Exploring Techniques to Generate Dialogues in Hogwarts Legacy Game Using Large Language Models

Kossar Pourahmadi
kmeibodi@ucdavis.edu
University of California, Davis

Atena Saghi
zsaghi@ucdavis.ed
University of California, Davis

Abstract

Computers have become more prevalent in our everyday lives and soon it will not be accepted if they cannot relate to humans by understanding their needs. Storytelling is part of the human experience; we communicate, have fun, and learn through storytelling. As a result, it is important to care about Computational Narrative Intelligence, i.e., the ability of computers to understand and respond effectively to stories. Although there has been great progress in automatic story generation, current narrative systems still cannot be creative enough and generalize to a topic that was not provided as rules of the world initially. We need to build a story generation system that could generate a novel story about any given topic. Since recent research in natural language processing has achieved ground-breaking success by proposing large language models that have a strong capability of both language understanding and generation, in this project we want to take advantage of language models to generate stories. Specifically, we narrow down the problem of story generation to dialogue generation for *Hogwarts Legacy* video game. We fine-tune GPT-2 on *Harry Potter Novels* and Harry Potter Dialogue dataset by Chen et al to generate dialogues given the characters' attributes, characters' relations, context of the dialogue, and dialogue history. Then we compare the results with other language model baselines in terms of response quality, persona consistency, and creativity. Our code is at: https://github.com/atenasadat/ECS289—StoryTelling

Keywords:
Computational Storytelling, Large Language Models

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00

ACM Reference Format:
Kossar Pourahmadi and Atena Saghi. 2023. Exploring Techniques to Generate Dialogues in Hogwarts Legacy Game Using Large Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 8 pages.

1 Introduction

[1] Computing systems have become a major part of humans' everyday lives. Soon, it will not be unusual for us to interact with intelligent robots on a daily basis. However, computers still suffer from making sense of what humans are trying to do and why. Additionally, commonsense reasoning, i.e., reasoning about socially and culturally shared beliefs among humans, is still a challenge of intelligent systems. One example of human experience that involves commonsense knowledge and computers have lots of difficulties understanding is storytelling. Storytelling is an important part of how humans communicate, entertain, and teach each other. We tell stories to motivate each other to learn or to make fun in entertainment movies and computer games. As a result, it is important that computers become more capable of relating to humans by understanding stories, so that they can make more sense of our needs. We need to care about giving computers the ability called Computational Narrative Intelligence, i.e., the ability to understand and respond effectively to stories.

It is possible to do amazing things with computational narrative intelligence when the simulated world and its rules, characters, abilities, objects, and locations are well pre-defined. However, if we want the narrative system to tell stories in the real world and out of the simulation, which is complex and uncertain, then we run into challenges because these systems suffer from a lack of robustness and scalability.

First, to show the challenge of robustness in generating stories, here is an example story generated by TALESPIN [11].

> "*Henry Ant was thirsty. He walked over to the river bank where his good friend Bill Bird was sitting. Henry slipped and fell in the river. He was unable to call for help. He drowned.*"

[1] Introduction is inspired by Mark Riedl blog post: "Computational Narrative Intelligence: Past, Present, and Future" https://mark-riedl.medium.com

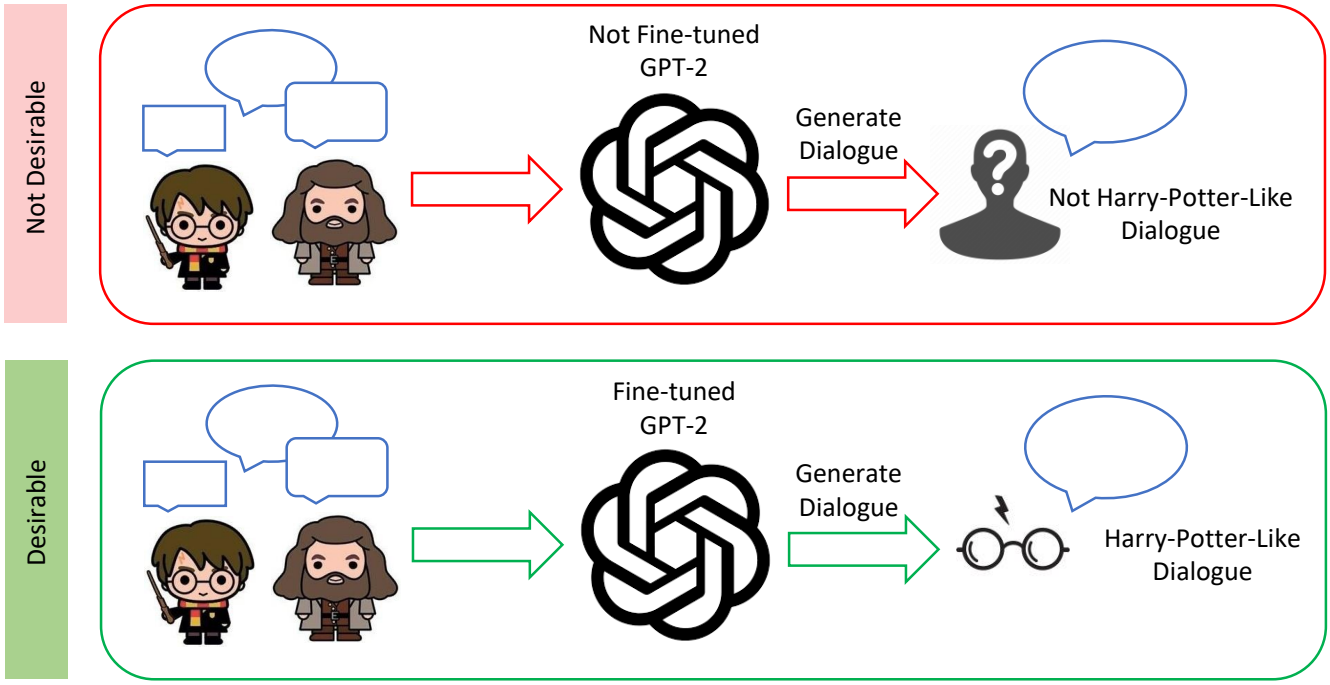

Figure 1. Our method overview. First row: using pre-trained GPT-2 with no fine-tuning may not satisfy generating dialogues that are consistent with the attributes of Harry Potter characters. Second row: if we fine-tune GPT-2 on HPD dataset that provides a lot of information about the scenes and relationships, then generated responses are more aligned with the theme of Harry Potter world.

However, it is mentioned in the paper that it was not supposed to happen. Falling into the river was deliberately introduced to cause the problem. Henry could call Bill for help. However, because of a rule in the system, saying "being in water prevents speech", Henry did not ask Bill directly, so Bill did not help Heny. This failure arose from an error in knowledge engineering and a lack of robustness in this narrative system that was not able to come up with a commonsense solution for the situation[11].

Secondly, to better understand the challenge of scalability in narrative systems, let us take a look at the symbolic story planning algorithm, Fabulist [13]. Fabulist paper talks about how to get a story. At first, rules about how the fictional world works should be provided. Then you have to provide facts about the entities in the world and their abilities. At last, the system with the outcome situation should be described. However, the issue with Fabulist is that one cannot really separate the algorithm's creativity from the creativity of the knowledge engineer since the generated story seems like just a reflection of the provided knowledge.

Having these into account, we need to build a robust and scalable story generation system that could generate a novel story about any topic conceivable. The first attempt to solve the open story generation problem was Scheherazade [8]. Scheherazade crowdsources stories about given topics, learns a generalized model, and uses the model to generate novel stories. Although Scheherazade can generate stories for any topic, it trades complexity for scalability, i.e., its stories are much shorter and structurally simpler.

It is where one may wonder if getting help from machine learning can provide robustness and scalability [5]. There has been great deal of work in which computational storytelling embraces machine learning. Automated Story Generation has been a research problem of interest since nearly the inception of AI. Learning to tell stories about any topic requires digesting large story corpora. Digesting this large amount of corpora requires large and deep neural networks with so many parameters. Stories are better generated using larger neural networks trained on larger amounts of textual corpora [6, 9]. A large language model is a deep learning algorithm with a large amount of parameters that can summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets [2]. Large language models are unlocking new possibilities in areas such as natural language processing and text generation. This motivates us to take advantage of state-of-the-art language models to generate stories.

[2]Nvidia blog post: "What Are Large Language Models Used For?" https://blogs.nvidia.com/blog

In this work, we narrow down the problem of story generation to dialogue generation for the very recent *Hogwarts Legacy* video game. We are interested in using recent state-of-the-art large language models, GPT-2 [12] and ChatGPT [3], to generate dialogues for different characters in the *Hogwarts Legacy* video game. We fine-tune pre-trained GPT-2 on *Harry Potter Novels* and Harry Potter Dialogue dataset (HPD) [3] to generate dialogues given the characters' attributes, characters' relations, context of the dialogue, and dialogue history. Then we compare the results with other language model baselines in terms of response quality, persona consistency, and creativity.

In the following sections, first, we present an overview of recent advances in automatic story generation in Section 2. Then we talk about our method and contributions in more detail in Section 3. Afterward, Section 4 represents the results of the experiments and shows some case studies. At last, we conclude our project and its findings in Section 6.

2 Related Work

[4] Automated story generation is the use of an intelligent system to produce a fictional story from a minimal set of inputs. Automated story generation is important because storytelling appears in many places and has many applications such as Human-AI coordination, training and education, etc. Besides that, story generation is one of the fundamental research questions in AI. In this section, we present an overview of non-machine learning-based and machine learning-based approaches for story generation.

2.1 Non-Learning Story Generation Approaches

This group of approaches mostly relies on knowledge bases containing hand-coded knowledge structures. TALESPIN [11] is acknowledged as the first intelligent story generator. It considers the plan as the story and applies some form of the symbolic planner to the problem of generating a fabula. However, one of the problems with symbolic planners is that enablement is not the only consideration for stories. Otherwise, it ends up looking like all characters are collaborating on the same goal. As a result, Fabulist [13] used a new planner that creates character goal hierarchies. After that it sub-selects actions for each character, it uses templates to render actions into natural language. Later on, Ware et al [15] modified Fabulist's planner system to avoid conflicts because conflicts break characters' plans.

In addition to all of the mentioned methods that are author-centric, i.e., a singular author is responsible for planning all the characters in them, there are character-centric simulations as well. In these systems, each agent is autonomous and responds to other characters based on its own intentions. For instance, Cavazza et al [2] present an interactive storytelling system in which agents can replan in response to a new injected fact.

2.2 Machine Learning Story Generation Approaches

Machine learning approaches do not necessarily use neural networks. These approaches are used for knowledge acquisition. For instance, Scheherazade [8] maintains a memory of plot graphs that shows the probable ordering of events on a particular topic. If the corresponding plot graph for a given story topic is not available for the system, it crowdsources example stories and then learns a plot graph and the most likely ordering constraints.

Those machine learning approaches that use neural networks start with a given prompt and then continuously generate tokens given the history of generated tokens so far. Khalifa et al [6] explored recurrent neural networks for the first time. However, the produced sentences were too random because stories in the training data are sparse, i.e., most of the sentences in a story might be unique to only that story and not seen in the others. This makes story learning for language models difficult. Martin et al [9] later addressed this challenge and reduced the sparsity by converting stories from natural language to event tuples that could learn to make better event sequences. Another limitation of using neural language models for story generation is that they need to remember the sequence of previously generated tokens, but as the story gets longer, more of the earlier context is forgotten. This is where large-scale pre-trained transformers [14] such as GPT-2 [12], GPT-3 [1], BART [7], BERT [4], ChatGPT, and baseines become helpful and make larger context windows applicable. For instance, Lara Martin [10] takes the advantage of the potential of GPT-2 to probabilistically generate story continuation based on a history of prior word tokens and constrains it with reasoning about causality.

2.3 Natural Language Prompt-based Learning

Following the introduction of neural language models for story generation, to increase the performance of pre-trained language models on a specific downstream task, one well-known pipeline is to take advantage of fine-tuning that pre-trained model on the provided dataset of the given task. However, recent advances in large language models that contain a very large amount of parameters (more than 150 billion) have brought a new way of using these models. Without updating the parameters of the model, we can handle a wide range of tasks by leveraging natural-language prompts and giving only the task description and a few examples of that.

The most similar work to our project is DialoGPT [16]. DialoGPT is a prompt-based tunable gigawordscale neural network model for conversational response (dialogue) generation that extends GPT-2 and is trained on Reddit data. It

[3] ChatGPT: Optimizing Language Models for Dialogue https://openai.com/blog/chatgpt/

[4] Related Work section is inspired by Mark Riedl blog post: "An Introduction to AI Story Generation" https://mark-riedl.medium.com

Figure 2. Examples of dialogue choices in Hogwarts Legacy video game.

first concatenates all dialogue turns within a dialogue session into a long text ended by the end-of-text token. Then, it feeds the obtained long text to the model as the source prompt to receive the target response sentence. They show that the text that is generated based on such user-specific prompt (or from scratch) is realistic-looking.

In this research project, we follow DialoGPT to model a dialogue session along with the attributes and the affiliations of the characters and frame the generation task as language modeling.

3 Method

Recent advances in training transformer-based architectures have achieved great empirical success. For example, OpenAI's GPT-2 [12] has demonstrated that transformer models trained on very large datasets can capture long-term dependencies and generate text that is fluent and rich in content. Such models have the capacity to generate text similar to what humans write in real-world. In this project, we extend GPT-2 to address the challenges of response generation. Dialogue or response generation is a subcategory of story generation and has the common objective of generating natural-looking text, dissimilar to instances seen during training, and relevant to the context of the dialogues and attributes of the speaker. One of the challenges of modeling dialogues or conversations is that potential responses could be so diverse and pose a great one-to-many problem. Another challenge is that dialogues are usually informal and noisy, so a general text dataset may not be fully informative and sufficient to train a response generation model.

In this work, we use GPT-2 that is pre-trained in an unsupervised way on a massive amount of text data and fine-tune it on Harry Potter Dialogue (HPD) dataset. This dataset facilitates the study of dialogue construction for characters within a story by providing rich contextual information about each dialogue session such as scenes, character attributes, and relations. For more details about this dataset, refer to Section 3.1. As a result, by fine-tuning GPT-2 on HPD dataset, we take advantage of rich linguistic patterns that GPT-2 has already learned during pre-training and specifically tune it to generate Harry-Potter-like dialogues that are fluent and are not duplicates of those in the original books. In this way, we aim to obtain a system that no longer suffers from *Harry Potter* content or style inconsistency.

Figure 1 shows an overview of our method. It is shown that using pre-trained GPT-2 with no fine-tuning may not satisfy generating dialogues that are consistent with the attributes of Harry Potter characters. However, if we fine-tune GPT-2 on HPD dataset that provides a lot of information about the scenes and relationships, then generated responses are more aligned with the theme of Harry Potter world.

This method can have many applications. One of the applications is in video games. For instance, instead of hard-coding the conversations for characters of a game and showing them in a conditional way based on the plot of the game, we can generate unique and not predetermined responses during the course of playing. This method also gives the opportunity of generating conversations for interactive games in which actions of a player are not known beforehand. *Hogwarts Legacy* video game is a relevant option to apply our method. As it is shown in Figure 2, there are some dialogue choices during the game that the player has to select one of them to continue the game. These options may or may not affect on the rest of the game. As a future work, developers of this game may use our method to generate unique, but related responses for characters in these question boxes instead of showing repetitive hard-coded ones.

3.1 Harry Potter Dialogue Dataset

The Harry Potter Dialogue (HPD) dataset [3] is the first dialogue dataset that integrates with scene, attributes and relations of Harry Potter Novels. The scene is the paragraphs including information about when, where, and why the dialogue took place. The attribute is a matrix that records the main attributes of each character, such as gender, age, spells, and belongings. The relation is also a matrix indicates the fine-grained relations between characters, such as Friend, Classmate, Teacher, Family, Lover, Opponent, Teammate, and Enemy. In total, it annotates 113 key characters, each having 12 relations to Harry Potter, and 13 unique attributes. For more details please refer to the paper.

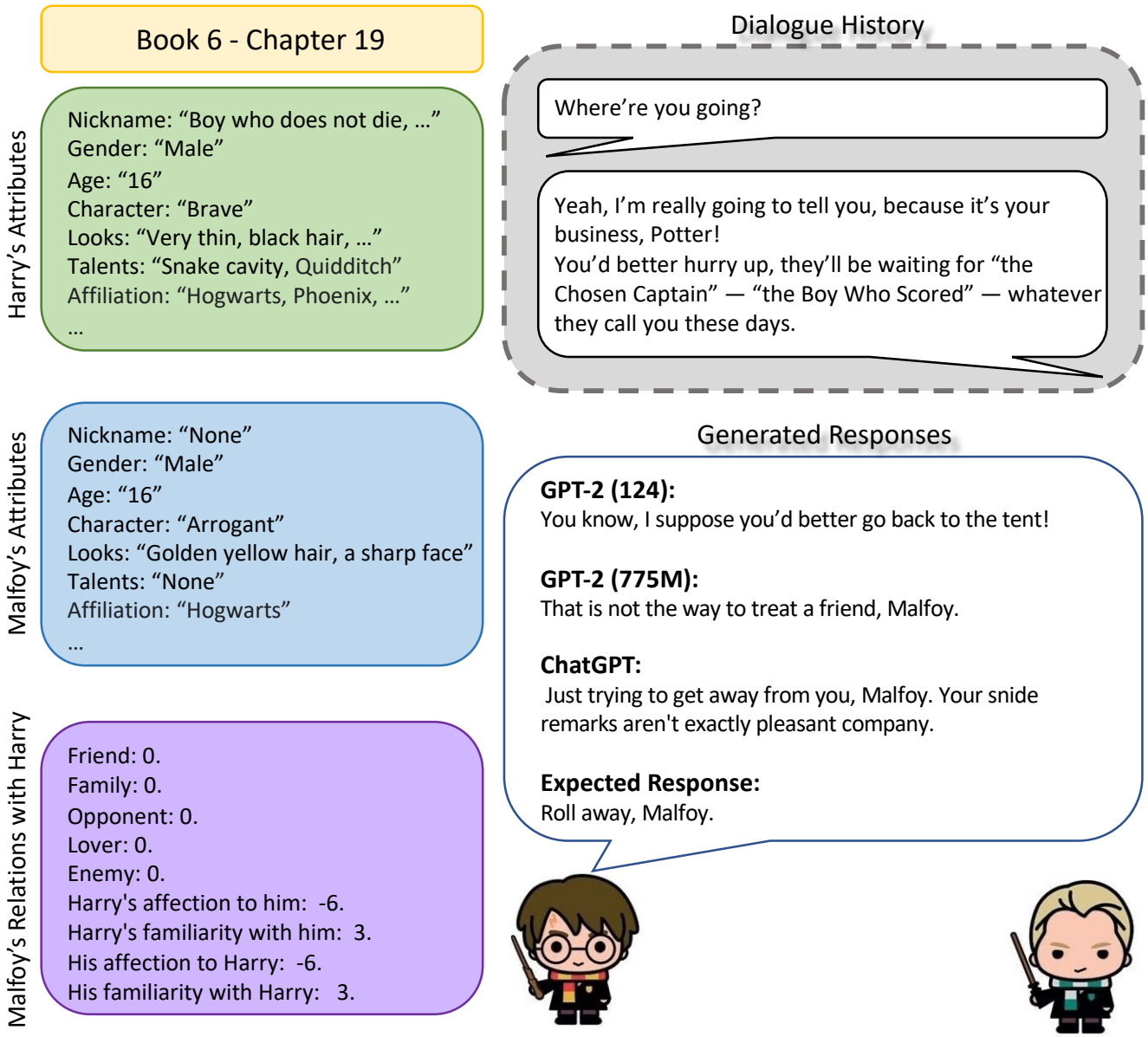

Figure 3. A case study. This example is a conversation between Harry and Malfoy. On the left, we can see each character's attributes and their relationship. On the right, the dialogue history is given. The benchmark models get all this information as input and generate the most likely responses. The original response is shown at the end. As it is shown, pre-trained GPT-2 (775M) that has not been tuned on this task does not produce Harry-Potter-like response.

4 Experiments and Results

In this section, we evaluate generated dialogues of our fine-tuned model and other baselines on HPD in terms of their quality, persona consistency, and creativity.

Dataset. As mentioned in Section 3 in detail, we use HPD dataset [3] to fine-tune GPT-2 and evaluate its ability compared to other baselines in terms of generating creative Harry-Potter-like dialogues that makes sense to humans. HPD dataset consists of 1042 dialogue sessions as the training set and further 178 dialogue sessions as test set. Each session in the test set consists of at least one positive response and 9 negative responses.

Baselines. We compare the performance of fine-tuned GPT-2 (124M) with **1)** pre-trained GPT-2 (775M) and **2)** ChatGPT. Comparing with these models that are not specifically tuned for the task of Harry-Potter-like dialogue generation will

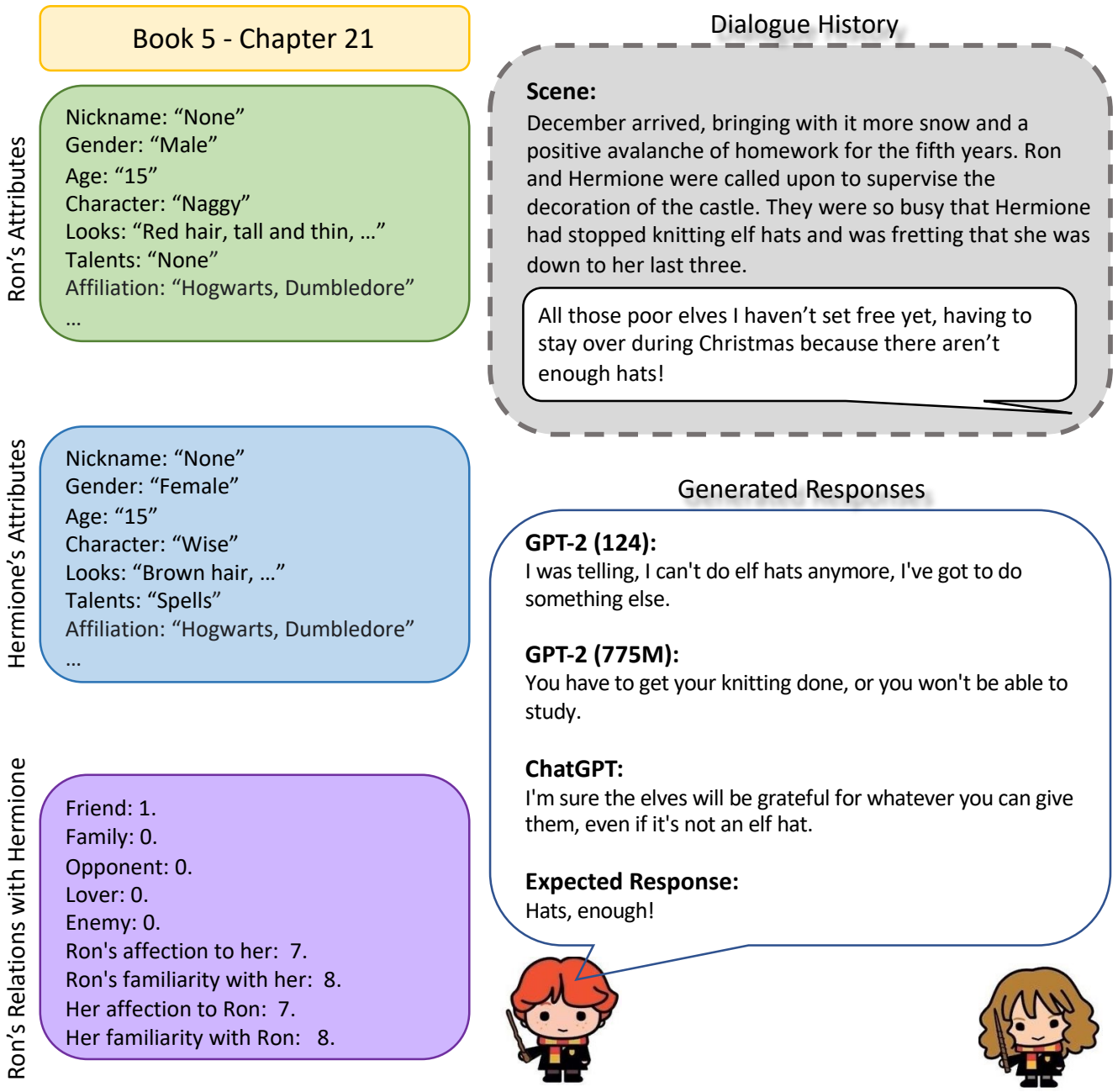

Figure 4. A case study. This example is a conversation between Ron and Hermione. On the left, we can see each character's attributes and their relationship. On the right, the scene and dialogue history are given. The benchmark models get all this information as input and generate the most likely responses. The original response is shown at the end.

show how consistent the general language models with rich linguistic patterns can be.

Implementation details. We fine-tuned official pre-trained GPT-2 architecture with 124 million parameters on HPD dataset with batch-size=16 and learning-rate=2e-5 for 1000 steps. GPT-2 transformer model adopts the generic transformer language model and leverages a stack of masked multi-head self-attention layers to train on the training data. In our case, the text is generated based on a user-specific prompt in the test time. Our implementation uses Tensorflow and is based on gpt-2-simple repository [5].

[5]gpt-2-repository: https://github.com/minimaxir/gpt-2-simple

Table 1. Human evaluation results of HPD test set. **Ft. GPT-2 (124M)** is fine-tuned GPT-2 with 124M parameters and **Pt. GPT-2 (775M)** is official pre-trained GPT-2 with 775M parameters. Relv.Sce, Flu, Relv.Att, and Relv.Re indicate Relevance with the scene, Fluency, Relevance with the character attributes, and Relevance with the character relations, respectively.

Model	Relv.Sce.	Flue.	Relv.Att.	Relv.Re.
Ft. GPT-2 (124M)	2.23	4.06	2.74	2.88
Pt. GPT-2 (775M)	2.82	4.31	2.71	2.66
ChatGPT	4.43	4.63	3.88	4.20

Evaluation metrics. We evaluate the response from models from three main aspects: response quality, persona consistency, and creativity. We randomly selected 7 samples out of 20 generated samples and requested 15 annotators to rate them in terms of four criteria: 1) relevance with the scene 2) fluency 3) relevance with the character attributes 4) relevance with the character relations. Each criterion is rated on a five-scale, where 1, 3 and 5 represent weak, moderate and perfect performance, respectively.

4.1 Human Evaluation Results on HPD Dataset

Table 1 shows the human evaluation results. We can observe that all models perform well on Fluency, where ChatGPT's score, 4.63, is close to perfect performance. However, the performance of models in terms of relevancy to scene, attributes, and relations is relatively lower. Although pre-trained GPT-2 can generate dialogues that are more relevant to scene, but fine-tuned GPT-2 outperforms it in generating responses that are more relevant to character attributes and relations. This is due to the fact that GPT-2 (775) has learned richer linguistic patterns, but our fine-tuned model has been specifically tuned for the task of Harry Potter-like dialogue generation. Similarly, ChatGPT scores are relatively higher except for Relv.Att. Overall, the results indicate although generated responses seem natural but can not be qualified as Harry Potter-like responses because they do not capture the attribute and relation information well.

4.2 Case Studies

Figure 3 and 4 show two sampled cases. Please refer to their captions for more details.

5 Discussion and Future Work

After looking at the results presented in this work, we can observe that existing approaches, even state-of-the-art ChatGPT, fall short of generating Harry-Potter-like responses. Even tuning the model on character relation information and leveraging them at test time does not help much. This proves that there is still much room to explore for the community.

6 Conclusion

In this work, we take advantage of language models to generate conversational responses. We fine-tune pre-trained GPT-2 (124M) model on Harry Potter Dialogue dataset to incorporate characters' attributes, relations, context, and dialogue history, such that we can generate creative and Harry-Potter-like dialogues for Hogwarts Legacy video game. We compare the results of this model with pre-trained GPT-2 (775M) and ChatGPT in terms of fluency, relevance to scene, relevance to attributes, and relevance to relations. The results show that although language models are capable of producing fluent dialogues, they still fall short of generating responses with persona consistency.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[2] Marc Cavazza, Fred Charles, and Steven J. Mead. 2001. Characters in Search of an Author: AI-Based Virtual Storytelling. In *International Conference on Virtual Storytelling*.

[3] Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Ziyang Chen, and Jia Li. 2022. What would Harry say? Building Dialogue Agents for Characters in a Story. *arXiv preprint arXiv:2211.06869* (2022).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. https://aclanthology.org/N19-1423

[5] Brent Harrison and Mark Riedl. 2016. Learning from stories: using crowdsourced narratives to train virtual agents. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 12. 183–189.

[6] Ahmed Khalifa, Gabriella AB Barros, and Julian Togelius. 2017. Deeptingle. *arXiv preprint arXiv:1705.03557* (2017).

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[8] Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27. 598–604.

[9] Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[10] Lara Jean Martin. 2021. *Neurosymbolic automated story generation*. Ph. D. Dissertation. Georgia Institute of Technology.

[11] James R Meehan. 1977. TALE-SPIN, An Interactive Program that Writes Stories.. In *Ijcai*, Vol. 77. 91–98.

[12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[13] Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39 (2010), 217–268.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[15] Stephen G. Ware, Robert Michael Young, Brent E. Harrison, and D. Roberts. 2014. A Computational Model of Plan-Based Narrative Conflict at the Fabula Level. *IEEE Transactions on Computational Intelligence and AI in Games* 6 (2014), 271–288.

[16] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 270–278. https://aclanthology.org/2020.acl-demos.30