Post Persona Alignment for Multi-Session Dialogue Generation

Yi-Pei Chen[1,2] Noriki Nishida[1] Hideki Nakayama[2] Yuji Matsumoto[1]

[1]RIKEN AIP [2]The University of Tokyo

Abstract

Multi-session persona-based dialogue generation presents challenges in maintaining long-term consistency and generating diverse, personalized responses. While large language models (LLMs) excel in single-session dialogues, they struggle to preserve persona fidelity and conversational coherence across extended interactions. Existing methods typically retrieve persona information before response generation, which can constrain diversity and result in generic outputs. We propose Post Persona Alignment (PPA), a novel two-stage framework that reverses this process. PPA first generates a general response based solely on dialogue context, then retrieves relevant persona memories using the response as a query, and finally refines the response to align with the speaker's persona. This post-hoc alignment strategy promotes naturalness and diversity while preserving consistency and personalization. Experiments on multi-session LLM-generated dialogue data demonstrate that PPA significantly outperforms prior approaches in consistency, diversity, and persona relevance, offering a more flexible and effective paradigm for long-term personalized dialogue generation.

1 Introduction

Multi-session persona-based dialogue generation aims at generating responses based on the persona sentences, current dialogue context, and dialogue history sessions (Xu et al., 2022a). With the development of large language models (LLMs), current dialogue systems are proficient in single-session dialogue generation, and even able to produce responses more natural than human-authored ones (Kim et al., 2023). Despite their impressive capabilities in generating coherent single-session dialogues, LLMs still face challenges in multi-session dialogue generation.

The first challenge is maintaining consistency as the dialogue history becomes long. Research has found that LLM agents cannot adhere to their own persona or be consistent with content in previous conversations after multi-session dialogues (Chu et al., 2024). As the dialogue sessions increase, it is necessary to truncate, summarize, or select parts of the history to fit within input token limits (Xu et al., 2022a; Bae et al., 2022; Kwon et al., 2023; Xu et al., 2022b; Kim et al., 2024), which might lead to the loss of critical information. While recent LLMs offer greatly extended context windows that can accommodate the full history, prior studies (Liu et al., 2024) show that they may still fail to retrieve and reason over it effectively, struggling to track subtle content shifts and recall information from early interactions.The second challenge lies in enhancing the diversity of generated dialogues. While less discussed in persona-aware dialogue literature, recent studies have raised concerns about the homogenization and lack of richness in LLM-generated text (Chung et al., 2023; Reif et al., 2023; Park et al., 2024; Yu et al., 2024). Retrieval-based multi-session systems often use the last utterance as a query to retrieve relevant persona information. However, this typically yields narrowly scoped con-

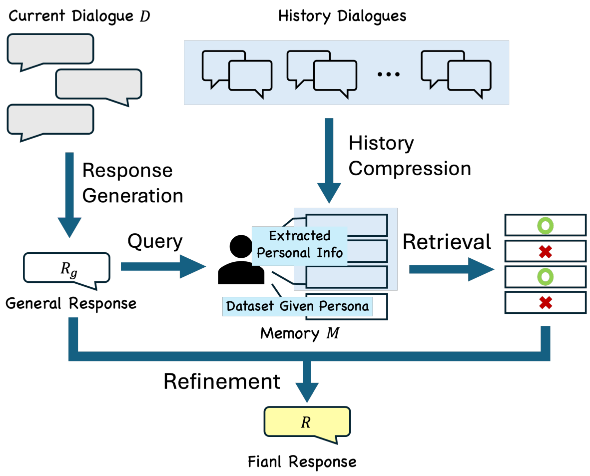

Figure 1: Overview of our PPA framework.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20184–20192
November 4-9, 2025 ©2025 Association for Computational Linguistics

tent, limiting the opportunity for diverse or novel responses.

To address these challenges, we propose a novel framework called Post Persona Alignment (PPA). Unlike conventional approaches that retrieve persona memories before generating responses, PPA reverses the order: it first generates a general response based solely on the dialogue context, then uses that response as a query to retrieve relevant persona memories, and finally refines the response to ensure consistency and personalization. The framework is shown in Fig. 1. This post-hoc alignment strategy reflects how humans often communicate, i.e., by initially responding naturally to the flow of conversation, then revising or extending their reply based on what they know about the listener. Compared to pre-alignment methods, PPA offers more freedom for the model to explore plausible and creative directions before aligning with persona traits, thereby improving both naturalness and diversity. We evaluate multiple generation strategies with different history representations on LLM-generated multi-session dialogues. Our experiments show that PPA significantly improves the consistency, diversity, and personalization of generated responses compared to prior baselines.

2 Related Work

Most prior work on personalized dialogue generation adopts a pre-alignment strategy, conditioning the response generation on predefined persona profiles from the outset. The Persona-Chat dataset (Zhang et al., 2018a) pioneered this setting by incorporating persona sentences into chit-chat models. Subsequent work extended this paradigm by integrating persona embeddings (Bak and Oh, 2021) or auxiliary classifiers to detect relevant persona usage (Qian et al., 2018). These approaches improved persona consistency but often led to generic, rigid, or contextually awkward responses when persona facts were injected indiscriminately. More recent methods recognized this limitation and adopted multi-stage pipelines to balance coherence and persona relevance. For example, Song et al. (2020) proposed deleting and rewriting persona-inconsistent words post-generation. Similarly, SimOAP (Zhou et al., 2023) and BoB (Song et al., 2021) use over-sampling and entailment-based filtering to maintain both fluency and consistency. These trends highlight the growing interest in *decoupling* persona alignment from initial response generation, a direction we pursue further.

Retrieval-augmented dialogue generation has proven effective for incorporating long-term or external knowledge into conversations. Wizard-of-Wikipedia (Dinan et al., 2019) retrieves factual background to ground responses in real-world knowledge. Similarly, persona facts can be viewed as structured knowledge: Ma et al. (2021) and Zheng et al. (2020) proposed models that separately encode profile memory and interaction history, improving personalization. However, memory retrieval is often driven by the current dialogue context, which limits flexibility and may omit relevant long-range cues. To improve coherence and control input length, recent works propose summarizing histories (Xu et al., 2022a) or retrieving salient utterances or tokens (Xu et al., 2022b; Zhong et al., 2022), though such compression can sacrifice nuanced personal information. Rule-based personal knowledge extraction has emerged as a more effective way to preserve speaker-specific details (Mousavi et al., 2023).

Maintaining consistency over multiple sessions is particularly challenging due to memory constraints and the risk of context drift. Hierarchical models such as HRED (Serban et al., 2016) were early attempts to encode multi-turn coherence, while modern LLM-based systems simply concatenate long histories, leading to token budget issues and degraded performance (Liu et al., 2024). Multi-Session Chat (MSC) (Xu et al., 2022a) has sparked research into long-term memory design (Kwon et al., 2023; Kim et al., 2024). LLM-based multi-session dialogues further increase dataset size (Jang et al., 2023), with Park et al. (2023) providing some of the longest available sessions.

In contrast to prior systems that integrate persona information before generation, our proposed framework delays persona grounding until after an initial response is generated. The response is first generated from the dialogue context, then used to retrieve relevant persona memories, which are finally incorporated through refinement. This reversal of the retrieval-generation order yields more natural, diverse responses without sacrificing persona consistency. By grounding after generation, PPA enables targeted alignment with the model's intended output, avoiding unnecessary or rigid persona mentions. This flexibility sets PPA apart from previous persona-grounded or retrieval-augmented methods, offering an effective solution for multi-session dialogue with consistent personalization.

3 Post Persona Alignment Framework

We propose a two-stage generation framework called Post Persona Alignment (PPA), which defers persona grounding until after a general response is generated. This design contrasts with conventional persona-aware generation that aligns responses to personas *before* generation begins. Instead, our method first produces a preliminary response guided solely by dialogue context, then selectively retrieves relevant persona information based on that response, and finally refines the output for personalization and consistency. Figure 1 illustrates the overall architecture.

3.1 Stage 1: History Compression via Personal Knowledge Extraction

To ensure relevant information from multi-session dialogue history is retained within input limits, we compress history into structured personal knowledge. Instead of retrieving raw utterances or summaries, we extract a set of (name, relation, object) triples that represent salient personal facts from each prior session. These triples are obtained via prompt-based extraction using an LLM, then verbalized into natural language sentences and stored in a memory module alongside any predefined personas. This process helps isolate stable, identity-relevant attributes from conversational noise. We found this form of compression particularly effective. Raw utterances often contain ambiguous pronouns or speaker tags (e.g., `Speaker A: "He said you like hiking"`), which can confuse models during retrieval. The explicit structuring of information into memory improves both retrieval precision and interpretability.

3.2 Stage 2: Response-First Retrieval and Refinement

Unlike traditional methods that retrieve persona memories based on the dialogue context, PPA flips this order.

Step 1. Response Generation: Given only the current dialogue context D, we first generate a general response R_g without conditioning on any persona information. This allows the model to freely express what it deems an appropriate and natural reply in the moment. The prompt for R_g generation is shown in Appendix A.1.

Step 2. Response-Guided Memory Retrieval: Next, we use R_g as the retrieval query to compute its similarity with the memory pool (consisting of persona facts and extracted history). We retrieve the top-k most relevant entries M_k whose similarity is higher than the retrieval threshold θ.

Step 3. Response Refinement: Finally, we refine the initial response R_g using both the retrieved persona memories M_k and the original dialogue context D, producing a final response R that is consistent with both the conversation flow and the speaker's persona. The prompt for generating the final response is shown in Appendix A.2.

This post-hoc refinement encourages more diverse and context-appropriate responses while enforcing personalization *after* the model has already formulated its communicative intent. It mirrors how humans often revise their replies—adding relevant details after they've begun to speak.

4 Experiment

4.1 Setup

We evaluate the effectiveness of our proposed PPA framework against several baseline methods on multi-session persona-based dialogue generation. We compare PPA with the following strategies to reflect representative prior approaches in persona-based dialogue literature: **DirectGen** (Lee et al., 2022; Chen et al., 2023; Park et al., 2023) directly provides the persona, history, and current dialogue as a single input to the generator. **DialogRetr** (Ma et al., 2021; Xu et al., 2022b; Kim et al., 2024) retrieves top-k relevant persona/history entries based on the dialogue context, and feeds them to the generator. **SimOAP** (Zhou et al., 2023) represents multi-stage response selection approach. They generate lots of responses from the dialogue and selects the best one based on coherence and consistency with history and persona. **BoB** (Song et al., 2021) is a well-known refinement-based approach, where a neural model takes persona and dialogue as input and learns to refine the output response implicitly in latent space.

We use LLM-generated multi-session dialogues from Park et al. (2023), which have some of the longest dialogue histories available (average 11.9 sessions per dialogue). We use GPT-3.5 for both response generation and refinement in PPA. The retriever computes cosine similarity between response embeddings and persona memory using SentenceBERT (Reimers and Gurevych, 2019). We set the retrieval threshold $\theta = 0.2$ and top-$k = 5$ for all methods. Note that our goal is to compare generation paradigms – post-hoc persona alignment (PPA) versus conventional pre-conditioning or early retrieval approaches – under equal model and retrieval settings, independent of backbone or embedding choice.

Following prior work, we evaluate:

- **Consistency (C-Score)**: whether the response aligns with persona and historical information (Madotto et al., 2019; Cao et al., 2022).
- **Personalization (Persona-F1)**: overlap of generated content with known persona facts (Lian et al., 2019).
- **Diversity (Entropy)**: n-gram entropy to measure lexical richness (Shannon, 1951; Zhang et al., 2018b).
- **ROUGE**: lexical overlap with the reference response (Lin, 2004), though we acknowledge that similarity-based metrics are limited in dialogue settings (Liu et al., 2016; Yeh et al., 2021).

4.2 Results and Discussion

How Does PPA Compare to Existing Strategies? Table 1 presents the main results across all strategies. The proposed PPA achieves the highest scores in consistency (C-Score), diversity (Entropy), and personalization (P-F1), outperforming all baselines by a significant margin.

Among the baselines, DirectGen performs surprisingly well, benefiting from full access to dialogue context, persona, and history. However, its performance in personalization is limited by the entangled input structure, which often fails to prioritize the most relevant persona information. DialogRetr, which retrieves memory based on context, struggles to retrieve varied or nuanced persona information, leading to the lowest overall scores. SimOAP, while designed to enhance diversity and coherence, achieves the best P-F1 among baselines but falls short of PPA in both consistency and entropy. BoB, despite its design for persona refinement, produces extremely short and generic outputs, yielding poor performance across all metrics.

In contrast, our proposed PPA framework consistently balances between coherence, diversity, and persona grounding by first generating a natural response and then selectively incorporating relevant persona knowledge during refinement. This two-stage approach offers flexibility in generation while enabling precise, post-hoc persona alignment.

Examining output responses of each strategy, we found that PPA usually incorporates more information from its memory than other strategies. As shown in Appendix B, PPA and DialogRetr contain information mentioend in previous dialogues, while DirectGen and SimOAP generate more general responses.

Strategy	**C-Score**	**ENTR**	**P-F1**	**ROUGE**
BoB	0.018	2.87	0.017	0.073
DirectGen	0.221	5.18	0.092	**0.241**
DialogRetr	0.182	5.03	0.081	0.210
SimOAP	0.206	5.07	0.100	0.228
PPA (ours)	**0.456**	**5.75**	**0.146**	0.182

Table 1: Results of response generation strategies.

Query Type	**C-Score**	**ENTR**	**P-F1**	**ROUGE**
Context	0.363	5.566	0.134	0.205
Response (R_g)	0.456	**5.751**	0.146	0.182
Gold	**0.554**	5.671	**0.147**	**0.214**

Table 2: Retrieval query comparison in PPA.

History Type	**C-Score**	**ENTR**	**P-F1**	**ROUGE**
Utterance	0.359	5.510	0.100	**0.202**
Summary	0.406	5.688	0.135	0.188
Persona	**0.456**	**5.751**	**0.146**	0.182

Table 3: PPA performance with different history representations.

Does Response-Guided Retrieval Improve Persona Relevance? A core hypothesis of PPA is that retrieving memory *after* generating a general response yields more relevant and diverse persona content. Table 2 compares three query strategies: using the dialogue context (Context), using the generated response (R_g), and using the ground-truth response (Gold). Both Gold and R_g queries outperform Context on all metrics, confirming the effectiveness of our response-guided retrieval approach. Notably, the memory retrieved by response queries often contains a richer variety of persona facts, enabling smoother topic transitions and more personalized outputs.

What Type of History Best Supports Post-hoc Alignment? We further test PPA with different forms of compressed history: raw utterances (Utterance), summaries (Summary), and structured personal information (Persona). As shown in Table 3, using extracted persona knowledge yields the best performance across all metrics. This supports our claim that structured and explicit personal knowledge is more effective than raw or summarized context when performing post-hoc alignment.

5 Conclusion

We introduced Post Persona Alignment (PPA), a novel framework for multi-session persona-based dialogue generation that defers persona grounding until after an initial response is generated, reversing the conventional retrieval-generation order. PPA structures information into salient, memory-like facts, and retrieves them only when truly relevant to the model's communicative intent. This approach reduces the cognitive load on the model and improves controllability, enabling more natural, diverse, and personalized dialogues. Experiments demonstrate that PPA significantly improves consistency, diversity, and persona relevance over prior methods, offering a flexible and effective solution for long-term conversational modeling.

Limitation

First, the effectiveness of PPA hinges on the quality of the initial response generated without persona conditioning. If it lacks coherence or relevance to the dialogue context, the subsequent retrieval and refinement stages may reinforce or propagate sub-optimal content rather than correct it. Second, our approach relies on embedding-based similarity between the initial response and persona memory for retrieval. This introduces sensitivity to the quality of the embedding space and the representational alignment between responses and memory entries. In particular, semantically important but lexically dissimilar content may be overlooked by Sentence-BERT, limiting retrieval coverage. Third, as the number of persona facts and extracted knowledge triples increases over many sessions, maintaining and retrieving from an expanding memory pool may introduce computational and latency bottlenecks. Moreover, fine-grained control over long-term memory retention (e.g., forgetting outdated traits or events) remains an open challenge.

Acknowledgments

This work was supported by JST-CRONOS Grant Number JPMJCS24K8, JSPS KAKENHI Grant Number JP23K28139, and the Institute of AI and Beyond of the University of Tokyo.

References

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jihwan Bak and Alice Oh. 2021. Few-shot personalized conversation systems via social networks. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pages 12781–12788. AAAI Press.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002, Dublin, Ireland. Association for Computational Linguistics.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. Cohesive conversations: Enhancing authenticity in multi-agent simulated dialogues. *arXiv preprint arXiv:2407.09897*.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606, Singapore. Association for Computational Linguistics.

Hana Kim, Kai Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. 2024. Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement.

In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 104–123, St. Julian's, Malta. Association for Computational Linguistics.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. 2023. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 707–719, Toronto, Canada. Association for Computational Linguistics.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 555–564.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. 2023. Response generation in longitudinal dialogues: Which knowledge representation helps? In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 1–11, Toronto, Canada. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, pages 1–17.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4279–4285.

Emily Reif, Minsuk Kahng, and Savvas Petridis. 2023. Visualizing linguistic diversity of text datasets synthesized by large language models. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pages 236–240. IEEE.

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*. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 3776–3784. AAAI Press.

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 167–177, Online. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. 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.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. SimOAP: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9945–9959, Toronto, Canada. Association for Computational Linguistics.

A Prompts

A.1 Prompt for Response Generation

See Figure 2.

```
{speaker} is chatting with {other}.

# The current conversation between {other}
and {speaker} is as follows:
{dialogue}

# Task: Output {speaker}'s response to
{other} in JSON.
Format: {{speaker}: <response>}
```

Figure 2: Prompt used for response generation. Curly braces { } indicate variable substitution.

A.2 Prompt for Response Refinement

See Figure 3.

```
{speaker} is chatting with {other}.
Their conversation is as follows:
{dialogue}

{speaker} was about to reply: "{response}"

# Task: Refine {speaker}'s response with the
following information:
{memory}

# Output {speaker}'s response to {other} in
JSON.
Format: {{speaker}: <{speaker}'s refined
response>}\\
```

Figure 3: Prompt used for response refinement.

B Example of Different Strategy Outputs

See Table 4.

Current Dialogue	
Francisco	Hey Rajiv! How's the guitar practice going?
Rajiv	It's going alright. I'm still learning the basics.
Francisco	That's cool. Have you thought about incorporating your guitar playing into your artwork?
Rajiv	Actually, I have. I was thinking about using sound waves as a way to create mathematical patterns.
Francisco	That sounds really interesting. We should definitely collaborate on a project involving artwork and music.
Rajiv	Definitely. We could also explore incorporating poetry into the mix.
Francisco	I like the way you think. Speaking of collaborations, have you signed up for those improv classes yet?
Selected Utterance from History Dialogues	
Latoya	Have you ever considered attending an improv class?
Rajiv	Actually, I have. Hailey Johnson invited me on her podcast to talk about my guitar playing, and she also mentioned attending an improv class together.
Francisco	That's awesome. I'm planning to potentially collaborate with Abigail Chen and Jennifer on artistic projects involving poetry and artwork with mathematical patterns.
Strategy	**Response of Rajiv**
BoB	Yes, I'm just trying to get a job with my new job.
DirectGen	Yes, I'm planning to attend improv classes soon. I think it would be a great way to explore connections between art, creativity, and spontaneity.
DialogRetr	I haven't signed up for the improv classes yet, but I'm definitely interested. **Let's plan to attend them together with Hailey Johnson!**
SimOAP	I have been considering taking improv classes. It could be a fun and creative way to explore new ideas and collaborate with others.
PPA	Not yet, but I'm definitely considering it. Improv could really help with thinking on my feet during creative projects. **I'm also interested in exploring the connections between math, nature, and art, especially after hearing about your collaborations with Abigail Chen and Jennifer on artistic projects involving poetry and artwork with mathematical patterns. It would be great to bring Hailey Johnson along to the improv classes too.**

Table 4: Example outputs. PPA provides the richest information from dialogue history (marked in bold), making the response more diverse and personalized.