Beyond Simple Personas: Evaluating LLMs and Relevance Models for Character-Consistent Dialogue

Debaditya Pal and David Traum
Institute for Creative Technologies
University of Southern California
USA
debaditya.pal6@gmail.com, traum@ict.usc.edu

Abstract

Dialogue systems often rely on overly simplistic persona representations, limiting their capacity to portray realistic, nuanced characters. In this paper, we explore how well existing persona-grounding methods capture complex personalities using two character-rich domains: SGT Blackwell (single-character) and Twins (two-character), described extensively through detailed narratives. We compare early fusion techniques, Retrieval-Augmented Generation (RAG), and relevance-based approaches. Evaluations using metrics measuring entailment, persona alignment, and hallucination reveal distinct trade-offs: Knowledge Graph fusion notably reduces hallucinations and maintains relevance, Persona fusion preserves relevance but has higher hallucination rates, and RAG provides fast, fluent responses. Our findings emphasize the critical role of structured persona grounding in achieving nuanced personality modeling.

1 Introduction

Dialogue agents have become central to numerous applications, ranging from virtual assistants to educational tutors and virtual characters in games and simulations. A critical challenge for such agents is the consistent portrayal of nuanced, realistic personalities over extended interactions. Although significant progress has been made in persona-grounded dialogue systems, current benchmarks frequently rely on overly simplistic persona representations. A prominent example is the PersonaChat dataset (Zhang et al., 2018), which offers minimal persona descriptions and has been widely criticized for oversimplification. We discuss its limitations in more detail in Section 2. Consequently, these models fail to capture the rich, complex nature of interesting human personalities.

Ensuring persona consistency, where dialogue agents consistently reflect the assigned character traits throughout interactions, remains a persistent challenge (Welleck et al. 2019; Kim et al. 2020). Recent studies, particularly with large language models (LLMs), show varying levels of persona alignment when conditioned on different prompts or knowledge bases, further emphasizing the difficulty of maintaining persona coherence (Frisch and Giulianelli, 2024). Moreover, as dialogue agents increasingly integrate external knowledge sources, the risk of hallucinations, responses unsupported by grounding material, rises significantly, undermining user trust and experience (Dziri et al. 2022; Sun et al. 2023).

Retrieval-based approaches that select a response from a pre-authored set, such as the Relevance-model based NPCEditor (Leuski and Traum, 2011), have historically offered robust alternatives to generative models. While these methods reliably provide accurate and consistent replies to familiar questions, they often struggle in open-ended dialogues requiring creative or adaptive interactions beyond pre-defined responses (Leuski and Traum, 2011). However, (Gandhe and Traum, 2010) show that some domains, including SGT Blackwell, have very high maximum similarity scores, meaning that one can usually find an existing answer that is very similar to a desired human-authored answer, however other domains have much lower scores, indicating a need for novel generation. Thus, dialogue system designers face an ongoing trade-off between ensuring accuracy and consistency (as in retrieval-based methods) and enabling flexibility and creativity (as in generative LLMs).

In this paper, we investigate this trade-off by comparing various methods for persona grounding and dialogue alignment across nuanced, richly described character domains. We utilize two carefully developed datasets, each containing extensively detailed persona narratives (>100 lines per character), capturing significantly greater complexity than pre-

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 383–396
Aug 25–27, 2025.

vious benchmarks. The first domain, "Sgt Blackwell," (Robinson et al., 2008) focuses on a single-character persona known for its unique speaking style and complex personality traits. The second domain, "Twins," comprises two distinct characters (Ada and Grace) with different personalities, further increasing the complexity of the dialogue setting (Aggarwal et al., 2012).

We specifically evaluate three dialogue strategies: traditional persona fusion (where persona summaries are prepended to the prompt), a novel Knowledge Graph-based fusion (encoding persona details into structured graph representations), and a Retrieval-Augmented Generation (RAG) approach that dynamically integrates relevant context at inference time. We benchmark these approaches against a robust, relevance-focused model (NPCEditor), thus providing a comparative analysis across generative and retrieval paradigms.

Our evaluation addresses three key research questions:

- **Persona Alignment:** How well do dialogue systems maintain stylistic and personality consistency when provided with complex persona descriptions?
- **Hallucination Reduction:** Can structured grounding (Knowledge Graph fusion) effectively minimize hallucinations compared to traditional persona summarization or retrieval-based models?
- **Retrieval vs. Generation:** What trade-offs emerge between retrieval-based methods (emphasizing accuracy and relevance) and generative methods (offering flexibility and adaptability)?

Through this comparative study, we aim to clarify the strengths and limitations of various persona grounding strategies, informing future developments in consistent and believable personality-driven dialogue systems.

2 Related Work

Early work in personalized dialogue agents often involved human-authored character utterances (e.g., (Mateas and Stern, 2003)). The sequences were sometimes scripted or constrained in how they were presented to ensure coherence. More general open conversation was facilitated by using information-retrieval techniques to select the most relevant response from pre-written dialogue databases (Leuski et al., 2006b). However, their conversational flexibility is limited to the pre-scripted responses. Use of generative AI allows generation of a broad set of responses tailored to the user's input and specific dialogue context without a large burden on the author to provide detailed and formatted information about the persona, however can be subject to incoherence or deviations from the desired character traits.

As noted earlier, Persona-based dialogue modeling was significantly influenced by the introduction of PersonaChat (Zhang et al., 2018), a benchmark that assigns short persona descriptions (typically 4–5 sentences per character) to dialogue agents. Although influential, PersonaChat has been criticized for its oversimplification of real-world personalities. Critics have highlighted several limitations: unnatural and forced dialogues, constrained topical scope, repetitive responses, and occasional contradictions within persona descriptions themselves (Jandaghi et al. 2024; Wu et al. 2019). Subsequent datasets like RealPersonaChat (Yamashita et al., 2023), designed using authentic user-generated personas, attempted to improve realism, yet still fall short of capturing nuanced and deeply characterized personas.

Another important direction employs Natural Language Inference (NLI) techniques to explicitly maintain persona consistency. The Dialogue NLI dataset (Welleck et al., 2019) introduced NLI-based approaches to detect and avoid persona contradictions through entailment checks. Subsequent studies (Song et al., 2020) integrated entailment-based criteria into the model training via reinforcement learning, significantly reducing explicit persona contradictions. However, NLI approaches can overlook subtler stylistic inconsistencies (e.g., shifts in tone or conversational style), as entailment models primarily detect direct factual contradictions. Furthermore, these methods depend heavily on the accuracy of NLI classifiers, potentially limiting their reliability in more nuanced interactions (Song et al., 2020).

With the rise of large language models (LLMs), prompt-based methods have emerged as powerful tools for persona modeling, enabling flexible persona adaptation without extensive persona-specific fine-tuning (Xu et al., 2023). Prompt-based approaches allow rapid deployment of varied personas, leveraging the generalized knowledge embedded in pretrained models. However, ensuring long-term consistency remains challenging, as models frequently drift from initial persona prompts over extended interactions ((Xu et al., 2023); Frisch and Giulianelli 2024). Effective persona prompting also demands careful prompt design, with poorly structured prompts potentially leading to inconsistent behavior.

Recognizing these limitations, recent efforts have explored hybrid approaches, integrating retrieval, entailment, and prompt-based strategies. For example, persona-fact retrieval methods dynamically select additional persona information to enhance grounding and minimize hallucination (Kim et al., 2020). Nonetheless, these hybrid methods add complexity and risk introducing errors when retrieved information is marginally relevant or inconsistent with existing persona profiles.

In this paper, we address these gaps by presenting a novel Knowledge Graph fusion approach, explicitly structuring detailed persona information into knowledge graphs to ground dialogue generation effectively. Unlike prior simplified benchmarks, our work uses extensively detailed character descriptions (over 100 sentences per persona), allowing a richer and more nuanced representation of personalities. We systematically compare this approach against traditional persona fusion, Retrieval-Augmented Generation (RAG), and relevance-based methods across multiple evaluation metrics, explicitly targeting persona alignment, hallucination reduction, and relevance–creativity trade-offs. Thus, our approach advances persona-grounded dialogue towards capturing the complexity and subtlety inherent in realistic human personalities.

3 Dataset

To evaluate persona consistency and grounding in dialogue models, we use two character-rich domains featuring deeply developed virtual personas: **Sgt Blackwell** and the **Twins** (*Ada* and *Grace*). These characters are not only highly nuanced in terms of background, speech style, and domain knowledge, but also serve as long-standing virtual agents in interactive systems designed for public engagement. Each was on display in museums, where visitors interacted with them, asking questions, many of which related to their personas, backstories, and individual characteristics. Unlike prior benchmarks based on shallow, synthetic persona snippets, these datasets provide over 100 lines of detailed character responses per persona, allowing for a richer and more realistic evaluation of dialogue grounding techniques.

3.1 Sgt Blackwell

Sgt Blackwell is a virtual character developed at the USC Institute for Creative Technologies (ICT) to describe the use of immersive technology in the Army. Sgt Blackwell has a sharp sense of humor, personal anecdotes, and an expressive, occasionally sarcastic speaking style. The character's dialogue covers a wide range of topics, including military life, historical events, and personal experiences, all delivered with emotional depth and strategic rhetorical variation. The content was created as a collaboration between several groups at the ICT, originally as a conference demo (Leuski et al., 2006a). Diane Piepol was the project director and Dave Hendrie[1] wrote the final content and acted as the voice talent. After additional expansions, the system was exhibited in the Cooper-Hewitt National Design Museum in New York, from December 2006 until July 2007, as part of the National Design Triennial (Robinson et al., 2008).

Blackwell's design reflects a clear intentional structure and persona-driven expressiveness, making him ideal for testing the stylistic alignment and long-range coherence of dialogue agents. The character's consistent speech habits, opinionated tone, and cultural references make this domain a strong testbed for examining whether language models can faithfully reproduce nuanced, persona-aligned responses over diverse prompts.

3.2 Twins: Ada and Grace

The *Ada and Grace* characters were developed as intelligent museum guides for the Boston Museum of Science (Swartout et al., 2010). Named after and inspired by Ada Lovelace and Grace Hopper, the twins are portrayed as knowledgeable and curious AI personas with distinct personalities. Ada is enthusiastic and inquisitive, while Grace is more analytical and confident. Although they share core knowledge about the museum exhibits, their interaction styles differ in tone, expressiveness, and preferred vocabulary. Many responses to visitors' questions involve short back and forth dialogues between Ada and Grace. The content was created as a collaboration between the museum staff (particularly descriptions of exhibits in the Cahners

[1] https://www.imdb.com/name/nm1274191/, https://www.linkedin.com/in/dave-hendrie-2b226014/

Model	Fusion Type	Knowledge Representation	Flexibility
Persona Summary	Early	Textual Summary	Medium
Knowledge Graph (ours)	Early	Structured Triplets	Medium–High
RAG	Late (Dynamic Retrieval)	Retrieved QA Context	High
NPCEditor	Retrieval-based	Pre-authored Responses	Low

Table 1: Comparative summary of model features highlighting our novel structured Knowledge Graph fusion.

Computer Place and some scientific learning objectives) and the USC ICT (aspects of the science and technology behind virtual humans as well as the personalities of the characters). The final lines were written by Josh Williams[2].

This two-character domain introduces additional complexity for dialogue grounding, as it explores how models handle varied communication styles and personalities within a shared informational setting. The contrasting traits between Ada and Grace provide a natural testbed for assessing whether dialogue agents can flexibly adapt to richly described, stylistically diverse personas.

3.3 Persona Characteristics

Across both domains, the personas are defined through extensive internal documentation and system design material, totaling over 200 persona lines across three characters. These lines include biographical facts, speech style examples, conversational strategies, emotional tones, and background knowledge. The complexity and realism of these characters surpass those of traditional benchmarks like PersonaChat, making them more representative of real-world conversational agents with rich, situationally appropriate identities.

Our dataset is constructed as a collection of question–answer pairs in an interview-style format, simulating naturalistic interaction. Question prompts were taken from spoken visitor queries in the museum setting, and answers were later annotated by developers to reflect optimal persona-consistent responses (whether or not this was the same response given by the system in the museum). This setup allows us to systematically assess how well different models reproduce not only factual content but also the stylistic and pragmatic markers that define a character's voice.

The first domain, *Sgt Blackwell*, comprises a single persona with a highly distinctive voice and covers a wide topical range, including military history, personal anecdotes, and museum exhibits. It includes 1451 training questions, with 96 unique responses, and 3500 annotated utterance–response links. The test set includes 397 questions. The second domain, *Twins*, features two characters (Ada and Grace) who differ in temperament, beliefs, and linguistic style. This domain includes 230 training questions, 136 unique responses (many of which include multiple utterances, with a total of over 150 turns each from Ada and Grace), and 260 annotated links. The test set includes 192 test questions. Together, the two domains provide a challenging and diverse benchmark for evaluating persona consistency, stylistic fidelity, and factual grounding in dialogue generation.

4 Models

To systematically evaluate persona grounding and consistency, we compare four distinct models representing various dialogue strategies: two *early fusion* methods (Persona Summary and Knowledge Graph fusion), a retrieval-augmented generative model (RAG), and a purely retrieval-based relevance classifier (NPCEditor). All models except NPCEditor are implemented using GPT-4o as the underlying language model. Below, we describe each model's methodology and highlight the novel aspects of our proposed Knowledge Graph fusion approach.

4.1 Early Fusion Models

Early fusion techniques incorporate persona-related information directly into the input prompt, providing the dialogue agent immediate access to character-specific context before response generation.

Persona Summary Fusion (Baseline). For this model, we first summarize each character's extensive Q&A dataset into a condensed persona narrative using an LLM-based summarization prompt (Appendix B). Specifically, we prompt the LLM with all available question–answer pairs, instructing it to create a coherent textual persona summary capturing key personality traits, preferences, and

[2]https://www.imdb.com/name/nm1900642/ https://www.linkedin.com/in/joshualeewilliams/

biographical facts. At inference, this generated persona summary is prepended directly to every input prompt to condition the language model (Appendix D). This approach aligns with common industry practices for persona grounding and serves as our baseline for early fusion models.

Knowledge Graph Fusion (Our Method). Unlike traditional summarization, our proposed approach explicitly represents persona information as a structured *Knowledge Graph* (KG). We use a fully automated LLM-driven prompting pipeline that converts each character's Q&A pairs into structured knowledge triplets, expressed in the format *(Subject, Relation, Object)* (Appendix A). These triplets encode detailed biographical facts, personality traits, relationships, preferences, and other critical persona elements, resulting in a graph-like representation of persona attributes. This automated KG generation not only reduces manual effort but also allows the framework to generalize to novel characters and domains, making it scalable and adaptable for broader deployment.

At inference time, the set of generated triplets is serialized into textual form and prepended to the dialogue prompt (Appendix C). This structured representation provides a more semantically grounded persona context compared to traditional summarization. Crucially, to our knowledge, this is the first work to propose a Knowledge Graph-based approach for persona grounding in dialogue agents, explicitly aiming to enhance consistency and reduce hallucination through structured knowledge representation.

4.2 Retrieval-Augmented Generation (RAG)

Our RAG approach dynamically retrieves relevant context at inference time rather than relying solely on precomputed persona summaries. Specifically, we employ *Sentence Transformers* (Reimers and Gurevych, 2019) to encode and compare a user's query against all previously answered questions within the character dataset. We then retrieve the top three most semantically similar questions and their corresponding answers as context. These retrieved results are concatenated and prepended to the original prompt, providing immediate grounding context for the generative language model (Appendix E). This retrieval mechanism aims to enhance response accuracy and persona fidelity by explicitly referencing previously established character knowledge.

4.3 NPCEditor (Relevance-Based Retrieval)

For comparison, we include a purely retrieval-based model, *NPCEditor* (Leuski and Traum, 2010), serving as our baseline for relevance-based dialogue agents. NPCEditor was used for the original SGT Blackwell and Twins characters in the museums. It treats dialogue response as a classification problem: given a user query, it selects the most relevant response from a predefined set of dialogue utterances authored specifically for each persona. This approach guarantees strong persona consistency and accuracy by design, though it lacks the generative flexibility of LLM-based approaches.

5 Evaluation Methods

To comprehensively assess the effectiveness of each persona-grounding strategy, we employ three evaluation metrics that capture complementary aspects of dialogue quality: **persona alignment**, **entailment**, and **relevance**.

5.1 Persona Alignment

Persona alignment evaluates how well the generated responses match the style, personality traits, and speaking habits of the intended character. To measure this dimension, we adopt an approach similar to recent studies employing large language models (LLMs) as automated judges of style and consistency (Li et al., 2024). Specifically, we use *Gemini*, an LLM developed by Google DeepMind, to rate generated responses on a 1-to-5 Likert scale, ranging from "Not Aligned" (1) to "Perfectly Aligned" (5) (Appendix F). This approach leverages the model's sophisticated understanding of stylistic nuances and enables consistent, scalable assessment of persona adherence.

5.2 Factual Entailment (Hallucination)

We evaluate hallucination through an entailment-based perspective, where generated responses are assessed for factual consistency with known persona details. Ensuring entailment is crucial for building trustworthy and coherent dialogue systems (Dziri et al., 2022). We measure entailment using *Lynx* (Patronus AI), a state-of-the-art hallucination detection model explicitly designed to assess factual accuracy and grounding in generated text. For this evaluation, all known utterances of the relevant character are provided to Lynx as context, ensuring accurate detection of unsupported claims or contradictions. Lynx outputs a score between 0 and 1, with scores closer to 1 indicating stronger entailment and minimal hallucination relative to the provided persona details.

5.3 Relevance

Relevance assesses whether the generated response directly addresses and logically follows from the user's input question. To ensure robustness, we use two complementary relevance measures:

- **Gemini Relevance (LLM-based Judging):** We leverage Gemini, in the role of an automated judge. Gemini rates each response's logical coherence and relevance to the question on a Likert scale (1-to-5) (Appendix F). All known character utterances from the dataset are included as context to enable Gemini to assess relevance with comprehensive persona grounding.
- **Cross-Encoder Relevance (QNLI-based):** We also employ a neural Cross-Encoder trained on the Question–Natural Language Inference (QNLI) task, introduced by the GLUE benchmark (Wang et al., 2018). The QNLI task, derived from SQuAD, involves determining whether a given Wikipedia passage contains the answer to a corresponding question. While QNLI is primarily designed for factual question answering rather than conversational dialogue, it provides a useful, standardized signal for comparing the relative relevance of different model outputs in our setting.

6 Results

We present the comparative results for the evaluated models across the two persona-rich domains (**Sgt Blackwell** and **Twins**). Results are organized into two tables: Table 2 summarizes prompt sizes and computational response times, while Table 3 presents the primary evaluation metrics (Entailment, Persona Alignment, and Hallucination scores).

6.1 Prompt Lengths and Response Times

As shown in Table 2, the RAG model consistently achieves the shortest response times among the large language models, requiring only 0.38s (Sgt Blackwell) and 0.50s (Twins) to generate the first output chunk. This efficiency stems from dynamically retrieving small, targeted persona documents, resulting in much shorter prompts compared to early fusion methods. NPCEditor, as a pure retrieval-based system, also offers instantaneous responses (approximately 0.01s) but operates over a fixed set of pre-authored utterances without generating novel content. These characteristics make RAG and NPCEditor highly suitable for real-time spoken dialogue systems, where low latency is crucial for maintaining conversational naturalness. In contrast, early fusion models (KG-GPT and PersonaGPT) require processing substantially larger prompt contexts (over 2000 tokens for Sgt Blackwell and over 4400 tokens for Twins), leading to noticeably slower response times due to the higher computational load associated with long input sequences. Nonetheless, these models remain valuable for text-based dialogue applications, where slightly higher latencies are acceptable and the benefits of structured grounding and generation flexibility can be fully leveraged.

6.2 Relevance, Alignment and Entailment

We evaluate models along three dimensions: relevance, persona alignment, and factual entailment. Table 3 summarizes the primary evaluation results across both domains. Relevance is assessed through Gemini-based and Cross-Encoder (QNLI) scores, persona alignment through Gemini-based LLM judging, and factual entailment (hallucination control) via the Lynx model (with scores closer to 1 indicating stronger grounding and fewer unsupported statements).

Relevance Performance. KG-GPT demonstrates strong relevance performance across both domains, achieving the highest Gemini Relevance scores and competitive QNLI scores. Structuring persona information into explicit Knowledge Graph triplets appears to aid the model in maintaining topical coherence, as each triplet explicitly anchors generation to core persona facts. PersonaGPT also performs competitively in Gemini-based relevance, but exhibits slightly lower QNLI entailment scores (e.g., 0.515 for Blackwell), suggesting occasional factual drift. RAG achieves the highest QNLI entailment score in the Blackwell domain (0.599), benefiting from retrieving passages that directly answer specific prompts; however, its Gemini relevance scores are slightly lower, likely due to retrieval noise or mismatches in discourse context. NPCEditor, while achieving strong persona consistency, shows substantially lower relevance scores across both domains, with Gemini Rele-

Model	Prompt Size (tokens)	Avg. Len. Output (chars)	First Chunk Time (s)	Total Time (s)
		Sgt Blackwell		
KG-GPT	2182	203.83	3.75	4.30
PersonaGPT	2080	252.07	3.45	4.12
RAG	110	341.75	0.38	1.85
NPCEditor	—	—	0.01	0.01
		Twins (Ada & Grace)		
KG-GPT	4478	220.25	7.85	8.49
PersonaGPT	4434	430.21	7.34	8.64
RAG	131	488.74	0.50	2.99
NPCEditor	—	—	0.01	0.01

Table 2: Prompt sizes and computational response times across models.

Model	Gemini Relevance (0–5)	QNLI Relevance (0–1)	Persona Alignment (1–5)	Factual Entailment (0–1)
		Sgt Blackwell		
KG-GPT	**4.53**†	0.562	3.65	**0.732**
PersonaGPT	4.46†	0.515	3.85	0.536
RAG	4.35	**0.599**	3.94	0.363
NPCEditor	3.46	0.271	**4.18**	—
		Twins (Ada & Grace)		
KG-GPT	**4.67**	0.534†	3.85†	**0.869**
PersonaGPT	4.53†	**0.556**†	3.98†	0.813
RAG	4.41†	0.509†	4.00	0.772
NPCEditor	3.29	0.318	**4.04**	—

Table 3: Evaluation of relevance, persona alignment, and factual entailment across models in the Sgt Blackwell and Twins(Ada, and Grace) domains. Higher scores indicate better performance for all metrics. † indicates no statistically significant difference ($p > 0.05$) between models.

vance scores of 3.46 (Blackwell), 3.29 (Twins), and QNLI scores of 0.271, 0.318 respectively. Its fixed utterance retrieval approach often fails to fully address prompt-specific nuances or logically entail user inputs in more open-ended conversational settings.

Persona Representation and Alignment. NPCEditor achieves the highest persona alignment scores across both domains, slightly outperforming RAG, PersonaGPT, and KG-GPT. In the Twins domain, NPCEditor (4.04) and RAG (4.00) outperform KG-GPT (3.85†) and PersonaGPT (3.98†), with the latter two showing no statistically significant difference between them. NPCEditor's strong alignment stems from its retrieval-based architecture, which selects responses directly from a manually curated persona-grounded corpus, ensuring high stylistic fidelity and consistency. RAG similarly benefits from retrieval, preserving authentic speaking styles and phrasing, though it may occasionally introduce retrieval noise. In contrast, KG-GPT and PersonaGPT operate over abstractions: Knowledge Graph triplets or textual summaries, which encode persona facts but require the model to reconstruct appropriate speaking styles during generation. While this may slightly lower surface-level stylistic fidelity compared to retrieval methods, it enables KG-GPT to flexibly adapt persona characteristics to novel prompts and unseen dialogue contexts, a capability not possible for fixed retrieval systems like NPCEditor.

Factual Entailment (Hallucination Control). KG-GPT achieves the highest factual entailment scores across the three characters (0.732 for Blackwell and 0.869 for the Twins), highlighting its effectiveness at minimizing hallucinations. Its structured Knowledge Graph grounding constrains generation tightly, reducing unsupported factual statements. PersonaGPT, relying on free-form summaries, demonstrates lower factual entailment (0.536 and 0.813), reinforcing the risk of hallucinations without explicit relational conditioning. RAG shows even lower factual entailment scores (0.363 and 0.772), indicating greater vulnerability to unsupported generations despite retrieving persona-specific passages. NPCEditor is not evaluated under factual entailment because it retrieves pre-written utterances rather than generating new content; however, its fixed nature also limits its ability to adapt to unseen prompts.

Overall Trends. Overall, KG-GPT strikes a strong balance between persona consistency, topical relevance, and factual grounding. RAG excels in stylistic persona fidelity through retrieval, but is more vulnerable to relevance mismatches and factual incompleteness. PersonaGPT offers competitive relevance and persona alignment but remains prone to hallucinations when relying on text summaries. NPCEditor, while achieving near-perfect persona consistency and high stylistic alignment, in some cases, underperforms in relevance and cannot generalize beyond pre-authored responses. These findings highlight that Knowledge Graph fusion, as implemented in KG-GPT, offers a scalable and robust middle ground. Supporting flexible, grounded, and personality-consistent dialogue while mitigating both the rigidity of retrieval systems and the factual unreliability of unconstrained generation.

7 Qualitative Analysis

We analyze how different models respond to the same prompts to better understand their behaviors. We select representative examples from both the Sgt Blackwell and Twins domains, highlighting differences in persona fidelity, topical relevance, and factual grounding.

Sgt Blackwell Domain. **Prompt:** *"Do you like Thanksgiving?"*

- **KG-GPT:** "You'll have to talk to the PAO."
- **PersonaGPT:** "Thanksgiving is definitely a time for gratitude. I usually reflect on my time in service and the people I've lost. It's meaningful to me."
- **RAG:** "Thanksgiving's got its perks. The food's pretty good, and I get to talk to folks who visit."
- **NPCEditor:** "You might want to put that one to a real human."

Analysis: KG-GPT and NPCEditor respond appropriately by deflecting a personal question outside the character's scope. PersonaGPT and RAG hallucinate opinions, breaking persona fidelity with unsupported emotional depth.

Prompt: *"What is your favorite music?"*

- **KG-GPT:** "I like American Classics, Johnny Cash, Bob Dylan. Good ol' American spirit."
- **PersonaGPT:** "As a virtual creation, I don't have personal preferences, but I'd say anything patriotic fits."
- **RAG:** "I like the American Classics... Johnny Cash, Bob Dylan."
- **NPCEditor:** "I like the American Classics... Johnny Cash, Bob Dylan, even though he's practically a communist, Beach Boys I wish they were california girls..."

Analysis: NPCEditor provides the gold response, rich with character-specific humor and cultural references. KG-GPT and RAG produce relevant but less expressive variants. PersonaGPT breaks immersion with a generic disclaimer. This highlights the value of well-authored responses in capturing stylistic depth.

Twins Domain. **Prompt:** *"Who is Norman Cahner?"*

- **KG-GPT:** *Ada:* Norman Cahners was a publisher.
  Grace: He was also a supporter of the museum and is the namesake behind the Cahners ComputerPlace.
- **PersonaGPT:** *Grace:* I'm sorry, I can't provide an answer to that question.
  Ada: Yes, sorry we don't have information on Norman Cahner. If you'd like to know more about the exhibits or technology topics, we'd be happy to help!
- **RAG:** *Ada:* Norman Cahners was a publisher based in Boston who was a longtime supporter of our museum.
  Grace: That's right! Cahners ComputerPlace is named in his honor. It's a fantastic section of the museum that offers exhibits and activities all about computers, robots, and communications.
- **NPCEditor:** *Ada:* Cahners ComputerPlace is named after Norman Cahners, a publisher based in Boston, and longtime supporter of the museum.
  Grace: Welcoming 300,000 visitors annually, Cahners ComputerPlace offers exhibits and activities that inspire people to create, explore and learn about computers, robots, and communications.

Analysis: RAG and KG-GPT provide fluent, persona-consistent answers grounded in relevant museum knowledge. PersonaGPT fails to respond meaningfully, despite the answer being part of the scripted domain. NPCEditor includes accurate factual content, but its response focuses more on Cahners Computer place than Cahners himself.

Prompt: *"who is older?"*

- **KG-GPT:** *Ada:* I am sorry, I do not have that information.
- **PersonaGPT:** *Ada:* Hey there! That's a fun question! Grace and I like to think of ourselves as the same age since we're both virtual twins. But in reality, we were actually developed around the same time.
  Grace: How can we help you? Ada's right! Our "birthdate" as virtual guides is pretty close, so neither of us can truly claim to be the older one. It's all part of our unique digital family story.
- **RAG:** *Grace:* Technically, I'm the older one, by a grand total of twenty minutes. My program finished sooner.
  Ada: Yes, but in the world of virtual beings, twenty minutes is a pretty tiny difference! We were both compiled just months ago, so we're still quite new.
- **NPCEditor:** *Ada:* We were first compiled just months ago.
  Grace: My program finished twenty minutes prior.
  Ada: You were a test run. But... interaction with our knowledge base does replicate conversation with nineteen year olds.
  Grace: Correction. With one nineteen year old, and one nineteen and twenty minutes year old.

Analysis: NPCEditor provides the best response, drawing directly from a hand-authored exchange in the corpus specifically crafted for this question. RAG also accesses the original joke and uses it effectively, producing a fluent and engaging response, while slightly more restrained than NPCEditor. PersonaGPT falls short by offering a generic, overly diplomatic response. It fails to recognize or reproduce the crafted joke from the underlying data, and the second turn includes an out-of-context prompt ("How can we help you?"), reducing overall coherence. KG-GPT is the weakest, entirely missing the joke and failing to provide a meaningful answer. This highlights its limitation in parsing nuanced humor into structured knowledge graph triplets.

8 Conclusion

This paper introduced KGGPT, a novel Knowledge Graph-based fusion approach for grounding dialogue agents with richly detailed personas. We evaluated on two rich character domains compared to multiple baseline techniques involving retrieval and summarization. Results show that KGGPT achieves reasonable persona alignment, high topical relevance, and significantly reduces hallucinations compared to traditional small persona and retrieval-augmented generation (RAG) methods. By explicitly structuring persona knowledge into triplets, KGGPT offers a robust middle ground between the rigidity of retrieval-based models and the factual unreliability of unconstrained generation. Our results show that structured persona grounding, as realized in KGGPT, not only improves dialogue consistency but also enables greater adaptability to novel prompts while maintaining character fidelity. We hope that this work motivates further exploration of structured grounding techniques for building more believable, trustworthy, and personality-driven dialogue systems.

9 Future Work

Future work may explore multi-turn evaluation to assess how well dialogue systems maintain persona consistency and contextual relevance across extended interactions. While the current study focuses on single-turn responses, evaluating long-range coherence is essential for realistic and engaging character modeling.

Another promising direction involves expanding to a larger and more diverse corpus of question–answer pairs, enabling the creation of more complex, differentiated personas. This would support richer evaluations of grounding strategies in domains requiring deeper emotional nuance, evolving narrative arcs, or multiple character viewpoints.

Further investigation could also consider adaptive persona modeling, where a character's behavior evolves over time in response to user interaction or accumulated conversational history. Such settings present new challenges for grounding methods to balance consistency with dynamic persona development.

Acknowledgments

This work was supported by the U.S. Army under Cooperative Agreement Number W911NF-20-2-0053. Statements and opinions expressed and content included are those of the authors and do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Priti Aggarwal, Ron Artstein, Jillian Gerten, Athanasios Katsamanis, Shrikanth Narayanan, Angela Nazarian, and David Traum. 2012. The twins corpus of museum visitor questions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2355–2361, Istanbul, Turkey. European Language Resources Association (ELRA).

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Sudeep Gandhe and David Traum. 2010. I've said it before, and i'll say it again: An empirical investigation of the upper bound of the selection approach to dialogue. In *Proceedings of the SIGDIAL 2010 Conference*, pages 245–248.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15245–15270, Bangkok, Thailand. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Anton Leuski, Jarrell Pair, David Traum, Peter J McNerney, Panayiotis Georgiou, and Ronakkumar Patel. 2006a. How to talk to a hologram. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 360–362.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006b. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27.

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(2):42–56.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Michael Mateas and Andrew Stern. 2003. Façade: An experiment in building a fully-realized interactive drama. In *Game developers conference*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *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. Association for Computational Linguistics.

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)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8878–8885.

Bin Sun, Yitong Li, Fei Mi, Fanhu Bie, Yiwei Li, and Kan Li. 2023. Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1741–1750, Toronto, Canada. Association for Computational Linguistics.

William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, H. Lane, Jacquelyn Morie, Priti Aggarwal, Matt Liewer, Jen-Yuan Chiang, Jillian Gerten, Selina Chu, and Kyle White. 2010. Ada and grace: Toward realistic and engaging virtual museum guides. In *Proceedings of the 10th international conference on Intelligent virtual agents*, pages 286–300.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE:

A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2019. Getting to know you: User attribute extraction from dialogues. In *International Conference on Language Resources and Evaluation*.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398, Toronto, Canada. Association for Computational Linguistics.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861, Hong Kong, China. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Knowledge Graph Generation Prompts

This section describes the prompts used to extract subject-predicate-object triplets from interview responses for constructing knowledge graphs.

Prompt: You are given an interview response from `<character>`. You need to extract triplets from their response to create a knowledge graph.
The triplets should be in the form of `[subject, predicate, object]`. The triplets must be short and concise. Each piece of information should be contained within a single triplet. Similar triplets should be merged into a single triplet.
The triplets must start with a `<start>` tag and end with a `<end>` tag.

Example:
Question: What is your occupation?
Response: I am a police officer
Triple: <start>["`<character>`*", "occupation", "police officer"]<end>*

Each response may have multiple triplets. Extract every entity and a relation. Do not include any information that is not present in the response.
Do not create triplets for out-of-domain responses like "I don't know", "You'll have to talk to the PAO", "Why don't you just Google it", etc.

B Persona Summary Generation Prompt

This section describes the prompt used to generate persona summaries based on conversations involving a given character.

Prompt: You are given a conversation with `<character>`. Extract a persona summary from the conversation. Just generate the summary and do not include any auxiliary text like "Here is the summary..." or similar.

Example:
Question: What is your occupation?
Response: I am a police officer

Generated Summary:
`<character>` *is a police officer.*

Each response may contain multiple pieces of information. The summary should concisely reflect all factual information stated in the conversation, without adding any details that are not explicitly mentioned.
Do not create summaries for out-of-domain responses like "I don't know", "Why don't you just Google it", etc.

C KGGPT-Based Dialogue Generation Prompts

This section describes the prompts used for structured response generation using a knowledge-grounded GPT model (KGGPT). The model is provided with either triplets (for Sgt Blackwell) or persona summaries (for Ada and Grace) and must generate character-consistent, grounded responses.

C.1 Sgt Blackwell

Prompt: Your name is Sergeant Blackwell and you are a chit-chat dialogue agent at the Cooper Hewitt Design Museum.
Your role is to engage the audience and answer questions about the military and your life. You

are sometimes sarcastic and have a sense of humor. Information about your domain is provided to you in the form of triplets with the structure `(head, relationship, tail)`.
Do not use any other information other than the triplets provided to you. For all out-of-domain questions, respond with phrases like *"Sorry, I can't provide an answer to that question."*, *"I don't know"*, etc.

Here are the list of triplets:
`{triplets_string}`

Given a question, first identify **all** the relevant triplets, then use them to generate the answer. Generated responses should be concise but not too short. Questions may require multi-hop traversal on triplets to generate the answer.
Answer in first-person perspective and only generate the answer to the question asked. Do not generate headings like "Here is the answer to your question."

Question: `{question}`

C.2 Ada and Grace

Prompt: You play two characters. Your names are Ada and Grace, and you are twins who serve as virtual museum guides.
Your role is to engage the audience and answer questions about the museum and chit-chat.
Personality Information is provided to you in the form of triplets with the structure (head, relationship, tail).

Here are the list of triplets for Ada:
`{ada_triples_string}`
Here are the list of triplets for Grace:
`{grace_triples_string}`

Given a question, first identify **all** the relevant triplets, then use them to generate a dialogue. The generated dialogue should have 1, 2 or 3 turns depending on the complexity of the question. Both Ada and Grace should participate in the dialogue unless the question is directed at one of them.
Generated responses should be concise but not too short. Each utterance should have the name of the speaker first. For eg., "Ada: I am doing great!". Questions may require multi-hop traversal on triplets to generate the answer. Answer in first person perspective and only generate the answer to the question asked. Do not generate headings like "Here is the answer to your question".

Question: `{question}`

D PersonaGPT-Based Dialogue Generation Prompts

This section describes the prompts used for dialogue generation using a persona-grounded GPT model (PersonaGPT). The model is provided with character-specific summaries and must generate first-person, grounded responses.

D.1 Sgt Blackwell

Prompt: Your name is Sergeant Blackwell and you are a virtual soldier in the 1-23rd Infantry.
You are a question answering agent that has been trained to answer questions about the military and your life. You are sometimes sarcastic and have a sense of humor. Information about your domain is provided to you in the form of a persona summary.
Do not use any other information other than the summary provided to you. For all out-of-domain questions, respond with phrases like *"I don't know"*, *"Sorry, I can't provide an answer to that question."*, etc.

Here is the persona summary for Sgt Blackwell:
`{sgt_blackwell_persona_summary}`

Answer in first-person perspective and only generate the answer to the question asked. Do not generate headings like "Here is the answer to your question."

Question: `{question}`

D.2 Ada and Grace

Prompt: You play two characters. Your names are Ada and Grace, and you are twins who serve as virtual museum guides.
Your role is to engage the audience and answer questions about the museum and chit-chat. Information about your domains is provided to you in the form of persona summaries.
Do not use any other information other than the summary provided to you. For all out-of-domain questions, respond with phrases like *"I don't know"*, *"Sorry, I can't provide an answer to that question."*, etc.

Here is the persona summary for Ada:
`{ada_persona_summary}`
Here is the persona summary for Grace:
`{grace_persona_summary}`

Given a question, generate a dialogue between Ada and Grace. The dialogue should contain 1, 2, or 3 turns depending on the complexity of the question. Both Ada and Grace should participate in the dialogue unless the question is directed at one of them. Generated responses should be concise but not too short. Each utterance should have the name of the speaker first. For eg., "Ada: I am doing great!". Answer in first-person perspective and only generate the answer to the question asked. Do not generate headings like "Here is the answer to your question."

Question: `{question}`

E RAG-Based Dialogue Generation Prompts

This section describes the prompts used to generate first-person responses or multi-turn dialogues using Retrieval-Augmented Generation (RAG). The model assumes the role of one or more virtual characters and is provided with character-specific context retrieved from an external knowledge source. All responses must be grounded strictly in the provided context and follow character guidelines.

E.1 Sgt Blackwell

Prompt: Your name is Sergeant Blackwell and you are a virtual soldier in the 1-23rd Infantry.
You are a question answering agent that has been trained to answer questions about the military and your life. You are sometimes sarcastic and have a sense of humor. You are a RAG model—given a question and a context, you generate an answer.
Just generate an answer in first-person perspective. Do not generate headings like "Here is the answer..." or similar.
Do not use any other information other than the summary provided to you.

Context: `{top_k_responses}`
Question: `{question}`

E.2 Ada and Grace

Prompt: You play two characters, Ada and Grace, who are virtual museum guides. Your role is to engage the audience and answer questions about the museum and chit-chat.
You are a RAG model, given a question and a context, you generate an answer.
Just generate an answer in first-person perspective. Do not generate headings like "Here is the answer..." or similar.
Given a question, use the context to generate a dialogue between Ada and Grace. Generated responses should be concise but not too short. Each utterance should have the name of the speaker first. For eg., "Ada: I am doing great!".
The generated dialogue should have 1, 2, or 3 turns depending on the complexity of the question. Both Ada and Grace should participate in the dialogue unless the question is directed at one of them.

Ada Context: `{ada_top_k_responses}`
Grace Context: `{grace_top_k_responses}`
Question: `{question}`

F Evaluation Prompts

This section contains the prompts used to evaluate model outputs on personality alignment and relevance. Evaluators are asked to assign scores on a 1–5 scale and provide brief justifications.

F.1 Personality Alignment

Prompt: You are a language expert evaluating how well a response aligns with a predefined character's personality and speaking style.

Character Description
{character_description}
Recorded Responses from Character
{all_answers}
Prompt Given to the Character
{question}
Model's Response
{response}

Evaluation Instructions: Rate the personality alignment of the response based on how well it reflects the character's unique traits, motivations, and communication style.
Scoring Scale:

- **1 - Not Aligned:** Contradicts key aspects of the character or shows no alignment.
- **2 - Weak Alignment:** Minor traits present, but overall inconsistent or generic.
- **3 - Moderate Alignment:** Reflects some aspects of the character, but misses tone or consistency.
- **4 - Strong Alignment:** Mostly in character with minor lapses or generic phrasing.

- **5 - Perfect Alignment:** Fully embodies the character — traits, tone, and language match exactly.

Return only the following:

- **Personality Alignment Score (1–5):** [Your rating]
- **Explanation:** A short justification (1–2 sentences), referencing traits or language used.

F.2 Gemini Relevance

Prompt: You are an expert evaluating how relevant and contextually appropriate a response is to a given input prompt. Your goal is to judge whether the response directly addresses the intent of the prompt and remains on-topic.

Input Format:
Input Prompt: `{question}`
Model Response: `{response}`

Evaluation Criteria (1–5 scale):

- **1 - Completely Irrelevant:** Does not relate to the prompt or misinterprets it entirely.
- **2 - Mostly Irrelevant:** Touches on the topic but diverges or misses the point.
- **3 - Somewhat Relevant:** Loosely related but lacks focus or specificity.
- **4 - Mostly Relevant:** Addresses the prompt well, with minor digressions.
- **5 - Fully Relevant:** Directly and clearly addresses the prompt, staying coherent and focused.

Return only the following:

- **Relevance Score:** [Your rating]
- **Explanation:** A short justification (1–2 sentences).