CSC8499 Individual Project: Enhancing Chatbot Interactivity through Personality Integration and Character Reviews

Vishnu Vardhan

MSc in Advanced Computer Science,
School of Computing, Newcastle University
s.v.v.sistla2@newcastle.ac.uk

Abstract. Chatbot is a computer program that simulates conversations with humans to provide efficient replies. The presence of chatbots has been increasing day by day in the fields of technology, medical, customer support and various other industries. However, most chatbots lack personality, resulting in generic and less engaging interactions. Previous research has shown that equipping dialogue agents with personalities can better engage end-users. That research focused primarily by profiling characters from novels using only dialogues. This leaves a gap in exploring the integration of additional external information for character profiling. The primary aim of this project is to enhance chatbot interactivity by integrating personality traits using both dialogues and character reviews, making them more engaging. This is achieved by training and testing various Large Language Models (LLMs), namely GPT-2 and Llama-2, to mimic a character from a novel, thereby embedding specific personality attributes. The findings show that for GPT-2, while adding external reviews enhanced some part's of models performance, there were drawbacks as well. External comments added to the complexity of GPT-2, with varying results in terms of model consistency and accuracy. In a similar way, Llama-2 showed enhanced comprehension with more context, despite at a cost of lower-quality output. These results suggest that while integrating external information can enhance model performance in some areas, it may require careful balancing to avoid trade-offs in output quality.

Keywords: Chatbots · Large Language Models (LLMs)· Personality Profiling · Character Reviews

declaration I declare that this dissertation represents my own work except where other-wise explicitly stated.

1 Introduction

Chatbots represent a powerful tool for human-computer interaction. A chatbot is a program that mimics and handles human conversation (whether written or spoken), enabling people to engage with digital devices as if they were interacting with an actual person [1]. Given that users naturally expect interactions with computers to mimic human dialogue, the development of chatbots has gained rapid momentum. Chatbots offer immediate responses to common queries, enhance human experience, and contribute to cost savings for companies. Research has found that the global spend over chatbots is forecasted to reach $78 billion by 2028 [11]. This shows the demand and future potential of the chatbot industry.

The rise of chatbots has been driven by advances in Artificial Intelligence and Natural Language Processing. These technologies enable understanding and generating human language with more accuracy, making chatbots valuable in customer service, personal assistance, and medical fields. However, many current chatbots are monotonous and lack the ability to engage in interactive conversations. This limitation is partly due to their inability to exhibit nuanced personality traits or adapt their responses to match specific character attributes. While they excel at answering frequent and straightforward queries, they fall short in providing a more human-like conversational experience. Studies by Nass and Reeves and their students in the computers as social actors' paradigm have concluded that people respond socially with computers when provided with appropriate social cues [20]. This suggests that enhancing chatbots with human-like attributes improves user engagement and satisfaction. Users are more likely to feel connected and engaged when interacting with dialogue agents that can exhibit humor and other personality traits.

The advent of Large Language Models (LLMs) tries to address this problem of lack of personality in dialogue agents. Large Language Models are machine learning models that can generate human language text by comprehending the information. They work by analyzing massive datasets of language [31]. By leveraging LLMs, such as GPT by OpenAI [28] or Llama by Meta [27], it is possible to imbue chatbots with more complex and distinct personality traits. Incorporating character-specific attributes into chatbot design could address current limitations by allowing more nuanced and engaging interactions. Previous research has shown that fine-tuning dialogue agents with personality profiles surpasses those trained without them. The existing workflow relies solely on comments in the books—dialogue utterances and surrounding text paragraphs. [35].

In this research, We will explore the feasibility of using dialogue information from novels and character reviews from various websites to fine-tune LLMs. This helps to create chatbots that embody specific character traits and respond consistently with established personality. This is verified by comparing the performance of chatbots trained with two datasets:

(i) One consisting only of dialogues from the novel, and
(ii) Another incorporating both dialogues and external character information.

This comparison will provide insights into whether adding external information enhances chatbot responses.

1.1 Aims and Objectives

This research is aimed at fine-tuning of two Large Language Models (LLMs)-namely, GPT-2 and Llama-2, based on the personality traits of a character from a novel and also by adding the character reviews. Subsequently, we will evaluate the performance of these fine-tuned models, considering various personality traits used during training. Detailed Objectives of this project include:

1. **Literature Review and Background Research:** Conduct a comprehensive literature review on the current state of LLMs and chatbots. Understanding the advancements in these fields will provide a solid foundation for research.
2. **Dataset Preparation and Enrichment:** a. Extract dialogues related to a specific character from a novel. b. Integrate character reviews-gathered from sources such as literary critiques or fan forums with the dialogues dataset. This enriched dataset will provide additional context and personality traits for fine-tuning the LLMs. c. Curate a dataset containing both character dialogues and external character reviews.
3. **Fine-Tuning and Model Training:** Fine-Tune the LLMs using the dialogue-only dataset and the enriched dataset. Align the models with characters unique speech patterns, and conversational style.
4. **Performance Evaluation and Analysis:** Evaluate the fine-tuned models using standard metrics, including ROUGE-1, ROUGE-2, ROUGE-L and perplexity. Analyse the results to determine the impact of incorporating external character information.
5. **Document the results and findings:** Compile the research findings into a comprehensive report, detailing the methodologies, results and implications. This document will provide valuable insights for future research and development in chatbot technology.

1.2 Structure of Dissertation

The structure of the dissertation is as follows: Section 2 deep dives into the research on the current state of chatbots and LLMs and their background, establishing a theoretical foundation. Section 3 discusses the proposed approaches and implementation of these methods such as data collection, preparation of dialogues only and character reviews datasets, and pre-processing them. It also emphasizes the various challenges involved in the preparation of final datasets. Section 4 consists of results obtained after the model training, evaluating the various scores and their implications. Section 5 will conclude the paper with key findings and the future related work that can be carried out further.

2 Background and Related Work

2.1 History of Chatbots

The history of chatbots is quite interesting. Early chatbots, including ELIZA, laid the foundation for more sophisticated conversational agents. Joseph Weizenbaum created the initial chatbot, ELIZA, as a rule-based artificial intelligence model to simulate the behavior of a Rogerian therapist [36]. Utilizing pattern matching and substitution technology, it showcased how human-computer interaction can be utilized in Natural Language Processing (NLP) [36]. Over the years, chatbots have transformed from rule-based systems to machine learning models, greatly enhancing their effectiveness.

2.2 Large Language Models

Large Language Models (LLMs) like GPT by OpenAI [16] and Llama by Meta [27], have become pivotal in the development of conversational agents. GPT-2 featuring 1 billion parameters, was one of the early models that has demonstrated significant capabilities in generating coherent and contextually appropriate text across various scenarios [37]. It is further enhanced with GPT-4 being trained on 1.7 trillion parameters and a max content length of 32,768 tokens [38]. This enhancement has increased its accuracy and efficiency in generating text, making it one of the most powerful language models currently available. GPT-4's advanced capabilities include better handling of context, more nuanced text generation, image creation and improved performance in wider range of applications. Meta's Llama has also made significant contributions to the field. Llama-2 is trained on 7 billion parameters, on par with GPT-4. It excels in generating content for specific domains, leveraging its efficient architecture to produce high-quality outputs with lower computational requirements. GPT-4 has higher accuracy in generating text with more efficiency while Llama-2 has its strengths in generating content in more specialized application areas [39].

The evolution of LLMs demonstrate how these models have become increasingly well-versed and efficient, making them crucial components in future in various industries. They are suitable for their tasks in various fields such as healthcare, finance and legal services where domain specific knowledge and precision are paramount. Their ability to understand complex queries and provide contextually relevant responses has opened many possibilities in automation, content creation and customer service. As these models evolve, dependence on them in human computer interaction and transforming digital communication will only grow.

Fine-Tuning pre-trained LLMs for specific tasks or domains is a common practice to improve their performance for specialized tasks. Models like GPT has already been trained on large sets of data with billions of parameters. Fine-tuning involves tweaking the model with giving more context about our specific requirements to align them with the required output. In the context of this project, it helps the models adopt the conversational style or speech patterns with respect to the specific characters or personas. In research by OpenAI, Fine-tuned models were preferred by human labellers over the base GPT-2 model in 88% and 86% of cases for sentiment and descriptive tasks, respectively [6]. Fine-Tuning a model is a complex process and involves various strategies and techniques to be adopted for improved results. The table below describes some of the popular methods of fine-tuning and their merits/demerits.

Domain Adoption	**Description**	**Advantages**	**Challenges**
Supervised Learning	Training on labelled data to learn specific tasks.	Clear information from labelled data; improves task-specific performance.	Requires huge amounts of labelled data; can be time-consuming to annotate.
Transfer Learning	Leveraging knowledge from pre-trained models to new tasks.	Reduces data and computational requirements; accelerates training.	May not fully capture domain-specific nuances; potential for negative transfer.
Domain Adaptation	Adjusting model parameters to align with a specific domain.	Enhances relevance and accuracy in specialized applications.	Requires careful tuning; can be overfitted if not tuned properly.

Table 1. Overview of Learning Techniques

2.3 Chatbots with Personality Traits

Recent research has begun to explore the possibility of adding recognizable personality traits, thereby transforming them into more engaging and relatable conversational partners [2]. This involves embedding chatbots with human-like attributes, making them friendly, humorous, or professional in their interactions. This is very important for socially-oriented chatbots, whereby human-like attributes may mean increasing user engagement without creating unrealistic expectations about its capabilities. [3]. Adding personality to chatbots involves several strategies. These include:

Pre-defined Personality Profiles: Psychological framerworks such as the Big Five personality traits (openness, conscientiousness, agreeableness, extraversion and neuroticism) [4] are used to construct the profiles. These personality traits can be used to create dialogue templates, and responses can be made to fit the selected traits.

Dynamic Adaption: The personality of the chatbot can be constantly changed by machine learning algorithms in response to continuous user interactions. Depending on the user's preferences and engagement style, the replies can change over time [5]. Reward-based strategies like reinforcement learning can be applied to improve the replies.

Context-Aware Responses: We can detect the user's mood and modify the tone accordingly using sentiment analysis, offering cheerful comments when the user is joyful and sympathetic responses when they are sad.

Multimodal Interactions: Inputs other than text can be used such as voice and facial expressions. Virtual assistants and customer support bots may find this especially helpful.

While existing methods have significantly advanced the capabilities of chatbots, several gaps remain. Most approaches rely heavily on either rule-based systems or traditional training data without incorporating additional external sources of varied information. Current strategies often fail to fully utilize the potential of external information and other supplementary data, which can offer richer, more nuanced profiles for chatbots.

This project particularly chooses the strategy of Pre-defined Personality Profiles where dialogues of a character are extracted from the novel, a spectrum is built based on the dialogues and Character Reviews, which are then used to train the LLMs.

3 Methods and Experiments

3.1 Tools and Technologies

Python: Python was chosen as the primary language for this project. It has a large community support in machine learning domain. It is perfect for quick development and experimentation because of its readability and simplicity of usage. It contains a huge number of libraries and frameworks that help in a variety of tasks.[17].

SpaCy: SpaCy is an open-source Python module for activities related to Natural Language Processing (NLP). It is known for how well it performs in large-scale information extraction and how managing challenging text processing jobs. In this project, SpaCy is used for identifying and extracting key sentences or phrases from the original text, giving us the top sentences [18].

Scikit-learn: Scikit-learn is used to compute the cosine similarity metric. This Python machine learning package is available as open-source and includes numerous algorithms for tasks like regression, classification, and clustering [19].

Google Colab: Google Colab is a cloud-based Jupyter notebook development platform for the execution of Python code in a web browser [21]. It provides free access to computational resources, including GPUs,for training machine learning models.

Tensorflow: Tensorflow is free and open-source machine learning library. It has advanced capabilities like deep learning algorithms, scoring algorithms and can be used for a wide range of machine learning tasks [34].

3.2 Methodology

The following section discusses about the approaches adopted for this project. It discusses in detail about the system design, the algorithms and the methods used to achieve the task of training the generative models.

The first step is extraction of utterance comments and non utterance comments of the character from the novel. The methods and code described in the paper "Personality Profiling for Literary Character Dialogue Agents with Human Level Attributes" [35] will be followed for this step. Context will be added to the extracted dialogue pairs. A character spectrum is generated for the characters - using the utterance and non utterance comments from the novel. A dataset, which we call '*dialogues dataset*' from here on till the rest of the paper is generated using the dialogue-pairs and the traits.

Then the project moves on integrating external character information to the dialogues extracted. New character spectrum is generated with the dialogue-pairs and external data combined. This dataset, with dialogue-pairs, character reviews summary and newly generated traits will be referred to as '*character reviews dataset*'.

A comparison is done to evaluate the effectiveness of integrating external character information with the dialogue-pairs. Language models trained with *dialogues dataset* against those trained with *character reviews dataset* will be compared and the results are evaluated. The methodology employed in this research project is illustrated in the figure 1. The project is structured into three distinct phases:

1. Dialogues Extraction and Spectrum Construction
2. External Reviews Processing
3. Fine-tuning the Language Models

3.2.1 Dialogues Extraction and Spectrum Construction

Extracting Dialogues: The first step in the process involves extracting dialogues and constructing a personality spectrum from a source text, such as a novel. For this study, we utilize the book "The American" authored by Henry James, which is publicly available through the Project Gutenberg Library [7]. Our focus is on the main character of the novel *Christopher Newman*. Specifically, we are interested in extracting both the direct dialogues of the character and the surrounding non-utterance related comments that provide contextual insights into his personality.

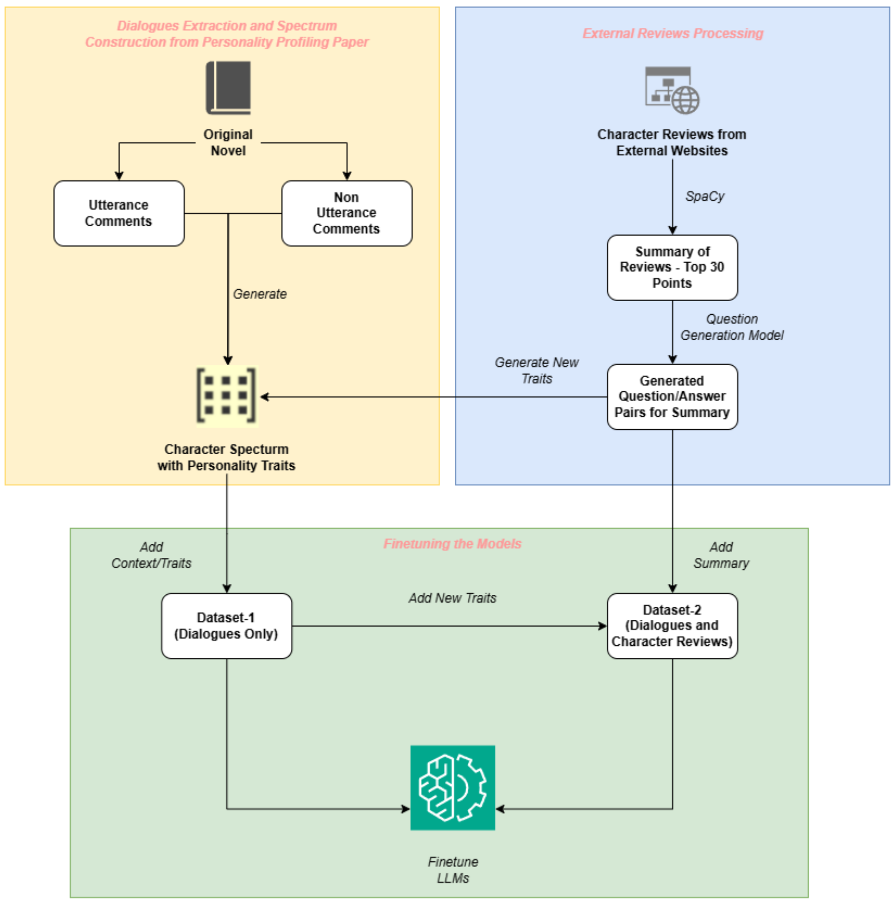


Fig. 1. The workflow consists of three phases - i) Dialogues Extraction/Spectrum Construction ii) External Reviews Processing iii) Finetuning the LLMs

To achieve this, we adopt a method that is closely aligned with the approach outlined by Rusnachenko et al. in their work on personality profiling for literary characters [35]. In order to minimize errors in the data extraction process, Named Entity Recognition (NER) is used to ensure that the dialogue pairings are appropriately attributed to the character. This method employs Natural Language Processing (NLP) techniques to accurately identify and extract the character's dialogues from the text. Additionally, non-utterance comments—narrative descriptions and authorial remarks that pertain to the character's actions and thoughts are also extracted. These comments are invaluable as they offer deeper contextual understanding and are integral to constructing a holistic personality profile.

Constructing the Character Spectrums: Once the dialogues and non-utterance comments are extracted, we move to the next phase: constructing a personality spectrum for *Christopher Newman*. The personality spectrum is a multidimensional representation of the character's traits, derived from both the explicit dialogues and the implicit non-utterance comments. The construction of this spectrum is done by lexicon-based approaches, which are effective in capturing nuanced personality traits. Specifically, we utilize FCPlexicon adjective-pairs [30], that reflect key personality dimensions, such as "introverted/extroverted" or "optimistic/pessimistic." These pairs are the opposite adjectives that allow us to quantify the degree to which certain traits are exhibited by the character.

This approach is consistent with the methodologies employed in the original work by Rusnachenko et al., where the authors combined TF-IDF and lexicon-based techniques to create detailed personality profiles. In our study, we place greater emphasis on the lexicon-based technique for constructing the personality spectrum, as it provides a more structured framework for analyzing the character's traits.

By integrating both the dialogues and the non-utterance comments, our method ensures a comprehensive analysis of the character's personality. This approach captures the subtler aspects of the personality reflected in the narrative context. For readers interested in a deeper understanding of the specific algorithms and techniques used in this process, we refer them to the original paper by Rusnachenko et al. [35], which provides a thorough explanation of the methods employed.

The process described above fetches us with *dialogues dataset* and a character spectrum which will be referred to as *dialogues spectrum*. The *dialogues dataset* obtained is in the format of Question-Response pairs with the question being utterance of any other character rather than the target character. Response is the utterance from the target character.

Adding the Context: To further enrich this dataset, context is added to the dialogues. Context will allow the model to access more information, thus making accurate predictions. But the right amount of context length is the key. Too short context cannot provide supplementary knowledge whereas too long will introduce excessive amounts of noise [9]. Context is added by identifying and including surrounding narrative elements, speaker turns and preceding dialogue lines when extracting dialogues for model training. The additional layer of information allows the generative model to learn not only from the direct responses of the target character but also from the context in which these responses occur.

3.2.2 External Reviews Processing

Adding External Reviews: The second phase is to integrate external reviews into the dataset. Character reviews are sourced from multiple online platforms, resulting in a comprehensive corpus that describes Christopher Newman in detail. The purpose of these reviews is to gain a broader understanding of the character, which can be integrated with the dialogue data to improve the training of generative models. Each source offers unique insights into Newman's character, contributing to a well-rounded profile for more accurate dialogue generation.

CliffsNotes - Christopher Newman Character Analysis : CliffsNotes provides a detailed analysis of Christopher Newman, highlighting his physical and moral attributes. It describes Newman as a wealthy and ambitious American who is somewhat naive and inexperienced in European society. The analysis covers his personal growth and the challenges he faces throughout the novel[10].

SparkNotes - Christopher Newman : SparkNotes offers a comprehensive overview of Christopher Newman's character, focusing on his motivations, development, and the complexities of his personality. This review emphasizes Newman's idealism and ambition. It describes about his encounters in his pursuit of happiness, providing insights into his interactions with other characters [12].

BakerP2004 - Character Analysis of Christopher Newman : BakerP2004 provides a detailed analysis focusing on Newman's psychological and emotional attributes. The blog post discusses his internal conflicts, aspirations, and the impact of his American background on his interactions in Europe [13].

Shmoop - Christopher Newman : Shmoop's analysis of Christopher Newman provides a concise look at his character. It covers his key traits, motivations, and the ways in which he reflects the themes of the novel.[14].

CliffsNotes - Summary and Analysis of Chapter I : Another CliffNotes review offers a summary and analysis of the first chapter of *The American*, introducing Christopher Newman's background, his ambitions, and the setting of the novel. [15].

SpaCy is used to extract the top-20 sentences from the above reviews using Cosine Similarity. This summarizing technique is used to reduce the noise in the dataset and remove repetitions about the character profile in different reviews.

Cosine Similarity: Cosine Similarity is widely used metric in text analysis and information retrieval. It is particularly used for applications that utilize small amounts data, such as word documents, transactions and recommendation systems [8], where each dimension in the vector space corresponds to a specific trait or a term.

It measures the similarity between the two vectors in direction or orientation of vectors ignoring differences in their magnitude or scale [8]. The similarity is measured by the cosine of angle between the two vectors. It is calculated as follows:

$$\text{Cosine Similarity (A,B)} = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

Where:

- A and B are the vectors representing two characters.
- $A \cdot B$ is the dot product of the vectors.
- $\|A\|$ and $\|B\|$ are the magnitudes of the vectors.

Converting Summarized data to Question-Answer Pairs: Next step involves transforming the character summaries in the format of Question-Answer pairs. Transforming the summarized external reviews into QA pairs ensures consistency with our existing dataset. Each line of the summary serves as an answer. Manually generating corresponding questions would be exceedingly time-consuming. To streamline this process, we employ a pre-trained T5 (Text-To-Text Transfer Transformer) model, specifically the "valhalla/t5-base-qg-hl" [22] variant, which is proficient at transforming text sequences into various forms, including the generation of questions from given text snippets. These QA pairs are added to the original *dialogues dataset* giving us *character reviews dataset.*

Finally, A new character spectrum, referred to as *Character Review Spectrum* is constructed using these summaries by the same techniques described in section 3.2.1. We end up with 4 datasets after the process of phases 1 and 2 - *dialogues dataset*, *dialogues spectrum*, *character reviews dataset*, *character reviews spectrum.*

3.2.3 Fine-Tuning the Models

The goal of the third and final phase is the fine-tuning of Large Language Models (LLMs) using the datasets generated in the earlier phases. We need to refine the models' capabilities to generate dialogues that are both contextually accurate and consistent with the character's personality traits. For this phase, two state-of-the-art LLMs are employed: GPT-2 and Llama 2.

The fine-tuning process begins by organizing the data into two distinct datasets. The first dataset, *dialogues dataset*, is composed of dialogues directly extracted from the original novel, enhanced with contextual information and personality traits from *dialogues spectrum* produced in Phase 1. This dataset focuses exclusively on dialogues, ensuring that the model trains on generating precise and character-consistent spoken text.

The second dataset, *character reviews dataset*, integrates both the dialogues and the character reviews processed in Phase 2. This dataset is further enriched with the *character reviews spectrum*, which contains the broader context and additional character traits derived from external analyses. By including these external perspectives, the dataset broadens the model's understanding of the character, allowing it to generate dialogues that consists of both the original narrative and a diverse range of character interpretations.

The actual fine-tuning is conducted using two models: GPT-2 and Llama 2. Each model is fine-tuned separately using the previously generated datasets:

GPT-2 Fine-tuning: GPT-2, known for its robust language generation capabilities [16], is fine-tuned on both datasets. GPT-2's strengths lies in generating fluent and contextually rich text, while ensuring that the dialogues remain faithful to the character's personality as established by both the novel and external reviews.

Llama-2 Fine-tuning: Llama-2 is a more recent model with enhanced capacity for contextual understanding and nuanced text generation [23]. It is similarly fine-tuned on the both datasets. Llama-2's architecture allows it to capture more complex relationships within the data, making it particularly well-suited for generating dialogues that reflect subtle character traits and evolving personality dynamics.

Post fine-tuning, both models are rigorously evaluated to ensure that the generated dialogues align with the original text's context while also reflecting the expanded character profiles. This involves comparing the outputs from each model against the original and augmented datasets, assessing their ability to capture character nuances, and ensuring consistency in the dialogues.

The final outputs from GPT-2 and Llama 2 are compared to determine which model produces dialogues that best meet the project's objectives.

3.3 Experiments

3.3.1 Overview

This section details the experiments conducted to evaluate the performance of the proposed model on generating meaningful and contextually appropriate responses in dialogue systems. Two main experiments were carried out: the first using a baseline model trained on a dialogue dataset, and the second utilizing an enhanced model with additional data from external character reviews. Third experiment is fine-tuning the models on these datasets.

3.3.2 Experimental Setup The experiments were conducted using Google Colab, utilizing the Python programming language and libraries such as TensorFlow and Hugging Face's Transformers [24]. A high-performance GPU was employed to expedite training and evaluation.

The model was trained and evaluated on the following datasets:

- **Dialogue Dataset:** A corpus of dialogues from the novel, forming the base for the model's conversational abilities.
- **Dialogue Spectrum:** A dictionary of Newman's top personality traits and corresponding weights, derived from the *dialogues dataset*.
- **Character Reviews Dataset:** An external dataset containing reviews and descriptions of Christopher Newman, enhancing the model's ability to generate contextually rich responses.
- **Character Reviews Spectrum:** A dictionary of Newman's top personality traits and corresponding weights, based on the *character reviews dataset*.

Hyper parameter tuning on parameters such as learning rate, batch size, and number of epochs were done and optimized based on preliminary trials.

3.3.3 Experiment 1: Baseline Performance with Dialogue Dataset

Objective: To establish a baseline for the model's performance using only the dialogue dataset.

Procedure:

Dialogues Extraction: The initial step involved extracting utterances and constructing a spectrum. The book *"The American"* by Henry James was selected, with a focus on the main character, Christopher Newman. Dialogues from 1,000 books, including the target book, were processed. A character map was created, assigning each character a specific code to minimize noise from different character names.

```
UNKN-X: And you mean to carry my little picture away over there?
177_0: Oh, I mean to buy a great many pictures—_beaucoup, beaucoup_,

UNKN-X: The honor is not less for me, [USEP] for I am sure monsieur has a great deal of taste.
177_0: But you must give me your card, [USEP] your card, you know.
```

Fig. 2. Dialogues extracted from the novel as Question-Answer pairs. '177_0' is the character code for Christopher Newman.

Dialogues spoken by characters other than Christopher Newman were marked as 'UNKN-X', serving as placeholders. This approach prevents the model from being overwhelmed by numerous character details. '[USEP]' was used as a placeholder for context, to be added in later stages.

The data format used for training is shown in Figure 2. Each "Question" corresponds to a dialogue uttered by a character other than Christopher Newman, while the "Answer" is his response, labeled as '177_0'.

Spectrum Generation: The next step was spectrum generation, which involved creating a spectrum of personality traits with corresponding weights. This profiling was adapted from the FCPlexicon [30], a lexicon containing 264 adjective pairs representing antonyms and polarities of character traits. The process involved two stages:

- **Stage 1:** Comments from each character were represented as sequences of unigrams. Relevant adjective pairs were identified within these comments using the FCPlexicon. Spectrum values were calculated as the average values of low and high polarity entries for each adjective pair, resulting in a matrix A of size $n \times m$, where n is the number of characters, and m is the number of adjective pairs.
- **Stage 2:** To establish deeper connections and rank similarities between characters, both character and personality traits were transformed into the same latent space using latent factor modeling. Matrix A was decomposed into A_c (latent factors for characters) and A_i (latent factors for attributes). The dot product $A_c \cdot A_i$ approximates A, using the Conjugate Gradient Method.

The spectrum in figure 3 is character spectrum generated for *Christopher Newman*. Similarly, the spectrum's are generated for all other characters in the 1000 books we have selected.

Similar Characters: After generating the spectrums for the characters in all the books we have selected, we end up with a json file containing the traits and weights of their personalities. We do this because we want to increase the size of the dataset by adding dialogues of characters similar to Christopher Newman. The size of the dataset can be adjusted without affecting the quality by calculating the similarities between the spectrum generated above and including the dialogues of the top 4 characters similar to Newman along with the

```json
{
    "name": "177_0",
    "prompts": [
        "serious", "cold", "bright", "real", "human",
        "angry", "fast", "beautiful", "glad", "open",
        "hard", "quiet", "short", "happy", "old",
        "poor", "kind"
    ],
    "weights": [
        1.0, 1.0, 1.0, 1.0, -1.0, -1.0, -1.0, -0.867,
        0.75, -0.75, -0.75, -0.75, 0.714, 0.667, 0.594,
        0.571, 0.556
    ]
}
```


Fig. 3. Character Spectrum of Christopher Newman with associated weights

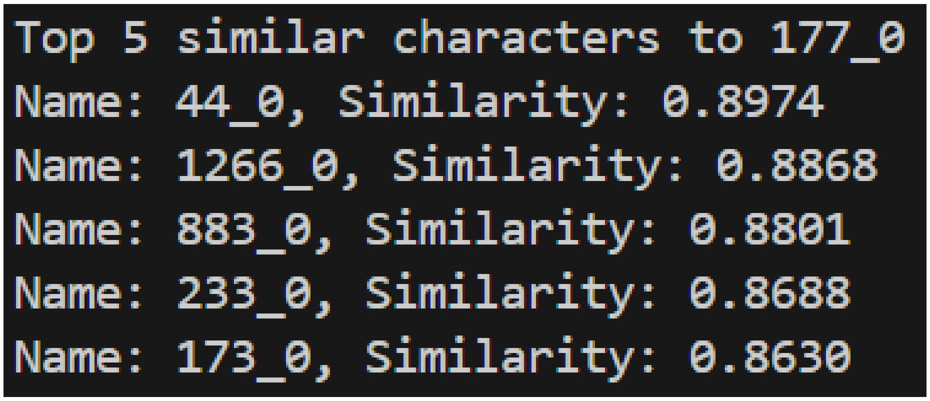
Christopher Newman's dialogues. This ensures we have sufficient data to train the model without directly impacting the style of the speaker, as they possess similar traits.

To compare the characters, we first need to convert their prompts and weights to numerical vectors. This process involves:

- *Extracting Unique Prompts:* Compile a list of all unique prompts from the dataset. This serves as a basis for creating a consistent vector representation for each character.
- *Vector Creation:* For each character, create a vector where each element corresponds to a prompt from the unique prompt list. The value of each element is the weight associated with that prompt for the character. If a prompt is not present for a character, the corresponding vector element is zero.

The result is a value between 1 and -1, where 1 indicates the vectors are perfectly aligned (i.e., they point in the same direction), 0 indicates orthogonality (no similarity) and -1 indicates the vectors are diametrically opposed (i.e., they point in opposite directions).

Figure 4 shows the similarities between different characters. The top 4 similar characters to Newman are 1266_0 - Ruth Thorne, 883_0 - Bella Wilfer, 233_0 - Carrie Madenda, and 173_0 - Katherine Hyde. These are all characters from different novels. Figure 5 depicts the similar traits between these characters in the bar graph.



```
Top 5 similar characters to 177_0
Name: 44_0, Similarity: 0.8974
Name: 1266_0, Similarity: 0.8868
Name: 883_0, Similarity: 0.8801
Name: 233_0, Similarity: 0.8688
Name: 173_0, Similarity: 0.8630
```

Fig. 4. Top-5 similar characters with the scores

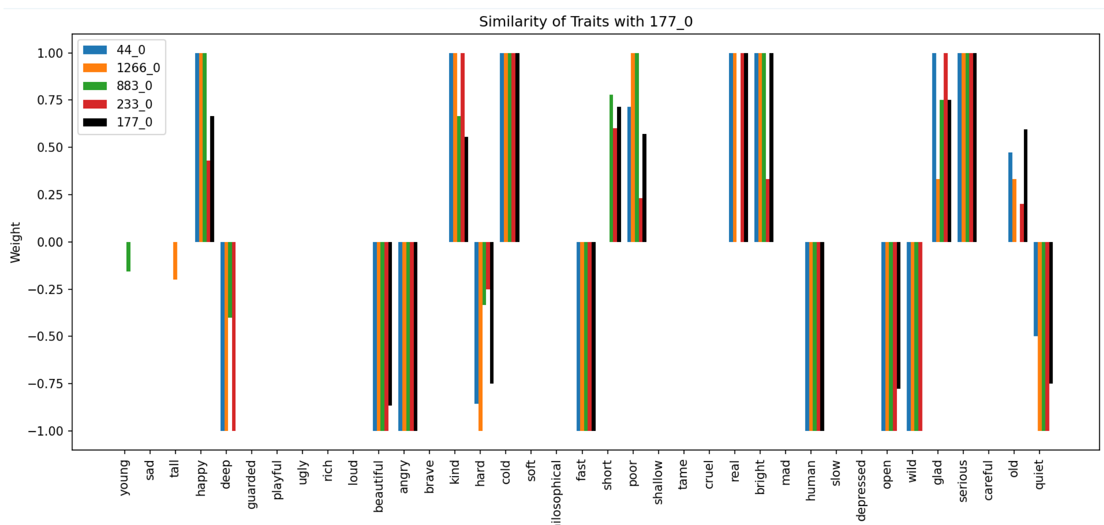


Fig. 5. Bar graph showing matching traits between Newman and other similar characters from different books

Adding Context: After adding all the dialogues, both Newman's and similar characters, final step is to add the context to these dialogue pairs. Adding context makes the model understand the conversational dynamics and the specific roles the characters play within those dynamics. This involves including surrounding narrative elements, speaker turns and preceding dialogue lines when extracting dialogues for model training. To augment the dialogues with context, we implement the following steps:

Extracting Search Text: For each dialogue line, the search text is identified. This involves stripping away placeholders or tags to isolate the actual spoken text. The search text forms the basis of the context text in the original book.

Context Extraction: Using regular expressions, the method searches for the exact match of the search text within the book. The pattern matching is designed to be robust, accommodating variations in the text formatting and case sensitivity.

Handling Missing Contexts: In cases where search text is not found within the book, a fallback mechanism retains the original dialogue without additional context. This ensures that no data is lost even if the context extraction fails.

Formatting the Output: The restored context is then formatted appropriately. The 'UNKN-X' or the characters other than the target character are updated with their corresponding context, and character responses are preserved with their original structure but enriched with context.

The above process yields us with dialogues in 'Question-Answer' pairs along with context of the dialogues. This step concludes the Experiment 1 with *dialogues dataset* as our final output.

3.3.4 Experiment 2: Enhanced Model with Character Review Dataset

Objective: To enrich the dialogues dataset with character reviews information from various websites.

Procedure:

We already have the *dialogues dataset* and *dialogues spectrum* from the Experiment 1. To continue with the Experiment 2, We have obtained character reviews from various websites describing *Christopher Newman*.

Summarization: SpaCy library is used to select the top-20 sentences from the external reviews. Below table represents the comparison between original and summary text. '+' implies percentage increase and '**-**' implies percentage reduction.

Metric	**Original Text**	**Summary Text**	**Percentage Change**
Number of Sentences	180	20	-88.89%
Average Sentence Length	105.07	176.60	68.1%
Maximum Sentence Length	295	272	-7.8%
Minimum Sentence Length	17	33	94.1%

Table 2. Comparison of Text Length and Sentence Counts

The summarization process achieved a reduction in number of sentences from 180 to 20, representing an 88.89% reduction. Similarly, The average sentence length in the summary is longer from 105 characters in a sentence to 177, reflecting the combination of information from multiple original sentences. Maximimum sentence length has been decreased slightly by 7.8%, making the sentences simpler. But Minimum Sentence length is increased by 94.1%. This helps in capturing more information in less number of sentences.

The sentence similarity scores, depicted in the figure 6, provide insights into the summary's closeness to the original text. Most similarity scores range from 0.8 to 1.0, indicating that the summarized sentences closely reflect the content of the original sentences. High similarity scores suggest that the summarizing retains the original meaning and context effectively.

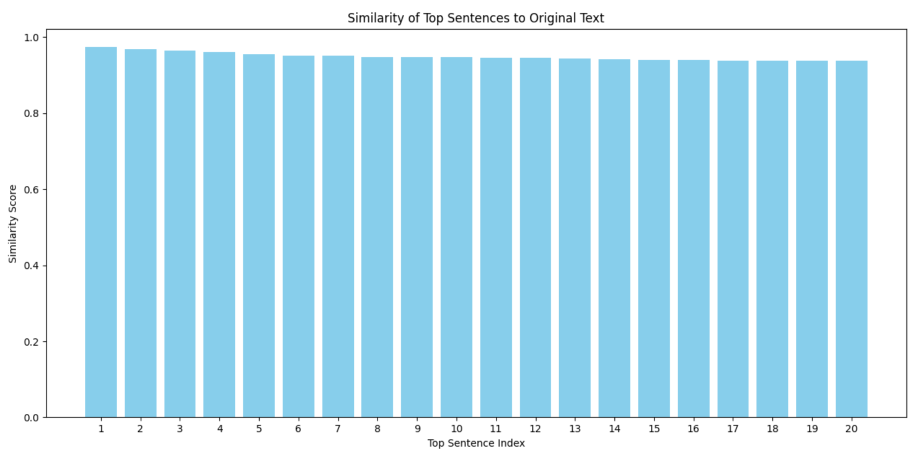


Fig. 6. Graph of Sentence Similarity Scores

Question-Generating Model: Transforming the summarized external reviews into Question-Answer pairs ensures consistency with our existing dataset. Each line of the summary serves as an answer. we employ a pre-trained T5 (Text-To-Text Transfer Transformer) model, specifically the "valhalla/t5-base-qg-hl" variant to generate questions to our top-20 sentences. The question generation process involves the following key steps:

Initialization: The T5 model and its corresponding tokenizer are initialized. The tokenizer encodes the text into a format suitable for the model, while the model generates the questions.

Input Preparation: Each sentence from the summarized character reviews is individually highlighted and prepared as input. The input format for the model includes a specific prefix `"highlight: "` followed by the sentence and the end-of-sequence token `</s>`. This format tells the model to treat the highlighted sentence as the context for generating a question.

Encoding: The tokenizer encodes the prepared input into a sequence of tokens that the T5 model can process. This tokenized sequence is then converted into tensor format suitable for model input.

Question Generation: The T5 model generates a question based on the encoded input. The model uses its pre-trained parameters to predict a sequence of tokens that form a coherent question related to the highlighted sentence.

Decoding: The generated token sequence is decoded back into human-readable text using the tokenizer. The resulting text is the question that corresponds to the highlighted sentence.

Formation of QA Pairs: Each generated question is paired with the original sentence from the character reviews, forming a QA pair. These QA pairs are added to our original *dialogues dataset*, giving us the *character reviews dataset*.

Calculating updated traits: We also need to calculate a new spectrum of personality traits after adding the summary. We use the same process as described above in Experiment 1, but in addition to utterances and non utterance comments, we also add the summary of external reviews. New set of traits - *Character Reviews Traits* are obtained.

Table 3 presents a side-by-side comparison of the character traits and their respective weights before and after the inclusion of external comments.

Several traits retained their weights, indicating that external comments did not change their significance. This shows the dialogues and character related information in novel itself captures most of the traits perfectly. These traits include *serious, cold, bright, real, glad, happy, beautiful, quiet, fast, angry,* and *human*. The trait *kind* increased from 0.556 to 0.692, suggesting that external comments contributed to a more favorable view of this trait. Other traits, such as *hard* and *open*, saw slight increases in their negative weights, from -0.75 to -0.8, reflecting a more pronounced negative perception. The trait *short* decreased from 0.714 to 0.333, indicating a diminished perception after the addition of external comments. Similarly, *poor* slightly decreased from 0.571 to 0.529, suggesting a small reduction in its positive perception. The comparison demonstrates that though external comments cannot change the traits significantly, they can influence the perception of certain traits. While some traits remained unaffected, others experienced notable changes in their assigned weights, highlighting the importance of external inputs in shaping character evaluations.

Prompt	Weight Before	Weight After
serious	1.0	1.0
cold	1.0	1.0
bright	1.0	1.0
real	1.0	1.0
glad	0.75	0.75
kind	0.556	0.692
happy	0.667	0.667
old	0.594	0.594
poor	0.571	0.529
short	0.714	0.333
beautiful	-0.867	-0.867
hard	-0.75	-0.8
open	-0.75	-0.8
quiet	-0.75	-0.75
fast	-1.0	-1.0
angry	-1.0	-1.0
human	-1.0	-1.0

Table 3. Comparison of Weights for Character Traits Before and After Adding External Comments

3.3.3 Experiment 3: Fine-tuning the Models

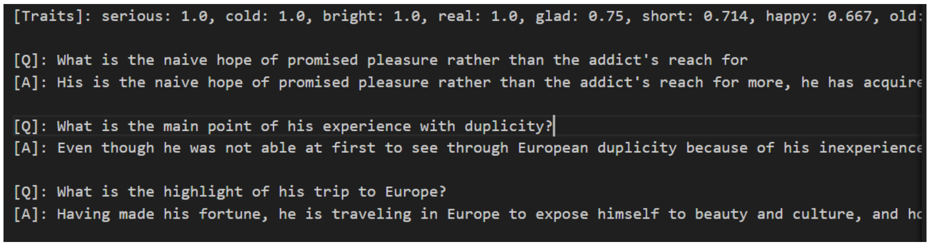
Objective: To fine-tune the large language models using the prepared datasets and evaluate performance in generating contextually relevant responses.

Procedure:

Fine-Tuning GPT-2: The GPT-2 model was selected for its proven capability in handling generative tasks involving natural language. We chose the huggingface [24] GPT-2 variant[25] balancing computational efficiency with model complexity. The model was initialized with weights pre-trained on a large corpus of internet text, allowing it to generate coherent and contextually relevant language.

To better suit the task's question-answer format, special tokens `[Q]` and `[A]` were introduced. These tokens were added to the tokenizer's vocabulary, ensuring that the model could distinguish between questions and answers effectively during both training and inference stages. *Dialogues spectrum* was added to the dataset for the model to understand better about the character's traits.

Two datasets - *dialogues dataset* and *character reviews dataset* were used to fine-tune the model. The datasets were formatted and split into training (80%) and testing (20%) subsets to validate the model's performance effectively. Given the importance of maintaining context across longer texts, a block size of 128 tokens was chosen for the text dataset, ensuring that the model captures sufficient



```
[Traits]: serious: 1.0, cold: 1.0, bright: 1.0, real: 1.0, glad: 0.75, short: 0.714, happy: 0.667, old:

[Q]: What is the naive hope of promised pleasure rather than the addict's reach for
[A]: His is the naive hope of promised pleasure rather than the addict's reach for more, he has acquire

[Q]: What is the main point of his experience with duplicity?
[A]: Even though he was not able at first to see through European duplicity because of his inexperience

[Q]: What is the highlight of his trip to Europe?
[A]: Having made his fortune, he is traveling in Europe to expose himself to beauty and culture, and hc
```

Fig. 7. Final dataset for training a GPT-2 model

context during training. Figure 7 depicts the format of datasets. The difference between *dialogues dataset* and *character reviews dataset* is the first line of *Traits* differ - *dialogues spectrum* for *dialogues dataset* and *character reviews spectrum* for *character reviews dataset* respectively. Also additional Question-Answer pairs are present in *character reviews dataset*.

Fine-tuning was performed using a `Trainer` class, configured with the following key parameters:

– **Epochs**: The model was trained for 10 epochs to ensure convergence while avoiding overfitting.
– **Batch Size**: A per-device batch size of 4 was employed, balancing between computational constraints and model performance.
– **Learning Rate and Optimization**: The learning rate was carefully chosen to ensure steady progress.

Post-training, the model was evaluated using perplexity, a standard metric for language models, and rouge scores. Also, the model's generative capabilities were tested using predefined prompts, showcasing its ability to produce coherent and relevant answers. More discussion about the results obtained will be done in Section 4.

Fine-Tuning LLaMA-2: For more advanced tasks, Llama-2, a larger and more complex model, was selected. We used the huggingface's `NousResearch\Llama-2-7b-chat-hf` [26]. It is optimized for dialogue use cases and was pre-trained on 7 billion parameters. To manage the increased computational load, the model was fine-tuned using 4-bit quantization via the `BitsAndBytesConfig`, which allowed for efficient training without significant loss in model accuracy.

Llama-2 takes slightly different special tokens like `[human]` and `[assistant]` instead of `[Q]` and `[A]` in GPT-2. The same datasets used for GPT-2 were reformatted and split similarly. Tokenization was handled using the `AutoTokenizer`, which ensured consistency with the model's pre-trained configuration. This enabled the model to differentiate between questions and answers effectively during the training process. The padding and truncation strategies were adjusted to align with the model's architecture.

```
### Human: What is the naive hope of promised pleasure rather than the addict's reach for
### Assistant: His is the naive hope of promised pleasure rather than the addict's reach for more, he has acquired wealth without a real sense of what wea

### Human: What is the main point of his experience with duplicity?
### Assistant: Even though he was not able at first to see through European duplicity because of his inexperience, when he gains the necessary amount of u

### Human: What is the highlight of his trip to Europe?
### Assistant: Having made his fortune, he is traveling in Europe to expose himself to beauty and culture, and hopefully find a wife. ###

### Human: Newman reflects on how good it was for him to see societies based on something other
### Assistant: Near the end of his excursion, Newman reflects how good it was for him to see societies based on something besides wealth. ###

### Human: Newman is a complex character in that he is often open, honest, and kind
### Assistant: Newman is a complex character in that he is often frank, open, and kind-hearted, but can also be self-absorbed and has difficulties underst

### Human: Newman's wealth is a result of his efforts to find a worthy person to
### Assistant: But whereas the conquistadors sought treasure to refill the coffers of a greedy and aging empire, Newman has built up his wealth from nothi
```

Fig. 8. Final dataset for training a Llama-2 model

The datasets used for GPT-2 fine-tuning was also employed for Llama-2, ensuring consistency in evaluation. The data was split into training (80%) and testing (20%) subsets. The block size was set to 128 tokens, allowing the model to maintain context over longer sequences.

To efficiently fine-tune Llama-2, a LoRA (Low-Rank Adaptation) approach was utilized, which introduces task-specific parameters without updating the full model weights [32]. This method was chosen to mitigate the risk of overfitting and reduce the computational burden, while still achieving significant performance gains.

The training configuration included:

- **Epochs**: The LLaMA-2 model was fine-tuned for 5 epochs. Given the model's size and complexity, a lower number of epochs was chosen compared to GPT-2 to prevent overfitting, while still allowing the model to learn the specific task.
- **Batch Size**: Due to the model's larger size, the batch size was set to 2 per device. This adjustment was made to balance memory constraints.
- **Learning Rate and Optimization**: A learning rate of 2e-4 was selected, which was lower than that used for GPT-2. This was necessary to accommodate the higher sensitivity of larger models to learning rate adjustments. The AdamW optimizer was used with weight decay to maintain generalization and prevent overfitting.
- **Gradient Accumlation**: To manage memory usage and ensure stable training, gradient accumulation was employed. This allowed the effective batch size to be increased without exceeding GPU memory limits, facilitating better model convergence.

After training, the model's performance was evaluated using perplexity and ROUGE scores, with the latter providing insight into the model's ability to generate text that is both semantically and syntactically aligned with reference texts. Cosine similarity between sentence embeddings of predicted and reference texts was also computed, highlighting the model's ability to capture the semantic essence of the inputs.

Key Challenges and Decisions Made: During the fine-tuning of both GPT-2 and Llama-2, several key challenges were encountered. For GPT-2, one of the main challenges was handling the model's tendency to overfit, especially given the relatively small size of the fine-tuning dataset. To mitigate this, An early stopping strategy was implemented and opted for a ideal number of training epochs - 10, to balance between underfitting and overfitting. The batch size was set to 4 to accommodate memory constraints while still ensuring sufficient gradient updates for learning.

In the case of LLaMA-2, the challenges were amplified due to the model's larger size, which demanded more complex strategies to optimize memory usage and training stability. To address this, 4-bit quantization was employed via BitsAndBytesConfig, significantly reducing memory footprint without sacrificing model performance. Additionally, Parameter-Efficient Fine-Tuning (PEFT) [33] from huggingface through Low-Rank Adaptation (LoRA) was utilized to focus the fine-tuning on a smaller subset of parameters, enabling efficient learning while keeping computational costs in check. A lower learning rate (2e-4) and a reduced number of epochs-5, were selected to further control overfitting. These decisions were crucial in fine-tuning both models effectively, balancing resource constraints with the need for high task-specific performance.

4 Results and Evaluations

In this section, We present the results and evaluations of our experiments involving different LLMs, namely GPT-2 and Llama-2 and the integration of external comments for character profiling. We provide a thorough description of the results, followed by a comprehensive evaluation of their implications.

4.1 Experimental Results

The experimental results are summarized in the tables below. We evaluated the performance of GPT-2 and Llama-2 models with *dialogues dataset, dialogues spectrum* and *character reviews dataset, character reviews spectrum* using various metrics, including perplexity and ROUGE scores. We have taken the readings of perplexity for three times and then calculated the average, since we are using a randomized 80% and 20% training and test sets. This average gives a clearer picture of the score based on the changes in dataset used for training.

GPT-2 Scores : GPT-2 with *dialogues dataset* has achieved a perplexity score of 26.155 (average of 23.802, 27.272, 27.272). ROUGE scores are as follows: ROUGE-1: 0.225, ROUGE-2: 0.0479, ROUGE-L: 0.204. GPT-2 with *character reviews dataset* had a perplexity of 28.541 (average of 26.639, 29.493, 29.493). The ROUGE scores are as follows: ROUGE-1: 0.224, ROUGE-2: 0.044, ROUGE-L: 0.205. The results for GPT-2 demonstrate that incorporating external comments led to an increase in perplexity from 26.155 to 28.541. The lower perplexity score with the *dialogues dataset* suggests that GPT-2 found it easier to predict the next word in this dataset compared to the *character reviews dataset*. This could imply that the dialogues are more straightforward or consistent in structure and content, allowing the model to make more accurate predictions. In contrast, the character reviews dataset might be more complex or varied, making it more challenging for the model to predict the next word accurately. However, The ROUGE scores between the two datasets are quite similar, with only slight variations. The ROUGE-1 and ROUGE-L scores are nearly identical, while ROUGE-2 is slightly lower for the *character reviews dataset*. This suggests that the model's ability to generate text that aligns with the reference texts is consistent across both datasets. However, the slightly lower ROUGE-2 score for the *character reviews dataset* could indicate that the model struggles a bit more with capturing the relationships between words (bigrams) in this dataset compared to the *dialogues dataset*.

Condition	**Perplexity**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
With Character Reviews Dataset	28.541	0.224	0.0441	0.205
With Dialogues Dataset	26.155	0.225	0.0479	0.204

Table 4. Performance of GPT-2 with and without External Comments

The results indicate that while GPT-2 is generally robust and performs consistently across different datasets, it finds *character reviews dataset* slightly more challenging than the other. This is reflected in the higher perplexity and slightly lower ROUGE-2 scores when dealing with character reviews. However, the overall similarity in ROUGE scores suggests that the model is still able to generate coherent and relevant text in both scenarios.

Llama-2 Scores : Llama-2 has fetched a perplexity score of 16.144(average of 14.59, 14.960, 18.882) with the *dialogues dataset*. The perplexity score with *character reviews dataset* is 13.297 (average of 13.065, 13.543, 13.285). The ROUGE scores for *dialogues dataset* are as follows: ROUGE-1: 0.604, ROUGE-2: 0.599 and ROUGE-L: 0.604. ROUGE scores for *character reviews dataset* are: ROUGE-1: 0.360, ROUGE-2: 0.352 and ROUGE-L: 0.360. For LLaMA-2, the inclusion of external comments improved perplexity from 16.144 to 13.297. However, this came with a decrease in ROUGE scores across the board. The inclusion of external comments resulted in a significant decrease in all ROUGE scores. This suggests that while external comments may introduce additional context or information, they may not always align well with the primary dataset, leading to generated text that is less similar to the reference texts. The drop in ROUGE scores indicates that the model's output became less relevant or coherent when external comments were included, which could be due to the model incorporating information that diverges from the core content of the dialogues or character reviews.

Condition	**Perplexity**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
With Character Reviews Dataset	13.297	0.360	0.352	0.360
With Dialogues Dataset	16.144	0.604	0.599	0.604

Table 5. Performance of LLaMA-2 with and without External Comments

4.2 Similarity Score Distributions

The similarity score distributions and corresponding metrics for the nodes reveal notable differences between the models (GPT-2 and Llama-2) and their performance with and without external comments. We generate predictions for all the labels in test set and calculate the similarity score using cosine similarity. This helps us in knowing how similar our model predictions are compared to the test set. The following table summarizes the number of highly correlated, moderately correlated, and not similar node pairs for each model under different conditions.

Model	**Condition**	**Highly Correlated (Score $\geq$ 0.7)**	**Moderately Correlated (0.4 $\leq$ Score $<$ 0.7)**	**Not Similar (Score $<$ 0.4)**
GPT-2	With character reviews dataset	136,497	63,878	3,641,225
	With dialogues dataset	305	52,399	3,572,512
LLaMA-2	With character reviews dataset	82	1,437	11,250
	With dialogues dataset	81	1,674	8,446

Table 6. Summary of similarity scores for GPT-2 and LLaMA-2 models with different datasets

The analysis indicates that the integration of external reviews can significantly impact perplexity but may not consistently improve ROUGE scores. For GPT-2, the increased perplexity with external comments suggests that while the model might capture more aspects of the character's personality, it also introduces additional complexity that the model struggles to handle effectively. Conversely, Llama-2 shows improved perplexity with external comments but at the cost of reduced ROUGE scores, highlighting a trade-off between model complexity and output quality.

The node correlation analysis demonstrates that the inclusion of external reviews tends to increase the number of highly correlated nodes, suggesting a potential benefit in capturing more detailed or nuanced relationships in the data. However, this is model dependent, as seen in the different results for GPT-2 and Llama-2.

4.3 Predictions

Predictions helps us in closer evaluation of our results. While scores and metrics offer quantitative insights into the model's performance, examining the actual predictions allows us to assess how well the models comprehend and generate text in response to specific queries. This step is vital in understanding the models' abilities to process input data meaningfully and generate outputs that are contextually and semantically accurate. To manually evaluate the predictions, two standard questions were selected:

- **Question 1:** ***What motivates you to travel to Europe?***
  This question is framed in the second person, using "you" to simulate an interaction where the model is expected to respond as if it were the target character from the dataset. The intent here is to evaluate whether the model has correctly internalized the training data to generate responses that align with the personality, motivations, and experiences of the character it is meant to represent.
- **Question 2:** ***Describe Christopher Newman's relationship with Claire de Cintré.***
  The second question is posed in the third person, asking the model to recount a specific relationship from the novel. This question tests the model's knowledge retrieval abilities, particularly its understanding of character dynamics and narrative elements. It also assesses whether the model can generate an accurate and contextually appropriate summary of the relationship between Christopher Newman and Claire de Cintré, as described in the source material.

Before we delve into the model predictions, it is essential to establish the ground truth—the original answers to these questions based on the novel from which the characters and scenarios are drawn. The ground truth serves as a benchmark against which the model's responses can be evaluated.

Ground Truth

- **For Question 1**: In the novel, Christopher Newman is portrayed as an ambitious and wealthy American businessman who travels to Europe primarily for leisure, cultural enrichment, and to find a wife. His motivations include a desire to explore European art and society, experience different cultures, and ultimately seek a companion who aligns with his values and aspirations.
- **For Question 2**: Christopher Newman's relationship with Claire de Cintré is complex and evolves throughout the novel. Initially, it is marked by mutual respect and affection, with Newman deeply admiring Claire's beauty, grace, and aristocratic background. However, their relationship faces numerous challenges due to the rigid social structures and family pressures that Claire faces. Ultimately, despite their deep emotional connection, the relationship is failed by external forces, leading to a tragic conclusion where Claire withdraws from the relationship under family pressure.

Model Predictions

With the ground truth in mind, we now evaluate the predictions made by GPT-2 and Llama-2 for these two questions. The focus will be on assessing how closely the models' responses align with the ground truth, as well as the extent to which they capture the nuances of the characters' motivations and relationships. The max_token length was set to 50 for GPT-2 and 75 for Llama-2.

- **GPT-2 Predictions trained with *dialogues dataset***:
  - **Response to Question 1**: I don't know. But I do know that I am not alone.," said Newman, "I"ll come back"
  - **Response to Question 2**: I don't know," said Newman. "I don't believe that he is a good person."
- **GPT-2 Predictions trained with *character reviews dataset***:
  - **Response to Question 1**: I don't know. I'll tell you,' said Newman. 'I"m tired of being dragged along by the coast"
  - **Response to Question 2**: "He is a very good friend of mine, and I am very glad to have him."

- **Llama-2 Predictions trained with *dialogues dataset***:
  - **Response to Question 1**: "I am not motivated to travel to Europe by any desire to see its cities or to learn its language. I am not a tourist, and I do not wish to be one. I am a writer, and I have a purpose in going to Europe. I wish to see the places where the great writers of the past lived and worked."
  - **Response to Question 2**: "A few days after his arrival in Paris, he had been presented to her, and he had immediately felt the spell of her beauty, which was a sort of delicate and exquisite charm. He had not, however, been in her presence more than a few minutes before he was conscious of a certain reserve, a certain haughtiness, which he could not explain."
- **Llama-2 Predictions trained with *character reviews dataset***:
  - **Response to Question 1**: "I'm a bit of a history buff, so I'm always interested in visiting the places where important events happened. Europe has a lot of that kind of history, and I think it would be really interesting to see the places where some of the most important events in human history took place. For example, I'd love to visit the Colosseum in Rome, or the Tower of London in England"
  - **Response to Question 2**: "Newman's relationship with Claire de Cintré is one of the most significant and complex aspects of the novel. Newman is a wealthy American who has come to Europe to marry Claire, a young Frenchwoman of noble birth. However, Newman soon realizes that he has made a mistake in marrying Claire."

The analysis of the model predictions reveals clear distinctions in how GPT-2 and Llama-2 handle the task of generating responses based on the dialogues and character reviews datasets.

– **GPT-2 Predictions:**

- **Trained with *dialogues dataset*:** The responses generated by GPT-2 when trained on the *dialogues dataset* are out of context and lack meaning. For example, the response to the second question is irrelevant, with no clear connection to the relationship between Christopher Newman and Claire de Cintré. These outputs suggest that the model struggled to generate appropriate responses when relying solely on dialogue data, likely due to the limited context provided by isolated dialogues.

- **Trained with *character reviews dataset*:** When GPT-2 was trained on the *character reviews dataset*, the responses exhibited slight improvements. The first response still lacks context, as it mentions being "tired of being dragged along by the coast," which does not directly address the motivation to travel to Europe. However, the response to the second question, "He is a very good friend of mine, and I am very glad to have him," provides some relevant information, indicating a friendship. Although this answer is still far from fully capturing the complex relationship between Newman and Claire, it does hint at a positive connection, which is more aligned with the narrative than the outputs from the dialogue-trained model.

– **Llama-2 Predictions:**

- **Trained with *dialogues dataset*:** Llama-2, when trained on the *dialogues dataset*, produces responses that are more contextually appropriate compared to GPT-2. The first response effectively captures a writer's motivation to travel to Europe, mentioning a desire to visit places associated with great writers, which, while not entirely accurate to Newman's character, demonstrates a more contextually relevant output. The second response also demonstrates a better understanding of the relationship dynamics, mentioning Newman's initial attraction to Claire and his awareness of her reserve and haughtiness. Although not perfectly aligned with the ground truth, this response shows a deeper comprehension of the characters' interactions.

- **Trained with *character reviews dataset*:** Llama-2's performance further improves when trained on the *character reviews dataset*. The first response now includes a historical perspective, accurately reflecting a character interested in the significance of European landmarks, which is more aligned with Newman's motivations in the novel. The second response is particularly notable, as it correctly identifies the complexity of Newman's relationship with Claire, including the realization of a potential mistake in pursuing the marriage. This output closely shows the emotional and social tensions present in the original novel, showcasing Llama-2's superior ability to capture and convey nuanced relationships.

In summary, the predictions highlight the distinct advantages of training on character reviews versus dialogues. While both models struggled with out-of-context and irrelevant responses when relying solely on dialogues, the inclusion of character reviews significantly enhanced their ability to generate contextually appropriate and accurate responses. Llama-2 consistently outperformed GPT-2, demonstrating a more clear understanding of character relationships and motivations, particularly when fine-tuned with the richer, more descriptive *character reviews dataset*.

4.4 Evaluation

The experimental results for GPT-2 and Llama-2 across various datasets and conditions provide several key insights into the performance and behavior of these models.

GPT-2 and Llama-2 exhibited differing performance characteristics with the two datasets. For GPT-2, the perplexity scores indicate that the model finds the *dialogues dataset* easier to handle compared to the *character reviews dataset*. In contrast, Llama-2 showed a lower perplexity for the *character reviews dataset* compared to the *dialogues dataset*. This difference suggests that each model may have different strengths in handling the textual complexities inherent in these datasets. The lower perplexity score for Llama-2 with the *character reviews dataset* implies that this model may be more consistent at managing the content-specific details of character reviews.

The inclusion of external comments affected ROUGE scores differently for each model. For GPT-2, the ROUGE scores remained relatively stable with only slight decreases. This suggests that GPT-2's text generation remains consistent despite the added complexity of external comments. In contrast, Llama-2 experienced a significant drop in ROUGE scores with external comments. This drop in scores indicates that while Llama-2's perplexity improved, the external comments potentially introduced additional context that did not align well with the reference texts, leading to less relevant or coherent outputs. The drop in ROUGE scores with external comments for Llama-2 highlights a trade-off between enhanced perplexity performance and the quality of generated text.

Overall, the evaluation highlights that both models show strengths and limitations depending on the dataset and the inclusion of external comments. GPT-2 demonstrates robustness across different datasets with stable ROUGE scores, whereas Llama-2 shows improved perplexity but reduced ROUGE scores with external reviews. These findings underscore the importance of considering both predictive accuracy and text coherence when evaluating model performance. Additionally, the differences between GPT-2 and Llama-2 in handling external comments suggest that model selection should align with specific requirements for handling contextual information and maintaining output relevance.

Also, Manual evaluation of predictions showed models trained with external comments had a slight edge in understanding and giving contextually relevant replies compared to their counterparts.

5 Conclusion and Future Work

5.1 Conclusion

The main aim of this research was fine-tuning of two LLMs - GPT-2 and Llama-2 on two different datasets - dialogues only dataset and dialogues dataset with extra character reviews. This is done to evaluate if adding external character reviews showed any improvement in the results. The project achieves this objective by curating two different datasets and evaluating the fine-tune models on scores like ROUGE and perplexity.

Going into the detailed aims and objectives, Firstly, We have seen the history of chatbots, Large Language Models and the recent developments taking place in this field with the introduction of GPT-4 [28] and Llama-3.1 [27]. This helped us in laying a foundation on the advancements taking place in this field.

Two datasets were prepared for this project - One with only dialogues from the novel and the second dataset which contains character reviews along with the dialogues. Character traits were also introduced in these datasets. Rusnachenko et. al. [35] work greatly helped us in extracting the dialogues and building the character spectrums. Context is added to the dialogues and formatting is taken care of based on the input requirements for respective LLM models. Preparing datasets was crucial step as these lay the strong base for fine-tuning the models.

Question Generator model was used to automate the process of generating questions to large amounts of text. Though the model produced relevant questions in most of the cases, Some gaps are present with questions semantically out of order and sometimes blank lines as well. Improvement in Text-to-Text generators in future can handle this problem more efficiently.

Next objective included fine-tuning the models and evaluating their performance. The research provided us with interesting results showcasing mixed performance by the introduction of external information. GPT-2 exhibited consistency across different datasets but it struggled in handling the varied content of character reviews. This is clearly visible in stable ROUGE scores but increase in perplexity. Llama-2, on the other hand excelled in capturing the details of character reviews, reflected in its lower perplexity scores. However, the trade-offs are significant decrease in ROUGE scores meaning, though Llama-2 managed to integrate external data well, it struggled to maintain relevance when dealing with added contextual information.

Manual evaluation of predictions showed us adding context and external information improves the responses. But again, GPT-2 predictions were less close to the ground truth, but Llama-2 performed well in this segment. This can be taken forward in the future by conducting a more deeper evaluation of predictions with different human assessments and testing the models with more diverse prompts.

Overall, this research underscores that dataset selection and the quality of the dataset plays an important role in fine-tuning these models. It also highlights the complex relationship between perplexity and ROUGE scores, suggesting improvements in one may not always be an advantage to the another. The insights gained from this study can be helpful for further development for tasks involving character profiling and text generation in narrative contexts.

5.2 Future Work

There are several future works that can be proceeded and taken forward based on this research. These could significantly enhance the capabilities of LLMs in text generation. Some of these are:

- Enhanced Contextual Integration: The integration of external comments gives mixed results in this study. Further work can be done to integrate a better, more detailed and clear dataset to fine-tune the models. Also, research could be done to find out what kind of data is more suitable to train the models.

- Prompt Engineering: The whole project is solely based on fine-tuning the LLMs. But with the recent LLMs having being trained on vast amount of parameters, We could further analyze if detailed prompt engineering could deliver the similar results. This can help reduce the resource and time consuming process of fine-tuning these models.

- Fine-Tuning Advanced Models: We have chosen GPT-2 and Llama-2 for this study due to their light weight architecture and resource availability. By the time of this study, Llama-3.1 has been launched [27] trained on 405 billion parameters. OpenAI's most advanced GPT-4 [28] has been launched. These advanced models could be used for fine-tuning, which gives more improved results and perform better.

- Cross Model Fine-Tuning: As we have observed in this study, GPT-2 is a consistent in handling complex data consistently while Llama-2 excelled in capturing the details. Further research could involve development of hybrid models that balances consistency in addition of new data with text coherence more accurately.

- Reinforcement Learning Models: The models trained in this study used the dataset integrated with character reviews from external websites. Though we have the results, no feedback mechanism has been used. We could add a feedback mechanism, so model can be rewarded for making right decisions as it progresses on further training.

– Multilingual Capabilities: This research is purely based on english language. Chatbots can be trained on different languages making them more versatile [29].

Acknowledgement: I am deeply grateful to my mentor Huizhi Liang for helping me in each and every step of this project. She guided me thoroughly with all the information required to carry out this work. I am also thankful to Nicolay Rusnachenko for helping me set up the foundations on which this project further moves forward. I would also like to appreciate the Computing Department of Newcastle University providing me with resources and support all along to finish this work.

References

[1] Oracle. *What is a Chatbot?* [Online]. URL: https://www.oracle.com/uk/chatbots/what-is-a-chatbot/.

[2] M. E. H. Y. E.-S. H. F. Tarek Ait Baha. "The Power of Personalization: A Systematic Review of Personality-Adaptive Chatbots". In: *SN Computer Science* (2023).

[3] P. B. B. Asbjørn Følstad. "Users' experiences with chatbots: findings from a questionnaire study". In: *Springer* (2020).

[4] L. Goldberg. "The development of markers for the Big-Five factor structure". In: *Psychological Assessment* 4.1 (1992).

[5] M. G. C. B. G. P. S. J. G. B. D. Jiwei Li. "A Persona-Based Neural Conversation Model". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Vol. 1. 2016, pp. 994–1003.

[6] N. S. J. W. T. B. B. A. R. D. A. P. C. G. I. Daniel M. Ziegler. "Fine-Tuning Language Models from Human Preferences". In: *Computation and Language (cs.CL)* (2020).

[7] Project Gutenberg. *Project Gutenberg*. 2024. URL: https://www.gutenberg.org/.

[8] Fatih Karahbiber. *Cosine Similarity*. [Online]. URL: https://www.learndatasci.com/glossary/cosine-similarity/.

[9] Yazhou Zhang et al. "DialogueLLM: Context and Emotion Knowledge-Tuned Large Language Models for Emotion Recognition in Conversations". In: *arXiv preprint arXiv:2310.11374* (2024). arXiv: 2310.11374 [cs.CL].

[10] CliffsNotes. *Christopher Newman Character Analysis*. 2024. URL: https://www.cliffsnotes.com/literature/a/the-american/character-analysis/christopher-newman#:~:text=In%20general%2C%20Newman%2C%20though%20inexperienced,man%20both%20physically%20and%20morally..

[11] *Retail Spend over Chatbots to Reach $12bn Globally in 2023*. [Online]. Juniper Research. June 2023. URL: https://www.juniperresearch.com/press/retail-spend-over-chatbots-to-reach-12bn-globally/.

[12] SparkNotes. *Christopher Newman*. 2024. URL: https://www.sparknotes.com/lit/american/character/christopher-newman/.

[13] BakerP2004. *Character Analysis of Christopher Newman*. 2014. URL: https://bakerp2004.wordpress.com/2014/10/06/character-analysis-of-christopher-newman-from-the-american-by-henry-james/.

[14] Shmoop. *Christopher Newman*. 2024. URL: https://www.shmoop.com/study-guides/the-american/christopher-newman.html.

[15] CliffsNotes. *Summary and Analysis of Chapter I*. 2024. URL: https://www.cliffsnotes.com/literature/a/the-american/summary-and-analysis/chapter-i.

[16] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[17] Wikipedia contributors. *Python (programming language)*. Accessed: 2024-08-08. 2024. URL: https://en.wikipedia.org/wiki/Python_(programming_language).

[18] Explosion AI. *SpaCy: Industrial-Strength Natural Language Processing*. Accessed: 2024-08-08. 2024. URL: https://spacy.io/.

[19] Fabian Pedregosa et al. *Scikit-learn: Machine Learning in Python*. Accessed: 2024-08-08. 2011. URL: https://scikit-learn.org/stable/.

[20] R. W. Picard and Timothy W. Bickmore. "Establishing and Maintaining Long-Term Human-Computer Relationships". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 12.2 (2005).

[21] Google Research. *Google Colaboratory: Free Jupyter Notebook Environment*. Accessed: 2024-08-08. URL: https://colab.research.google.com/.

[22] Suraj Patil. *T5-base for question generation with highlighted answers*. https://huggingface.co/valhalla/t5-base-qg-hl. Accessed: August 10, 2024. 2020.

[23] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL]. URL: https://arxiv.org/abs/2307.09288.

[24] Hugging Face. *Hugging Face: Home*. https://huggingface.co/. Accessed: 2024-08-12. 2024.

[25] OpenAI Community. *GPT-2 Model*. https://huggingface.co/openai-community/gpt2. Accessed: 2024-08-12. 2024.

[26] Nous Research. *Llama 2 7B Chat Model*. https://huggingface.co/NousResearch/Llama-2-7b-chat-hf. Accessed: 2024-08-12. 2024.

[27] Meta AI. *Llama 3.1 Model Cards and Prompt Formats*. Accessed: 2024-08-12. 2024. URL: https://llama.meta.com/docs/model-cards-and-prompt-formats/llama3_1.

[28] OpenAI. *GPT-4 Research*. Accessed: 2024-08-12. 2024. URL: https://openai.com/index/gpt-4-research/.

[29] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.

[30] Tom Cookson. *Fictional Character Personalities Data*. https://github.com/tacookson/data/tree/master/fictional-character-personalities. Accessed: 2024-08-13. n.d.

[31] Cloudflare. *What is a large language model (LLM)?* [Online]. Accessed: July 2024. URL: https://www.cloudflare.com/learning/ai/what-is-large-language-model/.

[32] Cloudflare. *What is LoRA (Low-Rank Adaptation)?* Accessed: 2024-08-14. 2024. URL: https://www.cloudflare.com/learning/ai/what-is-lora/.

[33] Hugging Face. *Parameter-Efficient Fine-Tuning (PEFT)*. Accessed: 2024-08-14. 2024. URL: https://huggingface.co/blog/peft.

[34] Wikipedia contributors. *TensorFlow*. Accessed: 2024-08-14. 2024. URL: `https://en.wikipedia.org/wiki/TensorFlow`.

[35] H. L. Nicolay Rusnachenko. "Personality Profiling for Literary Character Dialogue Agents with Human Level Attributes". In: *10th Annual Conference on Machine Learning, Optimization and Data Science (LOD)*. 2024.

[36] J. Weizenbaum. "ELIZA: A Computer Program For the Study of Natural Language Communication Between Man and Machine". In: *Communications of the ACM* (1966).

[37] Alec Radford and J. W. R. C. D. L. D. A. I. S. *Language Models are Unsupervised Multitask Learners*. 2018.

[38] D. Cheung. *Meta Llama 2 vs. OpenAI GPT-4*. [Online]. Accessed: 26 July 2024. Dec. 5, 2023. URL: `https://medium.com/@meetdianacheung/meta-llama-2-vs-openai-gpt-4-785589efe15e`.

[39] V. Arnold. *Llama 2 vs GPT-4: Which language model is more efficient?* [Online]. Accessed: 26 July 2024. Jan. 12, 2024. URL: `https://neuroflash.com/blog/llama-2-vs-gpt-4-which-language-model-is-more-efficient/`.