University *of Ljubljana*
Faculty *of Computer and Information Science*

Conversations with Characters in Stories for Literacy — Quick, Customized Persona Bots from novels

Leon Todorov, Jan Rojc, Andraž Zrimšek

Abstract

This paper explores the design and potential of PersonaBots, digital personifications of characters, as a novel way to engage with literature. We address the challenges of creating PersonaBots from user-suggested novels due to the limited token count of character descriptions. Our approach includes the use of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and In Context Learning (ICL). We also delve into the educational potential of PersonaBots, as pedagogical agents have shown promising results in enhancing children's learning. Furthermore, we explore the use of sentence comparison techniques and situation models to create a bot that responds in a way that is consistent with its backstory. The process of corpus analysis, where we extract books from The Project Gutenberg repository for our study, is also discussed. This comprehensive exploration of the theoretical systems informing PersonaBot design, the evaluation of pedagogical agents, and the existing services available for PersonaBot creation contributes to the ongoing discourse on the use of digital technologies in literature and education.

Keywords
PersonaBots, In Context Learning, Large Language Models

Advisors: Slavko Žitnik

Introduction

Digital technologies have transformed our interaction with literary characters through PersonaBots. These bots, mimicking characters from novels, offer a unique reading experience. However, creating PersonaBots from user-suggested novels is challenging due to limited token count. This paper explores PersonaBot design using Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and In Context Learning (ICL).

We also investigate the educational potential of PersonaBots as pedagogical agents, which have shown to enhance children's learning. This is crucial considering the current worldwide literacy crisis. Existing services like Khanmigo offer limited PersonaBot conversations. We propose quick, customized PersonaBots based on teacher-suggested novels, considering current retrieval and indexing techniques.

Additionally, we explore sentence comparison techniques and situation models for bot backstory consistency and discuss corpus analysis using The Project Gutenberg repository. This paper contributes to the discourse on digital technologies in literature and education.

Related work

Existing services

There currently exist many customizable personaBot services, such as character.ai, chatfai.com, dreamtavern.ai, moemate.io and many others. Most just use a basic character description ("How would your character describe themselves?"), but some implementations are extended by adding exemplary character greetings/dialogue or by directly adding additional backstories (sometimes even from external sources). While these platforms offer customization, the character descriptions usually have a very limited token count, which might be problematic when trying to create a custom personaBot from a character in a user-suggested novel.

Pedagogical agents evaluation

Pedagogical agents show promising results for enhancing children's learning ([1][2]). They can improve both question-asking skills and vocabulary learning depending on the design of the agent and the students' characteristics, hence pedagogical agents in the form of text conversation with characters from novels might motivate its users to read more and improve learning.

Theoretical systems informing PersonaBot design
When looking at conversational AI systems, there are multiple approaches of that have been studied. The model proposed in [3] takes a more rule-based approach, but is mostly made to retain a human's attention rather than mimic a personality. A more useful approach is considered in [4], where a character is represented by an embedding, which should be close to the embeddings of words that are commonly around the character. With the rise of large language models (LLMs), [5] which perform great with conversational tasks, we are given many pre-trained models that can be considered. They perform great on general topics but require additional information when adapting to a specific domain. For this, we require the use of Retrieval-Augmented Generation (RAG). They allow us to find information relevant to the task from outside sources. This information can then be added to the language model's context using In Context Learning. LLMs, in combination with ICL, represent a possible approach to mimicking a persona without any parameter updates. For ICL an open source toolkit OpenICL [6] has been released, which offers state-of-the-art retrieval and inference methods that can use to adapt ICL to a specific problem.

Sentence comparison When looking at comparing sentences, a good approach is embedding the sentence and using the spatial relation of sentence embeddings to find the closest ones. If only working with English sentences, [7] offers a sentence embedding model based on a pretrained BERT network. If we want to enable conversations with our model in multiple languages, using a language-agnostic sentence embedding model like LASER, presented in [8] offers a good solution.

Situation models
Situation models are mental representations built by readers to understand the characters, events, and overall setting described in the text. Zwaan et al. (1998) [9] investigated which aspects of a situation model are actively monitored during reading. Their findings suggest that readers primarily focus on dimensions like time, causality, goals, and the protagonist. By incorporating these aspects of situation models into personaBots, we can create a bot that responds in a way that is consistent with its backstory.

Corpus analysis
In this project, we will be extracting books from The Project Gutenberg repository [10] to serve as the text source for our analysis. The corpus consists of over 70,000 free eBooks. It focuses on older works whose copyright has expired in the United States, making them part of the public domain.

1. Proposed solution

The solution we propose is using ICL to help a pre-trained LLM produce answers that would relate to the character. The data we would use to give it context could be using question-answer pairs that will be automatically generated from the desired novel. A possible approach is proposed in [11]. An efficient method for finding relevant questions would be to use quick vector comparison methods to compare the question with questions in our dataset and use the question-answer pairs as examples to add to our model's context.

2. Methods

At the core of our conversational agent lies a pre-trained LLaMA2 chat model. We provide it instructions through a carefully crafted system prompt, which essentially tells the LLaMA2 model how to behave and what information to consider when crafting responses.

2.1 RAG

To improve the performance of our conversational agent, we employ two types of RAG. To better capture the style of speaking of the character, we search the book for relevant lines spoken by the character. To better capture the context for the answer, we also search the entire book segments for the most relevant parts.

Sentence extraction

When starting our model, we save the entire book into a class and extract lines spoken for each character. When the model receives a question, it embeds it using the *multi-qa-mpnet-base-cos-v1* sentence transformer, which was trained for semantic search, especially for question-answer sentence pairs. It also returns normalized embeddings with length 1, which allows for easier comparison. With this, we aim to find the lines from our character that could be a response to the question. We compare the embedding to embeddings of all character lines and return the ones that are most similar according to cosine similarity.

Context extraction

When importing the book, we also split the entire text into segments of length 500. When we receive a question, we also embed it using the *all-mpnet-base-v2* sentence transformer and compare it to all the extracted segments. We add the two most similar to our prompt for context.

ICL

The end product is a prompt which includes a system prompt telling our agent how to act, as well as top n character lines and extracted context from the book.

2.2 Evaluation

Contextual awareness

We can assess the conversational agent's ability to understand context by measuring its performance on questions that require using the surrounding information. Here, we propose several methods for obtaining context-dependent question-answer pairs:

- Utilize a pre-trained model like mojians/E2E-QA-Mining to automatically extract question-answer pairs directly from the book.
- Prompt LLM to generate follow-up questions based on a provided passage.

- Carefully select context-dependent questions from educational websites or other curated online sources.
- Manually craft context-dependent questions that specifically target the character's knowledge, experiences, or the overall story.

Now the context evaluation question and answer embeddings are extracted using the *all-mpnet-base-v2* sentence transformer. Then we prompt the conversational agent with the question and compare the output embedding with the known answer's embedding.

Character personality

We can evaluate how well the agent captures the character's personality by examining its responses to general conversational questions that don't necessarily have a specific answer in the book. These questions can be obtained by:

- We can prompt LLM to generate open-ended questions.
- Manually crafting open-ended questions.

Now the personality evaluation question embeddings are extracted using the *all-mpnet-base-v2* sentence transformer. For each extracted question the most relevant character line is extracted by comparing the question's embedding with the embeddings of all extracted character lines and selecting the one with the highest cosine similarity. Then we prompt the conversational agent with the question and compare the output embedding with the embedding of the most relevant character line.

References

[1] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. Pedagogical agents for fostering question-asking skills in children. *CoRR*, abs/2004.03472, 2020.

[2] Thijs M. J. Nielen, Glenn G. Smith, Maria T. Sikkema de Jong, Jack Drobisz, Bill van Horne, and Adriana G. Bus. Digital guidance for susceptible readers: Effects on fifth graders' reading motivation and incidental vocabulary learning. *Journal of Educational Computing Research*, 56(1):48–73, 2018.

[3] Ioannis Papaioannou et al. *Designing coherent and engaging open-domain conversational AI systems*. PhD thesis, Heriot-Watt University, 2022.

[4] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[5] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.

[6] Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*, 2023.

[7] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Holger Schwenk and Matthijs Douze. Learning joint multilingual sentence representations with neural machine translation, 2017.

[9] Rolf A Zwaan, Gabriel A Radvansky, Amy E Hilliard, and Jacqueline M Curiel. Constructing multidimensional situation models during reading. *Scientific studies of reading*, 2(3):199–220, 1998.

[10] Michael Hart. Project gutenberg, 1971.

[11] Jian Mo. E2e-qa-mining: Training transformers (t5) for end to end question answer pair mining, 2023. GitHub repository.