to RAG struggling to retrieve all of the context needed to label each pronoun with the character associated with it in the documents it finds. This is due to RAG retrieving the same number of context chunks regardless of how long the prompt is. With coreference resolution, RAG already has the context it needs to associate the actions that take place in the relevant documents with the characters that perform them. By resolving character references, the system had all of the context it needed to make accurate connections, improving prediction accuracy.

## V. Conclusion

In conclusion, this study demonstrates how Natural Language Processing techniques, particularly text pre-processing and LangChain, can contribute to literary analysis. By focusing on critical elements such as the protagonist, antagonist, victim, climax chapter, and murder weapon, we explored the capacity of computational models to capture essential narrative components both in short, single component queries and long queries with multiple components. The results of using coreference resolution suggest that while LLMs may be able to handle some aspects of coreference resolution on their own, the additional preprocessing step enhances the system's ability to connect and retrieve relevant information for the aforementioned long prompts.

Looking ahead, we see several avenues for future research. First, refining the prompts could help improve prediction accuracy by making the queries more closely match the details of the plot and narrative. It would also be interesting to try out different models within LangChain, like LLaMA, Gemini, or other OpenAI models such as GPT-4o and o1-preview, to see how different model sizes and abilities impact the results. A broader comparison of predictions across novels beyond Agatha Christie's work would give a better sense of how well the model performs across different genres and authors. Diving into more specific predictor metrics would also help us understand the model's strengths and weaknesses, giving us clearer insights into how it picks up on different plot elements. Lastly, it would be interesting to explore how accurate the predictions could be if the model isn't given a list of characters or multi-word expressions. This could shed light on how well the model can learn and make connections on its own, without being explicitly guided.

## VI. Acknowledgment

We would like to thank the University of Tennessee, Knoxville, and the EECS Department for their support. Special thanks to Dr. Edmon Begoli and Gomathi Lakshmanan for their guidance and mentorship throughout this course on Natural Language Processing.

## References

[1] Christie, A. (1926). *The murder of Roger Ackroyd*. Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/61262
[2] Christie, A. (1920). *The mysterious affair at Styles*. Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/863
[3] Christie, A. (1923). *The murder on the links*. Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/61075
[4] Peruski, R. Coffey, N. Eden, T. Duff, W. (2024). *nlpproject* [GitHub repository]. GitHub. https://github.com/Silverasdf/nlpproject2
[5] Project Gutenberg. (n.d.). Project Gutenberg. https://www.gutenberg.org/