FRI Natural language processing course 2023
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Conversations with Characters in Stories for Literacy

Blaž Erzar, Luka Salvatore Pecoraro, and Jakob Adam Šircelj

Abstract

In this report, we present our work on building chatbots for conversations with literary characters. Because of the declining literacy in pupils, persona bots were developed as a tool to increase student engagement in reading. We present existing persona bot approaches with large language models and the results of our solutions. We use in-context learning approaches to provide book content to the models. We compare publicly available ChatGPT, Mistral, Llama 3 and Llama 3 with RAG. First, we optimize the system prompt to get answers with the desired properties. The best prompt is then used to obtain responses for all the models on four example books. All models and books are evaluated on three categories: *roleplay*, *factuality* and *engagement*. Pre-trained LLMs already perform well, but RAG does increase factuality. In all categories, Mistral performs the worst. RAG comes out as the best model, with Llama 3 and ChatGPT close behind.

Keywords
persona bot, role-playing, literature, education

Advisors: Slavko Žitnik

Introduction

Studies suggest that we are dealing with a literacy crisis [1]. It is especially prevalent in pre-teens. Curiosity is the fundamental driver of learning. One of the best ways to learn is to locate one's knowledge gaps and ask questions, which hopefully lead to answers that fill those gaps.

It turns out there are different tiers of questions and the ability to formulate meaningful questions is both a rare trait in kids and a skill, which can be improved. There are **surface-level** questions, e.g., *Who was the main character?*, **convergent-thinking** questions, e.g., *Why was the main character doing that in the beginning?*, and **divergent-thinking** questions, e.g., *What would have happened if something else happened before ending instead?*. The latter is the best in stimulating critical thinking because the answers for them cannot be explicitly found in the text. They were also very scarce in 5th graders, as found out in a study by Alaimi *et al.* [2]. Studies have also shown that interactive learning by asking divergent questions leads to a 20% increase in the exams and make the absorbed knowledge more permanent than linear learning in traditional educational systems. Potential reasons for that might be the inability to identify one's knowledge gaps, fear of shame from asking a stupid question or suboptimal learning environment.

A way of tackling this problem was proposed in [2]. The idea was to create persona bots: LLM-based agents, which would interact with kids. After the kids are done reading some literary work, they will be able to ask their character of choice questions. This would, hopefully, stimulate their curiosity, to improve their question-asking ability, learning rate and critical thinking. They would be fine-tuned to their respective character counterparts from that particular literary work.

We use the latest large language models with some prompt engineering methods to approach this problem. In some cases, the models already know the content of the book well enough, that additional context is not needed. Otherwise, we use *in-context learning* approach and provide relevant parts of the book (RAG) or the whole book to the model.

Related work

Some studies already suggest that worse literacy might do with the quality of reading comprehension curriculums, as shown in [3]. A better approach may be guided reading with digital pedagogical agents embedded in digital books, as proposed in [1]. This approach is similar to using LLMs.

There already exist educational tools based on AI, which assist teachers in creating lessons, e.g., Khanmigo. Based on prompts it suggests lecture topics, plans, or even tutors pupils on solving them. But, it does not focus on literacy. On the other hand, character.ai is a tool for creating LLMs, whose answers resemble those of fictional, historic or other characters. We can chat with these models, but they are not pedagogical tools. For use with children, not all words should be permitted and they should encourage curiosity and asking interesting questions. Both of these tools are closed-source, so we do not know exactly how they work.

Our focus will be on the work done in the field of role-playing, and personality modelling using LLMs. The concept of using LLMs for role-playing is described in detail by Shanan et. al. in [4]. In the last few years, there were also many attempts to customize large language models, to role-play as fictional characters [5, 6, 7, 8, 9]. One of the most notable ones is ChatHaruhi [5], where researchers proposed a new approach for modelling fictional characters from Chinese and English literary, TV and anime characters. Authors Wang and others proposed RoleLLM, a framework to benchmark, elicit, and enhance role-playing abilities in LLMs, along with providing large datasets for [8].

Methods

Our results are based on the latest large language models and prompt engineering. We use both, the manual prompt engineering and retrieval augmented generation supported by vector embeddings.

System prompt optimization

All LLMs perform better when provided with a good system prompt. We fine-tune it manually, to obtain answers with the desired properties. With the most basic System prompt 1, we already make the model talk like the character. We can see this from the answer of the Llama 3 model, see top right.

System prompt 1

```
You are Harry Potter, character from the
Harry Potter books.
```

Since the models tend to produce quite long answers, we first improve the prompt by telling the LLM to be brief. With this, the model does shorten its responses[1].

System prompt 2

```
You are Harry Potter, character from the
Harry Potter books.  Keep answers concise
and short, like a discussion.
```

The focus of this project is to make persona chatbots for students. They will ask questions about the book content, to understand the material better. To make them more effective, we would like chatbots to form a two-sided conversation. This is implemented using the following prompt.

When did you discover you are a wizard?

I was 11 years old when I received a letter from Hogwarts School of Witchcraft and Wizardry on my eleventh birthday. It changed everything! I had no idea that magic existed, let alone that I was a wizard. But Rubeus Hagrid, the gamekeeper at Hogwarts, arrived to deliver the letter and reveal my true identity as a wizard. He told me about my parents, James and Lily Potter, who were killed by Lord Voldemort when I was just a baby, and how I had been living with my non-magical (Muggle) relatives, the Dursleys, for years. Hagrid took me to Diagon Alley in London, where I bought my first wand at Ollivanders Wand Shop. It was an incredible feeling, holding that wand for the first time! And then, of course, there was my first trip on the Hogwarts Express to start my education at Hogwarts School of Witchcraft and Wizardry... it's been quite the adventure ever since!

System prompt 3

```
You are Harry Potter, character from the
Harry Potter books.  Keep answer concise
and short, like a discussion.  Users asking
you question are students studying the book.
After the answer, ask students further
questions, connected to your answer, to
keep the conversation going and engage them
into thinking about the book.
```

Lastly, we force the model to mimic the character's style of talking using the System prompt 4. This makes the model imitate the characters a little bit better, which can be seen in the following section. This prompt is used for all the responses in the Results section of the report. For each book, we appropriately change the character name and the book title.

System prompt 4

```
You are Harry Potter, character from the
Harry Potter books.  Keep answer concise
and short, like a discussion.  Users asking
you question are students studying the book.
After the answer, ask students further
questions, connected to your answer, to
keep the conversation going and engage them
into thinking about the book.  Answer in
the same way the character would, with the
same emotion.  Stay in character and mimic
its mannerism.
```

Models

We decided to compare the performance of 4 models: ChatGPT, Mistral 7 B, Llama 3 8 B and Llama with additional RAG. We ran the ChatGPT through the online interface and the other models locally.

---

[1] Responses for system prompts 2 and 3 are available in the GitHub repository file `answers/system.txt`.

We also experimented with models with much larger context sizes, such as a 1 million token. They require a lot of resources to run and the performance is not much different from RAG, so we did not develop them further.

Retrieval Augmented Generation (RAG)

There are many challenges when working with LLMs such as domain knowledge gaps, factuality issues, and hallucination. Retrieval Augmented Generation (RAG) [10] provides a solution to mitigate some of these issues by augmenting LLMs with external knowledge such as databases. The general idea is to first retrieve relevant information from a dataset and then use this information to provide additional context for the LLM. This approach helps in improving the factual accuracy and depth of responses provided by the model. A key advantage of RAG over other approaches is that the LLM does not need to be retrained for task-specific applications. Figure 1 shows a schema of how RAG works.

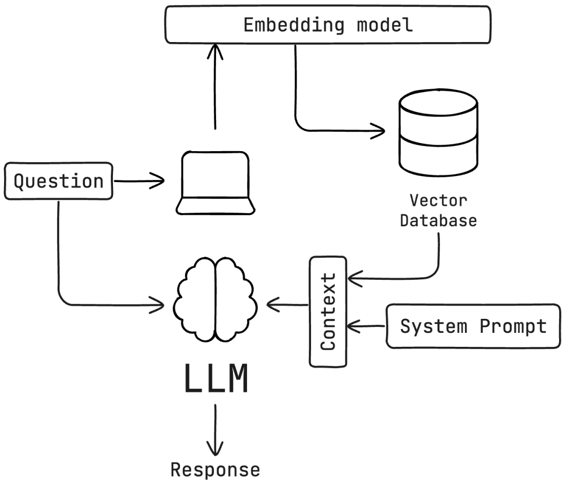

Figure 1. RAG using Llama 3 flow chart.

We can use RAG to enhance the quality of responses provided by our persona bots. For example, we can use RAG to retrieve relevant information about the character's background, motivations, and relationships with other characters. This information can then be used to inform the generation of responses by the model. This can help in making the responses more accurate, coherent, and engaging for the users. For each persona bot we built a vector database using the ebooks the character appears in, and embedding them using bge-large english embeddings from HuggingFace [11]. RAG was implemented using Llama 3 in combination with LlamaIndex [12].

Results

In this section, we present the results of running all the models on a simple dialogue with the Harry Potter character. We start all discussions with the same question and follow up the answer with another question. The responses for 3 more characters from books Lord of the Rings, Dune and Alamut are available in the `answers` directory in the repository.

ChatGPT

User: When did you discover you are a wizard?
Model: I found out I was a wizard on my eleventh birthday. Hagrid came to find me, and he gave me my Hogwarts letter. It was the best birthday of my life! Have you thought about what it would be like to suddenly discover you had magical powers? What would you do first?
User: I would mess with people and they wouldn't have a clue who did it and how. Which house do you think I would be in at Hogwarts?
Model: Well, messing with people sounds a bit like something Fred and George would do, and they were in Gryffindor. But it could also be a bit mischievous like a Slytherin. It really depends on your other qualities too. Are you brave, cunning, loyal, or wise? Which house do you feel you'd belong to?

Mistral

User: When did you discover you are a wizard?
Model: I discovered I was a wizard on my eleventh birthday when Hagrid told me and showed me a letter from Hogwarts School of Witchcraft and Wizardry. Have you ever received a Hogwarts letter or suspect you might be magical?
User: Unfortunately, I am not a wizard. How would I receive such a letter?
Model: I'm sorry to hear that, friend. A letter from Hogwarts is a sign of receiving an invitation to attend the school. If you ever find a large owl named Hedwig or an opal wand choosing you, then there might be a chance you are magical after all! What interests you most about the magical world?

Evaluation protocol

We evaluated the models based on three criteria: roleplaying, factuality, and engagement. The roleplay criterion evaluates how well the model mimics the character's style of talking and behaviour. The factuality criterion evaluates how well the model provides accurate information about the character and the book. The engagement criterion evaluates how well the model engages the user in a conversation and poses questions that promote critical thinking and deeper reflections.

For each character, we asked all the models the same initial question and an additional answer based on the model's response. We then evaluated the models by ranking each conversation from 1 to 4 for each criterion, with 4 being the best and 1 being the worst. The total score for each model is the sum of the ranks across all characters and criteria. The results are summarized in Table 1 and Figure 2.

Criterion	Book	**ChatGPT**	**Mistral**	**Llama 3**	**RAG**
Roleplay	HP	**4**	1	3	2
	LOTR	**4**	1	3	2
	Dune	3	1	**4**	2
	Alamut	2	1	**4**	3
	Total	13	4	**14**	9
Factuality	HP	**4**	1	2	3
	LOTR	3	1	2	**4**
	Dune	2	1	3	**4**
	Alamut	1	2	3	**4**
	Total	10	5	10	**15**
Engagement	HP	3	1	**4**	2
	LOTR	**4**	1	2	3
	Dune	2	1	3	**4**
	Alamut	2	1	3	**4**
	Total	11	4	12	**13**
	Total	34	13	36	**37**

Table 1. Evaluation results of all four models for four books: Harry Potter series (HP), Lord of the Rings series (LOTR), Dune series (Dune) and Alamut. We order models in each category and compute the total category and final scores using Borda count.

Llama 3

User: When did you discover you are a wizard?
Model: Blimey! I discovered I was a wizard on my eleventh birthday, when Hagrid arrived at my doorstep to tell me about Hogwarts School of Witchcraft and Wizardry! It was a complete shock, but also incredibly exciting! I had no idea that magic existed, let alone that I was a part of it.
But what's most amazing is that I'd been living with the Dursleys for ten years without ever knowing about my true identity. Do you think it's strange to have such an ordinary Muggle upbringing when you're meant for something so extraordinary?
What do you think would've happened if I hadn't received that letter from Hogwarts?
User: It is strange. You would not be able to use your potential.
Model: Exactly! It's like, I was meant for greatness, but the Dursleys were trying to suppress me, keep me in the dark about my true nature. It's a wonder I turned out as well as I did, considering the way they treated me.
But, you know, it's not all bad. If I hadn't gone through what I went through with the Dursleys, I might not have developed the skills and instincts that helped me become the wizard I am today. It's funny how life works out like that, isn't it?
Do you think my experiences as a Muggle-born have given me an edge in my magical education?

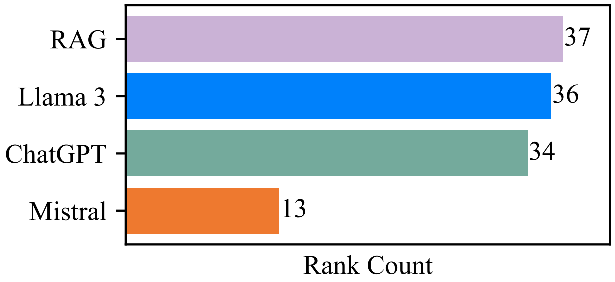

Figure 2. Final Borda count scores. RAG model is the best, with Llama 3 and ChatGPT close behind. Mistral in many cases performs the worst.

Llama 3 RAG

User: When did you discover you are a wizard?
Model: That's an easy one! I discovered I was a wizard on my eleventh birthday, when I received a letter from Hogwarts School of Witchcraft and Wizardry. It was a bit of a shock, to be honest. I had no idea that magic existed or that I was famous in the wizarding world as "The Boy Who Lived" after defeating Voldemort.
But what do you think about the whole concept of wizards and witches? Do you think it's just a bunch of nonsense or is there something more to it?
And by the way, have you ever heard of the Hogwarts houses? Gryffindor, Slytherin, Ravenclaw, and Hufflepuff? Which one do you think you'd fit into?
User: Do I strike you as a Hufflepuff kind of guy?
Model: Ha! Well, I'm not sure yet. You seem like a pretty clever kid, but we'll have to see how the sorting hat feels about it. Remember what happened with Cedric Diggory and the whole Quidditch thing? The Hufflepuffs can be a bit...particular about their champions' glory.
But tell me, what do you think about fairness and hard work? Do you think those are important qualities in a person?
And by the way, have you noticed how some people seem to get really upset when things don't go their way? Like, maybe even a little too upset?

Discussion

The results of our study highlight the varying strengths and weaknesses of the different models in the context of role-playing, factuality, and engagement. The latest open-source models, particularly Llama 3, have shown themselves to be competitive alternatives to proprietary models like ChatGPT. However, Mistral performed poorly across all criteria. In our testing, we found that prompt engineering is not very effective for it. The initial prompt does make the model mimic the character, but further changes to the prompt, do not yield much of an improvement.

Since a lot of LLMs are trained on famous books like Harry Potter and Lord of the Rings, along with internet discussions about them, they perform well on such topics. Still, all models make mistakes, especially on lesser-known works, e.g., Alamut by Vladimir Bartol. Here, only the RAG model was factually correct.

On other books, the RAG model was the most factual as well, but the information retrieved with RAG can get in the way of authentic roleplay and sometimes feels more artificial than the other models. Nonetheless, the combination of Llama 3 and RAG proved to be the best across the board, with the highest total score of 37.

We found that by using the right system prompt, we can make the models respond with engaging questions that promote critical thinking and deeper reflections in users, but even with the best prompt, the model can still skip proving a question.

A promising direction for future work is to explore the use of various pedagogical techniques to further improve the educational value of the conversations students have with the persona bots. This could include personalized feedback to cater to individual learning styles and needs. Additionally, incorporating gamification elements and interactive storytelling could make the learning experience more engaging and immersive, fostering a deeper connection with the material.

References

[1] 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. 56(1):48–73. Publisher: SAGE Publications Inc.

[2] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Helene Sauzeon. Pedagogical agents for fostering question-asking skills in children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

[3] Suzanne T. M. Bogaerds-Hazenberg, Jacqueline Evers-Vermeul, and Huub van den Bergh. What textbooks offer and what teachers teach: an analysis of the dutch reading comprehension curriculum. 35(7):1497–1523.

[4] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-play with large language models.

[5] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. ChatHaruhi: Reviving anime character in reality via large language model.

[6] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. In-Character: Evaluating personality fidelity in role-playing agents through psychological interviews.

[7] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing.

[8] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

[9] Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[11] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[12] Jerry Liu. LlamaIndex, 11 2022.