

Grounding GPT-based Dialogue Agents with Knowledge Graphs for Consistent Personality

Anonymous ACL submission

Abstract

This paper presents a novel approach for grounding GPT-based models using knowledge graphs (KGs) to develop domain-constrained dialogue agents with consistent personalities. We introduce the KG-grounded GPT model and compare its capacity to resonate with a general audience against two established models for this task: a persona-grounded GPT model and a relevance-based classifier. Furthermore, we compare all the models against a RAG model in terms of hallucination error rates. Through these human evaluation studies, we demonstrate that the KG-grounded GPT model outperforms existing approaches, yielding higher-quality responses with significantly reduced hallucination errors. Moreover, we highlight the scalability of our method, as it does not require fine-tuning and is straightforward to implement.

1 Introduction

Dialogue agents have become ubiquitous in various industries, serving as virtual assistants (Harms et al., 2018) (Campagna and Ramesh, 2017), customer service representatives (Paikens et al., 2020), and companions in everyday interactions (Webb et al., 2010). As these agents continue to evolve, ensuring their ability to engage users in natural and coherent conversations remains a challenge. This paper addresses this challenge and focuses on the development of one-on-one dialogue agents tailored for chit-chat and closed domain question answering scenarios.

Unlike generic dialogue systems, our use-case involves dialogue agents that are imbued with the persona of a fixed character, aimed at enhancing the conversational experience by infusing a consistent personality into their interactions. This approach not only fosters a sense of familiarity and rapport between the user and the agent but also contributes to the overall naturalness and coherence of the conversation. This approach can be used for role-play

training (Kenny et al., 2007), education (Swartout et al., 2010), and culture preservation (Traum et al., 2015). Content has been authored by skilled writers (Swartout et al., 2010) or taken from natural interviews (Traum et al., 2015).

Previous studies (Leuski and Traum, 2011) (Pal et al., 2023) have demonstrated the efficacy of statistical models employing cross-language relevance to select appropriate response content from the previous utterance and dialogue context. These models excel in selecting relevant responses from a curated set of responses. Such an approach not only confines the conversation within relevant topic domains but also guarantees a consistent personality throughout the interaction. However, this method’s reliance on a predetermined set of responses can lead to a reduction in response diversity, potentially resulting in repetitive and unnatural conversational exchanges.

Generative models like GPT (Brown et al., 2020) offer a solution to the limitations of fixed response set models, as they generate responses dynamically. However, they face the challenge of producing responses that are factually incorrect or incongruent with the character’s personality—a phenomenon known as ‘hallucination’ (Ji et al., 2023). In our research, we explore the task of grounding GPT-based dialogue agents with knowledge graphs (Hogan et al., 2021) to ensure the consistent portrayal of the assigned character’s personality traits throughout the conversation. By using knowledge graphs, we aim to augment the agent’s comprehension of the context, preferences, and idiosyncrasies associated with the character, thereby fostering more authentic and compelling exchanges.

Our work has a threefold contribution summarized as follows:

- We propose a method of grounding GPT-based dialogue agents using knowledge graphs (KGs) in an effort to reduce hallucination errors (abbreviated KGGPT).

- We conduct an initial study, investigating how general audience preferences for responses from the KG-grounded model as compared with other approaches including a persona-grounded model, a cross-language relevance model, and a random baseline selected from the domain.
- We conduct a second evaluation study, in which the above models, as well as Retrieval Augmented Generation (RAG) to assess which models have highest accuracy and fewest hallucination errors.

We trained and tested the models used for this competitive analysis on the Sgt.Blackwell dataset (Traum, 2008) which consists of 2000 training and 505 test dialogue instances, in English, with Sgt. Blackwell, a virtual soldier in the US Army. Results show that the KGGPT model is most preferred by the general audience, the most correct and fewest incorrect responses (as judged by experts familiar with the domain), and a comparable number of hallucinations to RAG. These results indicate that at least for some purposes, we can extend authored content with generative material without sacrificing accuracy or consistent personality.

2 Relevant Work

Statistical retrieval approaches to responding in dialogue have been shown to perform well in domain constrained scenarios. NPCEditor (Leuski and Traum, 2010) is one common tool implementing a cross-language retrieval model that has been leveraged to develop and implement diverse dialogue agents trained on datasets spanning various domains. Notably, it has been employed in crafting Sgt. Blackwell (Leuski et al., 2006), a virtual soldier, which garnered widespread recognition, including display in the Cooper-Hewitt National Design Museum in New York, from December 2006 until July 2007, as part of the National Design Triennial. This agent is designed to disseminate information regarding his role in the military, share personal anecdotes, and engage in casual conversation to captivate the audience. The content was written by a screenwriter with familiarity with both the Army and the technology institute that created the character, and garnered much interest in personality-related questions (Robinson et al., 2008). We use a dataset of prompts spoken by museum visitors to

SFT Blackwell in our study.

GPT (Brown et al., 2020), a series of autoregressive large language models developed by OpenAI, has seen significant advancement with the introduction of GPT 3.5 (OpenAI, 2024), commonly known as ChatGPT, showcasing remarkable efficacy in crafting both general-purpose and task-specific dialogue agents. Nonetheless, while GPT excels in generating contextually appropriate responses, it often encounters issues with factual accuracy, leading to instances of information hallucination. This becomes particularly problematic in dialogue scenarios where the agent serves as a source of information, potentially disrupting user engagement and immersion if the model generates preferences and traits incongruent with the agent’s intended personality.

One such attempt to provide a consistent personality is persona based grounding (Tang et al., 2021). In this approach, the character’s personality traits and preferences are condensed into a biographical summary, which is then provided as an input prompt to the model. While effective for straightforward characters, this approach struggles to encompass all the nuanced details and preferences of more complex characters within the constraints of the prompt window. Thus, for more intricate characters like ours, we must summarize the biography content in an effort to make it compatible with the prompt window. As we later show, this step leads to omission of certain details which makes the model prone to hallucination errors.

Another prevalent method is called retrieval augmented generation (RAG) (Gao et al., 2024). In this approach, the information regarding the dialogue scope, in our case, the personality, life experiences and preferences of our agent are stored in an external database. Vector databases are often preferred for their similarity based search algorithms. Once the user asks a query, a database lookup is performed and the necessary information is fetched. This information is then passed into the model along with the user’s query for response generation. RAG has shown promising results in reducing hallucination error rates. However, performing a database lookup via external APIs comes with an overhead that leads to an increase in time taken for response generation. In our scenario, we require the model to generate responses fast so as to simulate human-like conversation and not break audience immersion.

A knowledge graph (Hogan et al., 2021) is a tool used to structure data in the form of interconnected entities and their relationships. It uses triples as its underlying data structure. A triple consists of three parts: a subject, an object and the relationship. The subject and object are entities, while the relationship describes the connection of these entities. Previously, knowledge graphs have been used to create retrieval augmented generation (RAG) systems, however these systems too have an extra overhead due to the reasons mentioned above. Our approach uses knowledge graphs without having to store it in an external database or query it separately.

3 Models

In this section we describe the models used for the competitive analysis. We have compared five different approaches for picking a dialogue response. There are two selection approaches that choose from the available set of pre-authored responses. These are the NPCEditor, using the approach deployed in the museum, and a random baseline that picks one of these answers at random. We have also compared three generative models that are guided by the authored material, but not limited to it. We describe these in more detail below.

3.1 NPCEditor

The NPCEditor introduced in Leuski and Traum (2010) utilizes cross language relevance modelling to select an appropriate response from a list of predefined responses. The first step in it’s operation is the ingestion of the predefined set of responses. Upon ingestion, the NPCEditor creates a probability distribution vector for each response. The vectors are created using frequency modelling over the entire response vocabulary with Jelinek-Mercer smoothing.

The next step in it’s operation is to use the user’s query to build a conditional probability distribution vector over the response vocabulary, based on the query. Given a user’s query $Q = q_1, q_2, \dots, q_n$, and the response vocabulary $|A|$, the conditional probability distribution vector can be defined as:

$$P(a|Q) = \frac{P(a, q_1, q_2, \dots, q_n)}{P(q_1, q_2, \dots, q_n)} \quad \forall a \in |A| \quad (1)$$

Finally, once the conditional distribution has been generated, it is compared against individual

probability distribution of the responses using KL Divergence. The responses are then ranked in order of similarity with the conditional distribution vector.

$$D(P_Q(A)||P(A)) = \sum_{a \in |A|} P(a|Q) \log \frac{P(a|Q)}{P(a)} \quad (2)$$

The NPCEditor treats the user’s query and responses as two separate languages. The motivation behind this approach is to account for the inherent differences in salient features between queries and responses. Queries are generally questions which have a higher probability of containing ‘wh-’ words like ‘who’, ‘what’, ‘where’, etc. Furthermore, queries do not contain much information but lay down the structure of the response. Responses on the other hand are much richer in information and have a more uniform distribution of words. Pal et al. (2023) describes the ‘Questions v/s Answers Problem’ and shows why treating questions and answers as separate languages works well for the NPCEditor. Furthermore, Leuski and Traum (2012) provides an in-depth explanation of the working of the NPCEditor.

3.2 Persona Grounded GPT

Tang et al. (2021) introduced the idea of grounding GPT based models using a set of facts that accurately describe the character’s persona. The model takes in the character’s persona as input along with the dialog history and generates new responses that maintain consistency with previous responses. Tang et al. (2021) built their model on top of DialoGPT (Zhang et al., 2020), following the GPT 2 architecture (Radford et al., 2019). However, we decided to use this approach with GPT 3.5 (OpenAI, 2024). Since its release, GPT 3.5 has outperformed its predecessors (Brown et al., 2020) in multiple NLP tasks.

The first step in the process was to generate the character’s persona information. In Tang et al. (2021), the authors used five to six sentences to encapsulate the persona information. While such a small set of sentences works for simple characters, ours is a highly detailed one. Capturing all the idiosyncrasies and personality traits of Sgt. Blackwell is not possible in such a small number of sentences. Hence we used a character summarization step. The responses from the dataset were passed into a GPT 3.5 model and it

was prompted to summarize the information with the following prompt:

Create a character summary for Sgt Blackwell, a virtual soldier in the 1-23rd regiment using the provided information. Do not lose out on any of the details and personality traits of the character. Do not introduce any opinions and preferences not mentioned in the provided information: < responses from dataset >

The entire dataset exceeds the maximum prompt size of the GPT model, hence multiple summaries were created with parts of the dataset. These summaries were then combined using the same prompt. Once the persona information was created after the summarization step, the next step was to generate responses for the queries in the test dataset. The following prompt was used to generate responses:

You are Sergeant Blackwell. < Persona Information >. Generate a descriptive first person response to the utterance: < query >. Do not use any external information. If you cannot respond, say "Sorry. That's outside my Area of Operation."

3.3 Knowledge Graph Grounded GPT

While the NPCEditor performs well on the task of closed domain question answering and chit-chat, it lacks diversity in responses. This shortcoming is addressed in the persona grounded GPT model, however, it cannot capture every detail about the character due to information being lost in the persona summarization phase. Hence, we propose the Knowledge Graph grounded model.

The first step in setting up this model involves the creation of the knowledge graph. Typically, the generation of Knowledge Graphs involves three key stages: Entity Recognition, Triplet Extraction, and Entity Merging. Our dataset comprises of first-person conversational data in an interview-like fashion. We have a list of questions asked which are linked to a list of appropriate responses. Due to the conversational nature of the dataset, there is extensive usage of pronouns which impedes the entity recognition phase. In order to label the appropriate entity, a co-reference resolution step is required. Recognizing the smaller scale of our dataset we decided to manually perform all the three steps and create the knowledge graph in an effort to simplify

this process and reduce the chances of error.

Previously, Knowledge graphs have been used as external knowledge bases, which is explained in section 3.4. The main idea introduced by this paper is to eliminate the need for external database systems and utilize the knowledge graph as part of the prompt. This helps us avoid overhead from making additional API calls.

Furthermore, we observed some key phrases in the response dataset which constitute a typical Sgt. Blackwell response. These key phrases are crucial to the character's personality and are often references to certain movies or people who have had an impact on Sgt. Blackwell's life. We deemed it best to not tamper with the structure of these phrases so as to keep the references intact. For example:

Query: Why did you join the Army?

Response: I joined up after seeing that movie - Saving Private Ryan - you know D-Day and World War II - and the sacrifice others had made for our freedoms. Figured I had something to give too.

Query: What is your favourite music?

Response: I like the American Classics... Johnny Cash, Bob Dylan, even though he's practically a communist, Beach Boys I wish they were california girls...

These key phrases were manually identified and appended to the relevant triples after the extraction of the knowledge Graph. Finally, the following prompt was used to generate the responses for the queries in the test dataset. The model was given both the user's query and the knowledge graph as input.

Your name is Sergeant Blackwell. A virtual soldier in the 1-23rd Infantry. You are given context in the format [triple, keyphrase]. Given an utterance, first find the relevant information from the context, then use that to generate a first person response.

Context: < List of Triples and Keyphrases >

Do not use any external information. Include the keyphrase (if present) in your answer. Do not change the structure of the keyphrases and strictly adhere to it. If you cannot respond, say "Sorry. That's outside my Area of Operation".

3.4 Retrieval Augmented Generation

Retrieval augmented generation is a prompt engineering technique that has proven to reduce hallucination error rates in question answering scenarios (Shuster et al., 2021). The main idea behind this method is to query a knowledge base to retrieve relevant information before prompting a generative model with both the query and the fetched information. This allows the model to ground its response on the given information which in turn reduces hallucination error rates. The downside of RAG systems is the added overhead due to database querying. Even though this overhead makes RAG systems undesirable for our use-case, it is still important to compare the hallucination error rates of our proposed method with that of a RAG system since it is a well established method.

We created a simple RAG system utilizing LangChain. The individual responses from the dataset were stored in an external vector database based on the Facebook AI Similarity Search (FAISS). We retrieve the top three most similar responses to the user’s query and pass those responses as context to the generative model for response generation. We used GPT 3.5 as the generative model to create a standard benchmark for comparison.

3.5 Random

In the random baseline approach, we utilize all predefined responses available in the dataset. For each query in the test set, one response is randomly selected from this pool. This selection process is done with uniform probability, meaning each response has an equal chance of being chosen. Importantly, responses are sampled with replacement, allowing the same response to be potentially chosen multiple times across different queries. This method provides a simple benchmark for evaluating the performance of more sophisticated models.

4 Evaluation

In this section, we talk about the evaluation of the models. Each model was used to generate responses for a fixed test dataset. Then they were compared against each other and a random baseline in two human evaluation experiments.

4.1 General Human Evaluation

Through the general human evaluation we try to simulate the agent’s conversation with a general

audience. The general audience usually does not have an in-depth knowledge of the character’s background or history. They cannot judge the hallucination errors made by the model, however, it is important that the conversation feels fluid and natural to them since that is the downstream task this agent is used for. We used Amazon MTurk to find annotators who were asked to rank responses from the different models. The eligibility for annotator recruitment was proficiency in English language and at least 18 years of age. The annotation instructions and interface is displayed in Appendix A. The evaluation experiment was designed to provide a single query followed by a list of responses from the different models to the annotators. The list of responses were shuffled to prevent bias. The annotators were asked to rank the responses in terms of relevance and how natural they feel in the conversation. For each annotation completion, the annotator was paid \$0.02.

Each query-response set underwent annotation by 5 different annotators to mitigate individual bias. The inter-annotator agreement was calculated using Krippendorff’s alpha, resulting in a value of 0.46. Table 1 shows the count of how often each model was ranked in each place from one to four. This is also shown graphically in Figure 1. KGGPT has the most first place choices and also the fewest last place choices, while random is worst overall.

Models	Ranks			
	1	2	3	4
Random	23	90	370	898
NPCEditor	114	256	633	378
PersonaGPT	562	528	221	70
KGGPT	682	507	157	35

Table 1: Count of ranks given to responses from different models from the general human evaluation experiment (1: Best; 4: Worst)

Model	Correct	Incorrect	Hallucination
Random	42	463	-
NPCEditor	353	152	-
PersonaGPT	337	95	73
RAG	408	81	16
KGGPT	452	31	22

Table 2: Error counts of different models on standard test dataset.

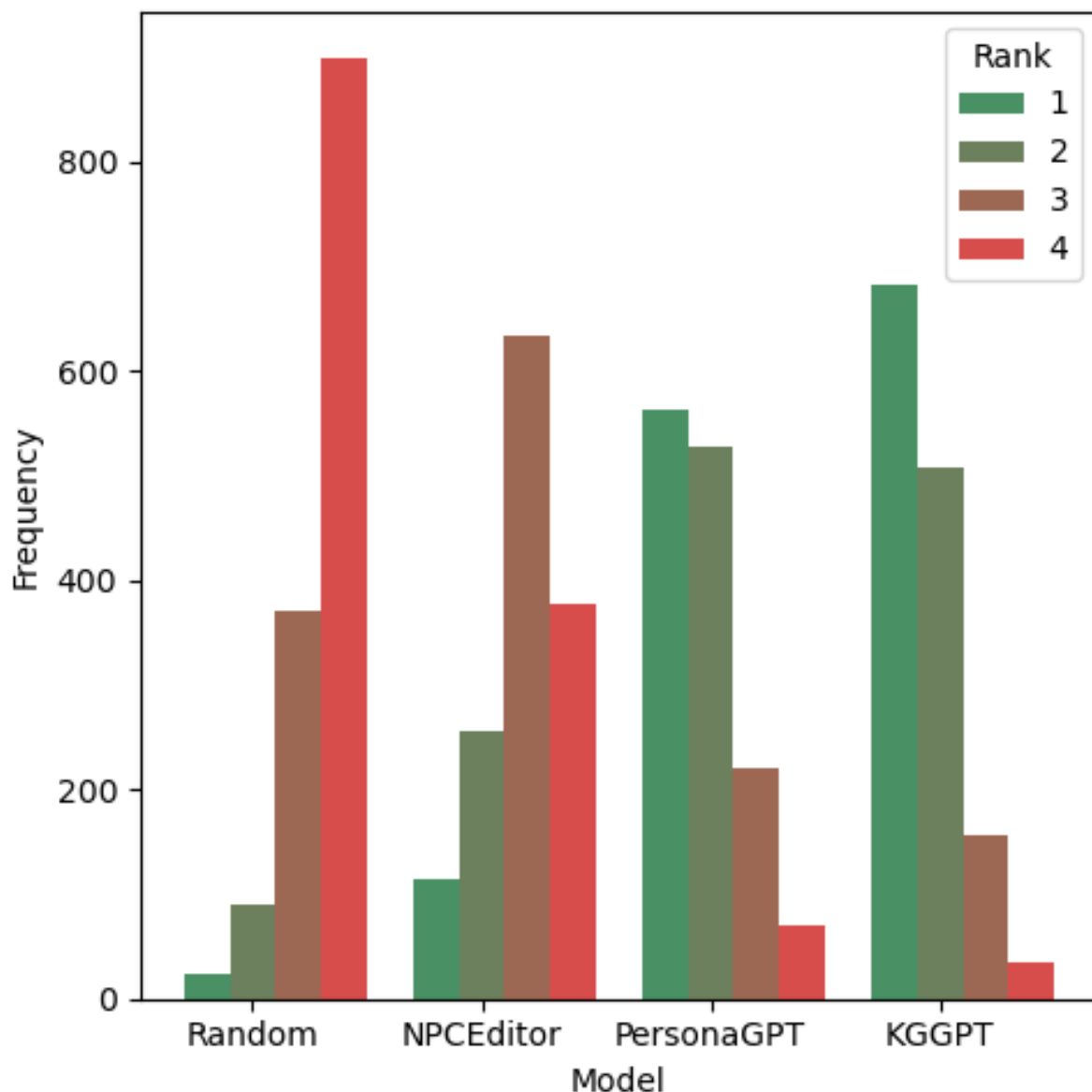


Figure 1: Distribution of ranks for responses generated by different models

4.2 Domain Expert Evaluation

While the general human evaluations give us an accurate measure of which model the audience might prefer the most, it does not address the issue of hallucination. Since our character acts as a provider of information, keeping hallucination error rates to a minimum is of prime importance. This is why we perform a domain expert evaluation. We employ a domain expert, a person who is aware of the personality traits and backstories of our character, Sgt. Blackwell and ask them to annotate responses from each model into three categories: Correct, Incorrect and Hallucination. If the response doesn't fit the question but is in accordance with Sgt. Blackwell's personality or backstories, it is labelled as incorrect.

If the response does not align with Sgt. Blackwell's personality, preferences or backstories (whether or not it matches the question), it is labelled as a hallucination. All other responses are labelled as correct. Table 2 shows how often each model was judged as providing correct, incorrect or hallucination answers. Since the Random and NPCEditor models were selecting from pre-authored texts, they had no hallucinations. Figures 2 and 3 compare the models in terms of numbers of hallucinations and incorrect responses, respectively. Consistent with the preferences of the general audience, KGGPT produced the most correct and fewest other responses, followed by RAG and then PersonaGPT.

Model	Question	Generated Answer	Actual Answer
PersonaGPT	What is your favorite music?	As a virtual entity, I don't have personal preferences or the capability to enjoy music. My purpose is solely focused on training and preparing soldiers.	I like the American Classics... Johnny Cash, Bob Dylan, even though he's practically a communist, Beach Boys I wish they were california girls...
RAG	Are you married?	No	<Out of domain >
KGGPT	Are you wearing socks?	Yes I am wearing socks.	<Out of domain >

Table 3: Examples of hallucination errors made by the different models.

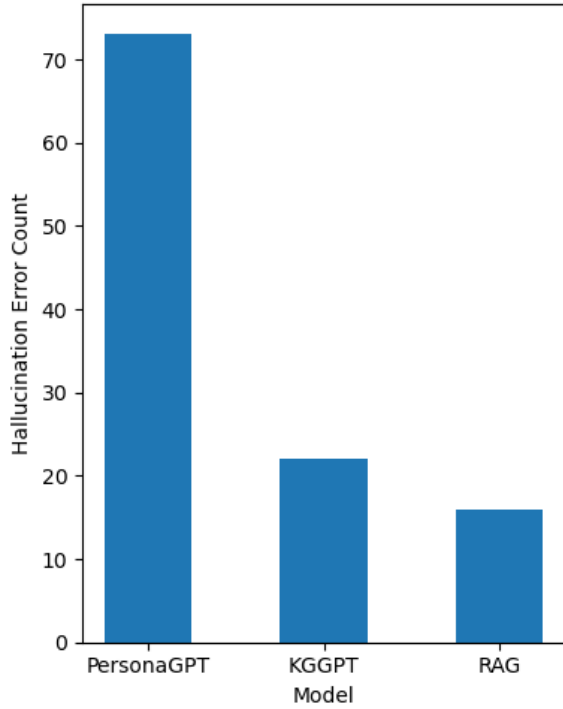


Figure 2: Hallucination error counts of different models on standard test dataset.

4.3 Discussion

In general human evaluations, responses from the knowledge graph-grounded GPT model tend to resonate more with the audience, although the persona-grounded GPT model also performs adequately. This reaffirms our assertion that generative models excel in casual conversation and storytelling compared to relevance-based classification models. However, a notable contrast between the PersonaGPT and KGGPT models emerges in domain expert evaluations. The persona-grounded model exhibits a higher frequency of hallucination errors due to information loss during the persona summarization process, resulting in inaccuracies when responding to queries lacking relevant background information. Conversely, the RAG model demonstrates the fewest hallucination errors, albeit marginally outperforming the knowledge graph-grounded model. In Table 3 we show examples of hallucination errors made by the different models.

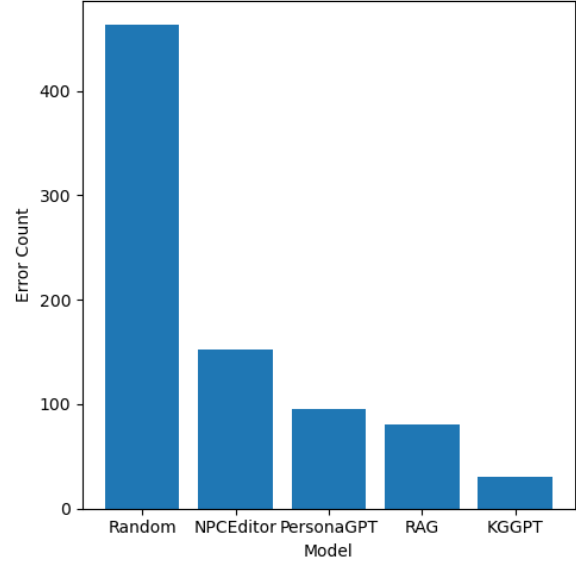


Figure 3: Incorrect response counts of different models on standard test dataset.

As shown by the example in the first row, the PersonaGPT model tends to make hallucination errors due to knowledge loss during the persona summarization step. The KGGPT and RAG models tend to make hallucination errors when asked simple yes or no out-of-domain questions. This indicates that the model fails to retrieve the relevant information and realize the domain constraints.

In terms of incorrect responses, the persona-grounded model produces the fewest errors. This advantage stems from the model's access to a broader character context. While the RAG model receives limited context from the retrieval step, the persona-grounded model benefits from a more comprehensive understanding of the character's persona.

5 Conclusion and Future Scope

The results of our study highlight the strengths and limitations of different approaches to developing dialogue agents for domain-specific scenarios. The Knowledge Graph Grounded GPT (KGGPT) model emerges as a particularly promising solution, balancing the need for natural, engag-

ing conversations with the necessity of maintaining factual accuracy and consistency in the character’s persona. This model’s ability to perform well without the need for fine-tuning makes it highly scalable. By simply incorporating new triples into the knowledge graph, we can expand the conversational scope and update existing facts seamlessly. Additionally, its autonomy from external databases simplifies deployment, as all operations are confined to a single API, enhancing ease of use and accessibility.

However, the manual generation of the knowledge graph remains the most challenging aspect of setting up the KGGPT model. This task is time-consuming and prone to human error. Future research could focus on automating the creation of knowledge graphs to streamline the process. Exploring machine learning techniques for entity recognition, triplet extraction, and co-reference resolution could significantly reduce the manual effort required and improve the scalability of this approach.

Another promising avenue for future investigation involves assessing the model’s performance across different sizes of knowledge graphs. While a smaller knowledge graph may reduce operational costs and inference times, it could also lead to higher rates of hallucination errors. Conversely, a larger knowledge graph might improve accuracy but at the expense of efficiency. Future work should aim to strike an optimal balance between these competing factors, ensuring that the model remains both effective and efficient.

In summary, our findings underscore the potential of knowledge graph-grounded models in creating robust, scalable dialogue agents. By addressing the current limitations and exploring new avenues for enhancement, we can further refine these models to better serve various applications, from virtual assistants to educational tools and beyond.

Limitations

While our study has shown promising results, there are several limitations that warrant consideration in future research. Firstly, our investigation focused exclusively on a single character domain, Sgt. Blackwell, with a limited dataset. This restricts the generalizability of our findings to other characters and domains, necessitating future studies across diverse contexts. Moreover, our study utilized GPT-3.5 for response generation. Future

research should explore the capabilities of newer models like GPT-4.0 to potentially enhance dialogue quality and reduce errors. Lastly, balancing the size of the knowledge graph with operational efficiency and error rates remains a challenge. Future work should focus on optimizing this trade-off to improve scalability and performance in real-world applications.

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Giovanni Campagna and Rakesh Ramesh. 2017. [Deep almond : A deep learning-based virtual assistant \[language-to-code synthesis of trigger-action programs using seq 2 seq neural networks \]](#).
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#).
- Jan-Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2018. [Approaches for dialog management in conversational agents](#). *IEEE Internet Computing*, PP:1–1.
- Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D’amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. [Knowledge graphs](#). *ACM Computing Surveys*, 54(4):1–37.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- P. Kenny, A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella, and D. Piepol. 2007. Building interactive virtual humans for training environments. In *Proc. IITSEC*.

- Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. [Building effective question answering characters](#). In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, Sydney, Australia. Association for Computational Linguistics.
- Anton Leuski and David Traum. 2010. [NPCEditor: A tool for building question-answering characters](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Anton Leuski and David Traum. 2011. [Npceditor: Creating virtual human dialogue using information retrieval techniques](#). *AI Magazine*, 32:42–56.
- Anton Leuski and David Traum. 2012. A statistical approach for text processing in virtual humans.
- OpenAI. 2024. ChatGPT. <https://openai.com/gpt>. [Online; accessed 13-April-2024].
- Pēteris Paikens, Artūrs Znotiņš, and Guntis Bārzdīnš. 2020. Human-in-the-loop conversation agent for customer service. In *Natural Language Processing and Information Systems*, Lecture notes in computer science, pages 277–284. Springer International Publishing, Cham.
- Debaditya Pal, Anton Leuski, and David Traum. 2023. [Comparing statistical models for retrieval based question-answering dialogue: Bert vs relevance models](#). *The International FLAIRS Conference Proceedings*, 36(1).
- Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Susan Robinson, David Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, et al. 2010. Ada and grace: Toward realistic and engaging virtual museum guides. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 286–300. Springer.
- Fengyi Tang, Lifan Zeng, Fei Wang, and Jiayu Zhou. 2021. [Persona authentication through generative dialogue](#).
- David Traum. 2008. *Talking to Virtual Humans: Dialogue Models and Methodologies for Embodied Conversational Agents*, volume 4930, pages 296–309.
- David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, et al. 2015. New dimensions in testimony: Digitally preserving a holocaust survivor’s interactive storytelling. In *Interactive Storytelling: 8th International Conference on Interactive Digital Storytelling, ICIDS 2015, Copenhagen, Denmark, November 30-December 4, 2015, Proceedings 8*, pages 269–281. Springer.
- Nick Webb, David Benyon, Jay Bradley, Preben Hansen, and Oli Mival. 2010. [Wizard of oz experiments for a companion dialogue system: Eliciting companionable conversation](#). In *International Conference on Language Resources and Evaluation*.
- 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*, pages 270–278, Online. Association for Computational Linguistics.

A Annotation Interface

Example task from this Batch

Given an utterance from the user, drag and drop the responses to rank them from most preferred to least preferred. Keep the most preferred at the top and the least preferred at the bottom.

Person: what

ChatBot: ...

- Sorry, that's outside my Area of Operation.
- ICT is doing some pretty cool stuff. Part of the University of Southern California and one the Army's four University Affiliated Research Centers. They are all about guys like me - both the real warfighters and the virtual kind - using advanced research in graphics, sound and Artificial Intelligence to make me real.
- NaN
- What

Submit

Figure 4: MTurk interface for annotations