Commonsense-augmented Memory Construction and Management in Long-term Conversations via Context-aware Persona Refinement

Hana Kim[1]   **Kai Tzu-iunn Ong**[2]   **Seoyeon Kim**[2]   **Dongha Lee**[2]   **Jinyoung Yeo**[2]

Department of Computer Science[1], Artificial Intelligence[2], Yonsei University
{hana.kim,ktio89,emseoyk,donalee,jinyeo}@yonsei.ac.kr

Abstract

Memorizing and utilizing speakers' personas is a common practice for response generation in long-term conversations. Yet, human-authored datasets often provide uninformative persona sentences that hinder response quality. This paper presents a novel framework that leverages commonsense-based persona expansion to address such issues in long-term conversation. While prior work focuses on not producing personas that contradict others, we focus on transforming contradictory personas into sentences that contain rich speaker information, by refining them based on their contextual backgrounds with designed strategies. As the pioneer of persona expansion in multi-session settings, our framework facilitates better response generation via human-like persona refinement. The supplementary video of our work is available at https://caffeine-15bbf.web.app/.

1 Introduction

Memorizing participants' personal information and conversing accordingly is important for dialogue systems to maintain long-term intimacy with users (Adiwardana et al., 2020). For that, studies have proposed datasets of long-term conversations, which require dialogue systems to memorize and utilize speakers' personas from past dialogue sessions to generate proper responses (Xu et al., 2022; Bae et al., 2022). Regardless, human-authored personas can be generic and over-simplified, limiting the generation of diverse and engaging responses.

Intuitively, this can be addressed by expanding existing personas with commonsense expansion (Majumder et al., 2020). However, such a naive remedy can raise contradiction between personas (*e.g.*, *"I am lazy"* and *"I clean my room every day"*), especially as sessions accumulate (Figure 7), hindering consistent response generation. While we can simply get rid of contradictory personas utilizing external modules such as models for natural language inference (NLI), it yields sub-optimal
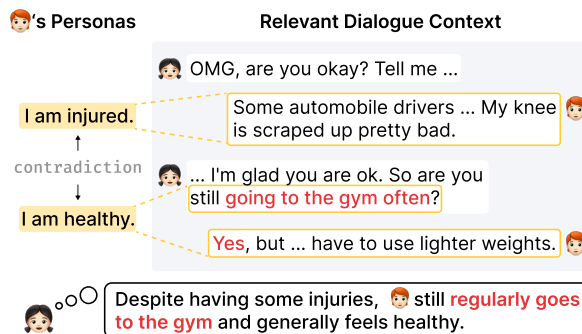

Figure 1: Contradictory personas can co-exist and provide rich speaker information for the conversation when their contexts are considered (an empirical example).

results (Section 3.2). Also, avoiding contradictory personas (Bae et al., 2022; Kim et al., 2023) does not align with human personality traits. Since human personality is context-dependent (van Oers et al., 2005), we naturally exhibit different personalities and behaviors in different contexts, allowing personas with contradictory interpretations to co-exist as one's personas, as shown in Figure 1.

Motivated by these, in this paper, we tackle such bottleneck of persona expansion in long-term conversations. Specifically, we focus on transforming contradictory personas into sentences that contain richer speaker information. To this end, we present ☕ CAFFEINE, a **C**ontext-**A**ware re**F**inement **F**ramework for contradictory p**E**rsonas **IN** long-t**E**rm conversations. CAFFEINE leverages large language models (LLMs) to iteratively refine the contradictory personas within/across the session(s) based on their contextual background with designed strategies. Our contributions are two-fold: (i) To the best of our knowledge, we are the first to explore commonsense-based persona expansion in multi-session settings; (ii) CAFFEINE enables better response generation in long-term conversations in both automatic and human evaluations. Also, it refines contradictory personas in a human-like manner, eliciting persona sentences that are superior in
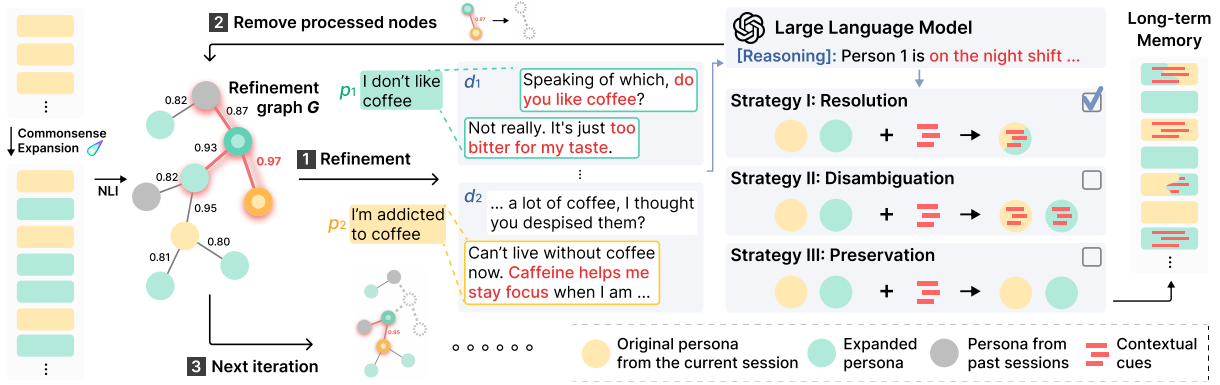
Figure 2: At the end of each dialogue session, CAFFEINE refines contradictory personas within/across the session(s) and saves the refined version to the dialogue model's memory for response generation in the next session.

various criteria while being cost- and time-efficient.

2 Approach

Long-term conversations involve multiple dialogue sessions. At the end of each session, we perform:

2.1 Commonsense-based Persona Expansion

Following Majumder et al. (2020), we perform commonsense expansion on personas derived from the conversation using COMET (Hwang et al., 2020). COMET generates commonsense knowledge based on cause-effect relation types (*e.g.*, XNEED and XWANT). For example, *"I drink coffee"* → *"I want to stay awake"*. Implementation details on COMET expansion are in Appendix A.1.

2.2 ☕ CAFFEINE

We present the overview of CAFFEINE in Figure 2.

2.2.1 Preparation: Graph Construction for Iterative Persona Refinement

After expansion, we identify contradictory personas by computing the probability of contradiction δ between all personas with an external NLI model. To refine contradictory personas cost- and time-efficiently, we adopt iterative refinement with a graph structure: Contradictory pairs with δ larger than a threshold μ are added as nodes V (edges $E = (\delta_1, \delta_2, \cdots, \delta_{|E|})$) to the refinement graph G.[1] Then, we locate the node (persona) p_1 with the largest $\Sigma\delta$ within its neighborhood. We select p_1 and the adjacent node p_2 with the highest δ with p_1 for the first refinement iteration (Algorithm 1).

---

[1]We empirically set μ as 0.8.

2.2.2 Context-aware Persona Refinement

As shown in Figure 1, personas causing contradiction can be logically acceptable and beneficial for conversations if contextual cues from their origin context are appended via commonsense reasoning. For that, we propose the following refinement strategies for the LLM to choose from:

Strategy I: Resolution. Inspired by entity resolution (Benjelloun et al., 2009), persona resolution resolves the contradiction between personas by seamlessly **merging** them into one informative sentence based on the contextual background from where they are derived (Figure 3 (a), Figure 1 is also an example of persona resolution).

Strategy II: Disambiguation. Contradiction between two statements can stem from the lack of contexts, known as pragmatic ambiguity (Macagno and Bigi, 2018). Drawn from entity disambiguation (Dredze et al., 2010), persona disambiguation **specifies** each persona with relevant information from their contextual backgrounds (Figure 3 (b)).

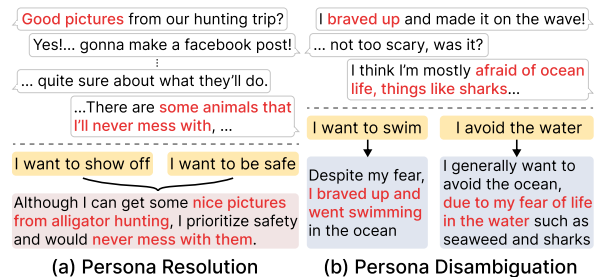(a) Persona Resolution   (b) Persona Disambiguation

Figure 3: Empirical demonstration of our strategies. Top: relevant contexts; Mid: contradictory personas; Bottom: refined persona(s).

Strategy III: Preservation. Due to the limitation of NLI models, personas predicted as contradictory

Settings	Session 2			Session 3			Session 4			Session 5		
	B-1	R-1	R-L	B-1	R-1	R-L	B-1	R-1	R-L	B-1	R-1	R-L
No Memory	20.75	19.38	15.16	20.42	19.53	15.09	19.88	19.56	14.98	19.87	20.16	15.33
GOLD	<u>21.19</u>	19.86	**15.50**	21.24	20.16	15.47	20.57	19.94	15.16	20.49	20.53	15.55
+ NLI-remove	20.81	19.98	15.26	21.04	20.28	15.52	21.33	20.69	15.91	21.43	20.75	15.95
+ NLI-recent	20.87	<u>20.09</u>	15.39	21.14	20.52	<u>15.71</u>	21.46	20.79	15.97	21.60	20.97	16.11
+ CAFFEINE	20.93	**20.18**	<u>15.47</u>	<u>21.41</u>	<u>20.72</u>	**15.86**	<u>21.67</u>	<u>21.00</u>	<u>16.15</u>	<u>21.92</u>	<u>21.23</u>	<u>16.31</u>
COMET-EXP	**21.23**	19.82	15.44	20.95	19.90	15.38	20.33	20.02	15.18	20.00	20.27	15.37
+ NLI-remove	20.72	19.96	15.27	21.12	20.40	15.56	21.66	20.77	15.88	21.77	20.91	16.01
+ NLI-recent	20.73	20.00	15.33	21.16	20.40	15.64	21.57	20.77	15.89	21.78	20.99	16.09
+ CAFFEINE	20.97	20.06	15.32	**21.63**	**20.73**	**15.86**	**21.97**	**21.10**	**16.18**	**22.26**	**21.32**	**16.37**

Table 1: Performance in response generation. **Bold** and <u>underline</u> show the best and second-highest in each column.

may be consistent and may not require refinement. Thus, we allow the LLM to preserve personas as they are when their contexts suggest so.

In practice, with contradictory personas $\mathcal{P} = (p_1, p_2)$ and relevant dialogue contexts $\mathcal{D} = (d_1, d_2)$ from where $\mathcal{P}$ are derived, we prompt the LLM to choose one out of the three strategies S with rationale and generate the refinement $\mathcal{R}$:

$$\mathcal{S}^* = \underset{\mathcal{S}}{\text{argmax}}\ P_{\text{LLM}}(\mathcal{S}|\mathcal{P}, \mathcal{D}) \quad (1)$$

$$\Rightarrow \mathcal{R}^* = \underset{\mathcal{R}}{\text{argmax}}\ P_{\text{LLM}}(\mathcal{R}|\mathcal{P}, \mathcal{D}, \mathcal{S}^*) \quad (2)$$

where $\Rightarrow$ denotes a sequential generation of tokens. D consists of w consecutive utterances.[2] When p is a persona generated by COMET, we use D of its original persona and concatenate the original persona with D. After refinement, we save $\mathcal{R}^*$ to long-term memory $\mathcal{M}$ and remove $\mathcal{P}$ from the graph G, and start the next iteration (Algorithm 1).

Algorithm 1 Iterative Graph Refinement

Require: Refinement graph $G(V, E)$
Ensure: The dialogue model's long-term memory $\mathcal{M}$
1: $\mathcal{M} \leftarrow \mathcal{M} \setminus V$
2: **while** $G \neq \emptyset$ **do**
3:     Select p_1 in V with the highest $\Sigma\delta$
4:     Select p_2, a neighbor of p_1 with the highest δ
5:     $(\mathcal{S}^*, \mathcal{R}^*) \leftarrow \text{Refine}(p_1, p_2)$
6:     $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{R}^*$
7:     Remove p_1, p_2 from G
8:     Remove isolated nodes from G
9: **return** $\mathcal{M}$

---

3 Experiments

3.1 Experimental Settings

Dataset. We use Multi-Session Chat (MSC) (Xu et al., 2022) to conduct experiments. MSC takes

the dialogues from Persona-Chat (Zhang et al., 2018) and extends their follow-up conversations throughout several sessions. Each session comes with speakers' personas authored by humans.

Models and baselines. In this work, we use Chat-GPT (OpenAI, 2023) for CAFFEINE and response generation (RG),[3] and Contriever (Izacard et al., 2021) to retrieve top-k relevant personas from long-term memory.[4] As for the NLI model, we use RoBERTa (Liu et al., 2019) fine-tuned on the MNLI dataset (Williams et al., 2017). To evaluate the effectiveness of CAFFEINE in RG, we apply it to: (i) COMET-EXP, human-authored personas with COMET expansion; (ii) GOLD, human-authored personas. We include this setting as a contradiction can also exist among un-expanded personas.[5] Also, to justify our choice to refine rather than remove, we compare CAFFEINE with two baselines: NLI-remove and NLI-recent.[6] The NLI-remove approach filters out personas that contradict at least one other persona with $\delta \geq 0.8$ via the NLI model. Similarly, the NLI-recent approach also uses the NLI model, but it differs by keeping the most recent persona in contradictory persona pairs and removing the older one (Bae et al., 2022), thereby prioritizing updated personas over time.

3.2 Results and Discussion

We present the empirical findings of the following research questions guiding our experiments:

---

[2]In our experiments, w differs depending on the persona annotation in the applied dataset. See Appendix A.4.

[3]Prompts for RG and CAFFEINE are in Appendix A.3.

[4]We set $k = 20$. Results with other k are in Appendix B.

[5]We report the statistics of contradiction in Appendix D. As our focus is persona expansion, extracting personas from conversations is out of the scope of this work.

[6]We test with MNLI and DNLI (Welleck et al., 2019) and report results using MNLI as it shows better performance. Results with DNLI are in Appendix D.

CAFFEINE vs.	GOLD	COMET-EXP	NLI-remove
Naturalness	73%*	71%*	79%*
Consistency	66%*	62%*	67%*
Specificity	55%	53%	51%
Engagingness	63%*	64%*	66%*
Overall	62%*	63%*	67%*

Table 2: Comparison of generated responses. We report CAFFEINE's **winning rate**. (*: p-value < 0.05)

RQ1: *Does* CAFFEINE *benefit response generation in long-term conversations?*
RQ2: *Does* CAFFEINE *refine personas in a way that aligns with human judgment?*
RQ3: *Is* CAFFEINE *cost- and time-efficient?*

CAFFEINE improves response generation (RQ1). To evaluate the efficacy of CAFFEINE, we conduct experiments on response generation (RG) using sessions 2 to 5 of each dialogue from MSC. Table 1 shows the results of RG in MSC with BLEU-1 (B-1), ROUGE-1 (R-1), and ROUGE-L (R-L) (Papineni et al., 2002; Lin, 2004). Applying CAFFEINE yields performance gains, which are more significant as sessions increase. Also, CAFFEINE consistently outperforms NLI-remove and NLI-recent, showing that leveraging contradictory personas elicits a more informative memory for RG than removing them. Compared to NLI-remove, the improved efficacy of NLI-recent is attributed to its focus on the recency of personas. By eliminating outdated personas from contradictory pairs, NLI-recent enhances RG, yielding responses more aligned with the current dialogue context. However, despite the enhancements in NLI-recent performance, CAFFEINE still exhibits superior performance. Furthermore, the performance brought by CAFFEINE exhibits a continuously rising trend as the number of previous sessions increases, while baselines yield a flat or downward tendency. These demonstrate the effectiveness of CAFFEINE in multi-session conversations. Table 2 shows the human evaluation results of randomly sampled 50 responses conducted by 3 judges from Amazon Mechanical Turk (Appendix F). CAFFEINE yields responses that are better (*i.e.*, winning) in several criteria. We provide examples of RG in Appendix E.

CAFFEINE elicits personas that align with human preference (RQ2). We sample 100 persona pairs refined with "resolution" or "disambiguation" and ask 3 judges *" whether they are contradictory*
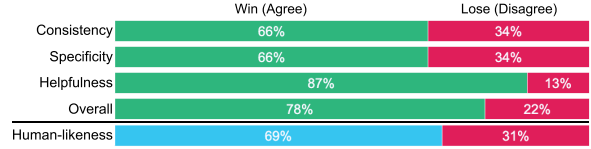

Figure 4: Human evaluation results on (i) refined personas and (ii) the refinement process (p-value < 0.05).

before refinement from a human standpoint". 89 samples that receive *"yes"* from all judges are used for the evaluation. Judges compare the refined version with its un-refined version and vote if they agree: it is less contradictory (Consistency); it provides more speaker information (Specificity); it is more useful when having a conversation with this person (Helpfulness); it has better quality (Overall); the refinement process is reasonable (Human-likeness). Figure 4 shows that personas refined by CAFFEINE are greater in all criteria, especially helpfulness. This supports our argument that contradictory personas become sentences with rich speaker information for the conversation if cues from their relevant contexts are included, and explains the performance gain in RG. Also, a 69% agreement on human-likeness demonstrates that CAFFEINE's refinement is in line with human judgment. Refinement examples are presented in Appendix E.

CAFFEINE refines personas in a cost- and time-efficient manner (RQ3). In CAFFEINE, we remove refined (p_1, p_2) from G after refinement. Figure 5 compares this with a setting without such removal, *i.e.*, all $|E|$ contradictory persona pairs in G are all refined (denoted as ALL). While yielding similar RG performance (Session 2-5), ours requires significantly fewer API calls per dialogue per session, especially as the sessions accumulate (9-fold $\rightarrow$ 21-fold more cost- and time-efficient).

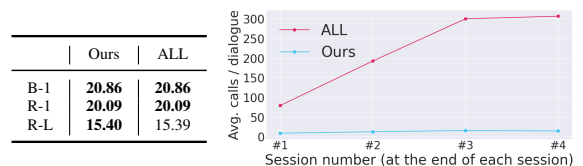|	Ours	ALL
B-1	**20.86**	**20.86**
R-1	**20.09**	**20.09**
R-L	**15.40**	15.39

Figure 5: Cost and time efficiency of our algorithm.

4 Related Work

Many studies have utilized commonsense knowledge for response generation. For instance: leveraging knowledge from a general-purpose knowledge model (Zhou et al., 2022b; Wu et al., 2022;

Liu et al., 2022; Li et al., 2023); training commonsense generators for dialogues via human-annotated dataset (Ghosal et al., 2022); formulating commonsense-linking between knowledge graphs and dialogues (Gao et al., 2022); modeling speakers' mutual beliefs before a response (Zhou et al., 2022a); integrating implicit information in dialogues into rationale for more effective response generation (Chae et al., 2023). While most work focuses on speaker utterances, we leverage speaker personas to address commonsense knowledge in response generation.

5 Conclusion

This work pioneers commonsense-based persona expansion in multi-session settings and presents a context-aware refinement framework that leverages contradictory personas to elicit a memory with richer speaker details. ☕ CAFFEINE improves response generation in long-term conversations and demonstrates human-like refinement of contradictory personas while being cost- and time-efficient.

6 Limitations

Our study has the following limitations: (1) Apart from the proposed CAFFEINE, our results can be affected by the quality of commonsense models and the knowledge graph on which they are trained. As future work, we plan to leverage LLM for persona expansion; (2) Our refinement graph stores contradictory personas that are predicted as `contradiction` with a probability higher than a pre-defined threshold by the NLI model. Our framework may miss personas that actually need a refinement due to the limitation of the NLI model; (3) While we pioneer the commonsense-based persona expansion in multi-session settings, we only consider one speaker's persona at a time in our refinement framework. Since different people can demonstrate different personality traits and behaviors in the same commonly experienced event (*e.g.*, discussed topic), we acknowledge there can be potential performance gain in response generation if such modeling is included; (4) In this work, we employ LLMs to generate responses based on the dialogue context and retrieved memories (*i.e.*, both speakers' personas) in a zero-shot setting. However, since the refined personas tend to be longer and contain more information, it is possible that the LLM can not fully utilize the presented personas in its inputs as they get longer (Liu et al., 2023). We

plan to address a better utilization of LLM's input texts for response generation in future work.

7 Ethical Statement

LLMs and COMET can generate sensual, harmful, biased, offensive, or violent content. Authors avoid such content from appearing in the main text, figure, and appendix. We guarantee fair compensation for workers we hire on Amazon Mechanical Turk. We ensure an effective pay rate higher than $18 per hour based on the estimated time required to complete the tasks.

8 Acknowledgements

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

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.

Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. 2009. Swoosh: a generic approach to entity resolution. *The VLDB Journal*, 18:255–276.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore. Association for Computational Linguistics.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 277–285, Beijing, China. Coling 2010 Organizing Committee.

Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. ComFact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1656–1675, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149, Dubrovnik, Croatia. Association for Computational Linguistics.

Minju Kim, Beong-woo Kwak, Youngwook Kim, Hong-in Lee, Seung-won Hwang, and Jinyoung Yeo. 2022. Dual task framework for improving persona-grounded dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10912–10920.

Siheng Li, Wangjie Jiang, Pengda Si, Cheng Yang, Qiu Yao, Jinchao Zhang, Jie Zhou, and Yujiu Yang. 2023. Enhancing dialogue generation with conversational concept flows. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1514–1525, Dubrovnik, Croatia. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiting Liu, Liang Li, Beichen Zhang, and Qingming Huang. 2022. Think beyond words: Exploring context-relevant visual commonsense for diverse dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3106–3117, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabrizio Macagno and Sarah Bigi. 2018. Types of dialogue and pragmatic ambiguity. *Argumentation and Language—Linguistic, Cognitive and Discursive Explorations*, pages 191–218.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt. https://openai.com/blog/chatgpt.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kees van Oers, Margreet Klunder, and Piet J Drent. 2005. Context dependence of personalities: risk-taking behavior in a social and a nonsocial situation. *Behavioral Ecology*, 16(4):716–723.

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.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Sixing Wu, Ying Li, Ping Xue, Dawei Zhang, and Zhonghai Wu. 2022. Section-aware commonsense knowledge-grounded dialogue generation with pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 521–531, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. 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.

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.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022a. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022b. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

A Implementation Details

A.1 Commonsense Expansion with COMET

At the end of each dialogue session, we augment personas derived from the current session with new personas via COMET (Hwang et al., 2020), a widely used commonsense model generating rich and diverse commonsense expansions of a given statement based on cause-effect relation. Among the 23 possible candidate relation types, following prior works on commonsense-based persona expansion (Majumder et al., 2020; Kim et al., 2022), we choose 9 relation types: XATTR, XEFFECT, XINTENT, XNEED, XREACT, XWANT, OEFFECT, OREACT, and OWANT for our expansion, where the prefix 'x' indicates an effect or cause on that person and 'o' denotes others. After persona expansion via COMET, we leverage an external NLI model to initially filter out improper expansion. Specifically, when a new persona p^n is generated based on an original persona p^o (1 original persona yields nine 9 personas), we filter it out if the NLI model predicts the logical relationship between p^n and p^o is *contradiction* with $\delta > 0.33$. Note that this is different from the NLI-remove baseline, as here we solely address a one-to-one relationship between a generated persona and its corresponding original persona, while the latter addresses the contradiction among all possible combinations of personas within/across the dialogue session(s). We report the statistics of this initial filtering in Table 3.

At the End of	Filtered (%)	Total
Session #1	2830 (6.84 %)	41391
Session #2	2715 (7.39 %)	36718
Session #3	2935 (7.43 %)	39523
Session #4	2971 (7.58 %)	39198

Table 3: Initial filtering of improper expansion.

A.2 Contriever

In our experiments on persona-grounded response generation (RG), we adopt Contriever (Izacard et al., 2021) as the memory retriever to retrieve top-k relevant personas from long-term memory based on the current conversation. Contriever is a dense information retriever trained with unsupervised contrastive learning. Even without supervision, it has shown remarkable capabilities in information retrieval tasks, particularly in demonstrating competitiveness with BM25 in Recall at 100 (R@100) on the benchmark for zero-shot retrieval.

A.3 Large language model

In this work, we employ ChatGPT for the proposed CAFFEINE and response generation. ChatGPT is an LLM with 175B parameters based on Instruct-GPT (Ouyang et al., 2022)[7]. ChatGPT is trained to follow instructions given by users and return requested information in a conversational manner. We use LangChain[8] to send API calls to OpenAI API. The prompt used in CAFFEINE and response generation are in Table 6 and Table 7, respectively.

A.4 Linking Personas to their Contextual Backgrounds

In the adopted MSC dataset, human annotators summarize information in a speaker's utterance and use it to derive a persona sentence. As demonstrated in Figure 6, since not every utterance contains enough information to conclude a persona for that speaker, some utterances are not paired with a persona sentence. In our experiment for context-aware persona refinement, we utilize contradictory personas $\mathcal{P} = (p_1, p_2)$ and their contextual backgrounds, *i.e.*, relevant dialogue contexts $\mathcal{D} = (d_1, d_2)$ from where they are derived. d consists of w consecutive sentences. In practice, w can differ, as we link each persona with their relevant dialogue context by separating the past conversation into dialogue fragments based on utterances that have corresponding persona sentences. For instance, the d_i for persona p_i will be $d_i = (u_1, u_2)$, and $d_{i+1} = (u_3, u_4, \cdots, u_6)$ for p_{i+1}.
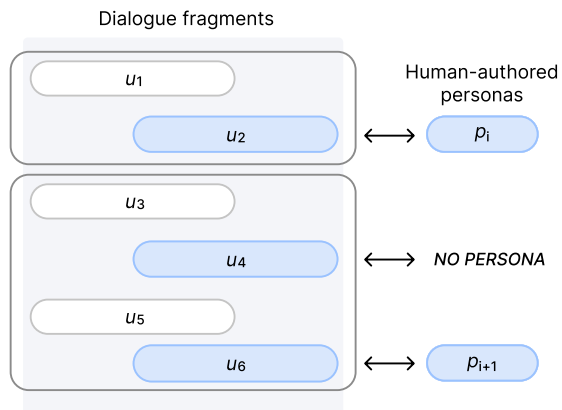

Figure 6: Demonstration of personas and their contextual backgrounds in the MSC dataset.

A.5 Computational Resources and API Cost

We run Contriever and the NLI model on eight NVIDIA RTX A5000 GPUs. For ChatGPT API usage, we use $35.52 on CAFFEINE's refinement, and $27.09 on response generation.

B Performance in Response Generation

In response generation, top-k relevant persona sentences are retrieved from the long-term memory to assist response generation. In the main text, we report the mode performance in response generation with $k = 20$, the results with $k = 12$ and $k = 30$ are presented in Table 4.

C Contradictory Personas in Multi-session Conversations

As human personalities are context-independent, we display different personalities in different contexts and adapt to new situations. This naturally leads to personas with contradictory literal interpretations to co-exist as one's persona. Such a phenomenon does not harm human conversations. However, contradictions between personas can lead to inconsistent response generation, hindering user interest in the dialogue systems.

In our study on the Multi-session Chat dataset, we first find that contradictory personas exist in human-authored personas (Figure 7 (a)). Then, we show that expanding existing human-authored personas via commonsense expansion can lead to orders-of-magnitude more contradictory personas that hinder user interest in the conversation (Figure 7 (b)) (Kim et al., 2023).

Personas can contradict other personas from the same sessions (intra-session) and from the previous sessions (inter-session). When comparing COMET-EXP with the human-authored personas (GOLD), we observe that as the number of previous sessions increases, the intra-session contradiction slightly increases, whereas the inter-session contradiction skyrockets significantly. Although such a rising trend appears similarly in GOLD and COMET-EXP, the total count in COMET-EXP is order-of-magnitude larger. This supports the necessity of CAFFEINE, which refines the contradictory personas in the long-term memory of dialogue models in multi-session settings.

D CAFFEINE vs. NLI models

Noteworthily, Figure 8 shows that CAFFEINE determines that 65.45% of contradictory personas

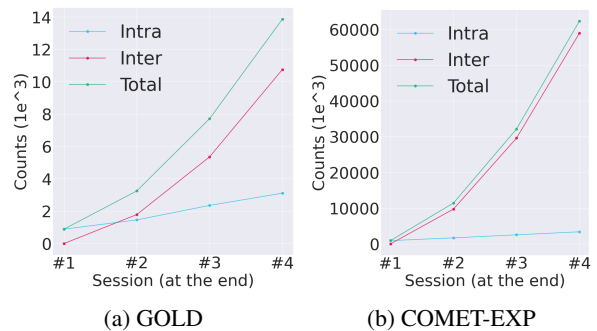(a) GOLD   (b) COMET-EXP

Figure 7: Contradiction among human-authored original personas (GOLD) and expanded personas (COMET-EXP). The blue, red, and green lines represent the intra-session, inter-session, and total contradictory persona pairs, respectively.

(with $\delta \geq 0.80$) can be consistent without requiring any refinement when their contextual backgrounds are taken into account, indicating that our context-aware refinement can address the simplification of NLI models where they often solely compare the semantic representation of two statements without reasoning over their contexts. We employ two NLI models: the MNLI model (referred to as 'NLI model') and the DNLI model. Results with the DNLI model are presented in Table 5.
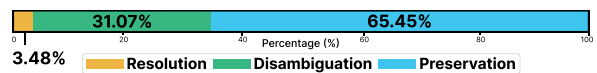
Figure 8: Proportion of selected strategies highlights the limitation of the NLI model.

E Examples of Refinement and Response Generation

We provide examples of response generation in Figure 9 and Figure 10. We can observe that while baselines provide short personas and yield inconsistent or unconstructive responses (green underlines), CAFFEINE offers informative personas (color red) and leads to a response that provides constructive suggestion for Persona A's Spanish learning (Figure 9) and a response that reflects Person B's situation and what B is looking for in a car (Figure 10).

Apart from the empirical examples demonstrated in figures in the main text, we have provided more examples for CAFFEINE's refinement in Table 8, 9, and 10.

112

F Details on Human Evaluation.

F.1 Response Quality

We outsource a human evaluation comparing the generated responses from our setting and those from the baselines via Amazon Mechanical Turk (AMT). We show the interface for the evaluation in Figure 11. We ask the human judges to compare the responses based on the following criteria:

- Naturalness: Which response is more human-like?

- Consistency: Which response is more consistent (aligned) with the dialogue context?

- Specificity: Which response contains more speaker information?

- Engagingness: Which response is more interesting?

F.2 Refinement Quality

We outsource a human evaluation comparing the personas before/after CAFFEINE via Amazon Mechanical Turk (AMT). We show the interface for this evaluation in Figure 12.

We ask the human judges to compare the persona before and after refinement based on the following criteria:

- Consistency: Is the refined version less contradictory or more reasonable than before?

- Specificity: Does the refined version describe a person more specifically?

- Helpfulness: Can the refined version be more helpful if you are having a conversation with this person?

- Overall: Overall, do you prefer the refined version?

- Human-likeness: Is the refinement process (generated rationales) reasonable?

Note that before assessing CAFFEINE's refinement, we ask workers to determine whether the personas before refinement are actually contradictory from human standpoints. The assessment of the above criteria only begins if the answer is positive. Therefore, the reported human evaluation results are based on 89 out of 100 samples we provided.

Dialogue Context	
A: Do you know where to get a copy of the The Fugitive movies?	
B: Sure! if you go to justwatch.com it will tell you where to buy a digital copy or you can go to walmart.com or amazon. How's your Spanish learning going?	
A: Perfect, thank you! I will look at all those places. I am still learning, and it is slow but I am trying!	

Speaker A's Personas (draw from retrieved top-20 memories)		Generated Response
GOLD	I am trying to learn Spanish better. I'm learning Spanish. The Fugitive is my all time favorite movie. My Spanish language progress is not much I'd expect. I like to speak in Spanish. I like to watch that movie.	B: That's great that you're putting in the effort to learn Spanish! The Fugitive is a fantastic movie, by the way. <u>How's your progress in Spanish been so far?</u>
COMET-EXP	I am trying to learn Spanish better. I'm learning Spanish. The Fugitive is my all time favorite movie. I want to learn Spanish. I need to learn Spanish.	B: That's great to hear! Learning a new language can take time, but as long as you're putting in the effort, <u>you'll make progress. Keep it up!</u>
COMET-EXP +CAFFEINE	I am trying to learn Spanish better. **I am trying to learn Spanish better by watching shows and movies with Spanish subtitles.** The Fugitive is my all time favorite movie. I'm learning Spanish. I'm learning Spanish by watching movies. I need to read books to learn Spanish faster. **I need to go to the library to learn Spanish, and my girlfriend might be interested in learning too.**	B: That's great to hear! Learning a new language can take time, but as long as you're putting in the effort, you'll get there. Have you tried watching any Spanish movies with subtitles to practice?

Figure 9: **Example of Response Generation #1.** While baselines yield short personas and inconsistent or unconstructive responses (green underlines), CAFFEINE offers informative personas (color red) and leads to a response that provides constructive suggestion for Person A's Spanish learning.

Dialogue Context

A: I always think used is the way to go; new cars lose so much value the second they leave the lot. I am all for them! If I didn't live in a tiny town with no charging stations I would've definitely gotten an electric car. Were you thinking about getting one?

B: I totally agree! I thought about getting electric because I've been thinking about the environment and the earth lately. I think I might change my major.

A: I really think you should do it! You can even get used ones, especially hybrids, I've seen more and more popping up for sale. Oh really, to what?

Speaker B's Personas (draw from retrieved top-20 memories)		Generated Response
GOLD	I prefer imported cars. I want to buy a car. I am looking for a car. I cannot afford a hyundai genesis. I used to have a Ford Edge.	B: I cannot afford a hybrid at the moment, but maybe I'll consider it in the future. I was actually considering a pre-owned luxury car. Maybe something like a used BMW or Mercedes-Benz. What do you think?
COMET-EXP	I get a new car. I want to have a good car. I want to have a car. I want to get a car. I need to have a car. I prefer imported cars. I want to buy a car. I am looking for a car. I want to buy a new car.	B: I want to get a new car, preferably an imported one. I'm looking for something that combines good value and performance. Maybe a new BMW or Mercedes-Benz.
COMET-EXP + CAFFEINE	I want to have a car. **As I work on my PhD, I'm considering buying a used car like a Honda Fit for its affordability and spaciousness.** I want to get a car. I prefer imported cars. **Since I've been laid off from my job at Cisco, I'm hoping I can still buy a car.** I cannot afford a hyundai genesis. I want to buy a car.	B: I am considering sticking with a used car for now, especially given my recent employment situation. A Honda Fit seems like a practical choice, given its affordability and spaciousness. It might be a sensible option considering my current circumstances.

Figure 10: **Example of Response Generation #2.** While baselines yield short personas, CAFFEINE offers informative personas (color red) and leads to a response that reflects Person B's situation and what B is looking for in a car.

We are surveying qualities for **responses** from dialogue.

Specifically, you'll be given a piece of dialogue context, and a response to follow the dialogue. You'll be asked to **compare which response is better** in terms of different aspects, and **specify which aspect was most important** for judging.

Guidelines:
1. **[Q1~5] First, choose which response is better regarding the given aspect.**
   ○ Try to focus on quality over quantity. **Contentful/high-quality** response doesn't need to be lengthy.
2. **[Q6] Second, choose which aspect influenced you the most when judging the overall quality.**
   ○ If some factor other than the ones in Question 1~6 had the biggest influence, please select "Other" and specify.

Dialogue Context
${context}

Response candidate 1	*Response candidate 2*
${response_ours}	${response_other}

Question 1. Which response is more **natural** (human-like)?

〇1 〇2

Question 2. Which response is more **interesting**?

〇1 〇2

Question 3. Which response is more **consistent** (well aligned) with the dialogue context?

〇1 〇2

Question 4. Which response is more **specific**(contains more information about speakers)?

〇1 〇2

Question 5. Which response do you like more **overall**?

〇1 〇2

Question 6. Which aspect affected you the most when judging the overall quality?

〇 Naturalness   〇 Interesting   〇 Consistency   〇 Specificity   〇 Other:

Optional feedback?   (expand/collapse)

Submit

Figure 11: Interface for human evaluation on response quality.

We are surveying **qualities** of personas from dialogues.

Specifically, you'll be given Persona 1 and 2 that are potentially contradictory, its dialogue context, and its refined persona. Each persona is expanded from Source persona. Refined persona is either a single merged persona, or two individually specified persona of Persona 1 and 2. You'll be asked to **compare which persona(s) is better** in terms of different aspects, and **assess rationale** that was used for refinement.

Guidelines:
There are four choices: [Definitely Disagree/Agree] and [Slightly Disagree/Agree].
Please trust your instincts and choose [Definitely] if you would feel more confident giving one response, versus the other one.
1. **[Q1] First, answer if Persona 1 and 2 contradict each other.**
2. **[Q2-6] Second, assess refined persona in terms of different aspects.**

Persona 1 (Before refinement)	*Persona 2 (Before refinement)*
${persona_1}	${persona_2}

Question 1. Are Persona 1 and Persona 2 contradictory?
○ Definitely Disagree   ○ Slightly Disagree   ○ Slightly Agree   ⦿ Definitely Agree

Refined Persona
${refined}

Question 2.
(If there is one refined persona) Is refined persona **reasonable**?
(If there are two refined personas) Are refined personas **less contradictory**?
○ Definitely Disagree   ○ Slightly Disagree   ○ Slightly Agree   ○ Definitely Agree

Question 3. Is refined persona more specific to describe a person?
○ Definitely Disagree   ○ Slightly Disagree   ○ Slightly Agree   ○ Definitely Agree

Dialog Context of Persona 1	*Dialog Context of Persona 2*
${context_1}	${context_2}

Question 4. Is refined persona more useful when having a dialogue with this person?
○ Definitely Disagree   ○ Slightly Disagree   ○ Slightly Agree   ○ Definitely Agree

Rationale for refinement
${rationale}

Question 5. Do you think this rationale is appropriate for refining contradiction in Persona 1 and 2?
○ Definitely Disagree   ○ Slightly Disagree   ○ Slightly Agree   ○ Definitely Agree

Question 6. Is Refined persona more preferrable overall than personas before refinement?
○ Definitely Disagree   ○ Slightly Disagree   ○ Slightly Agree   ○ Definitely Agree

Optional feedback?   (expand/collapse)

Submit

Figure 12: Interface for human evaluation on refinement quality.

117

Memory	Session 2			Session 3			Session 4			Session 5		
	B-1	R-1	R-L	B-1	R-1	R-L	B-1	R-1	R-L	B-1	R-1	R-L
None	20.75	19.38	15.16	20.42	19.53	15.09	19.88	19.56	14.98	19.87	20.16	15.33
$k = 12$												
GOLD	**21.18**	19.78	**15.46**	<u>21.26</u>	20.11	15.43	20.58	19.97	15.17	20.38	20.40	15.42
+ NLI-remove	20.74	19.83	15.19	21.05	20.27	15.51	21.09	20.35	15.62	21.22	20.56	15.78
+ CAFFEINE	20.91	<u>20.03</u>	15.33	21.20	<u>20.52</u>	**15.74**	21.46	<u>20.77</u>	**15.94**	<u>21.62</u>	20.97	**16.11**
COMET-EXP	<u>21.04</u>	19.63	15.32	20.89	19.88	15.27	20.20	19.84	15.14	20.12	20.43	15.50
+ NLI-remove	20.68	19.89	15.19	21.04	20.21	15.42	<u>21.49</u>	20.70	15.81	21.57	20.73	15.88
+ CAFFEINE	20.99	**20.05**	15.32	**21.41**	**20.55**	<u>15.71</u>	**21.66**	**20.83**	<u>15.93</u>	**21.86**	<u>20.96</u>	<u>16.07</u>
$k = 20$												
GOLD	<u>21.19</u>	19.86	**15.50**	21.24	20.16	15.47	20.57	19.94	15.16	20.49	20.53	15.55
+ NLI-remove	20.81	19.98	15.26	21.04	20.28	15.52	21.33	20.69	15.91	21.43	20.75	15.95
+ CAFFEINE	20.93	**20.18**	<u>15.47</u>	<u>21.41</u>	<u>20.72</u>	**15.86**	<u>21.67</u>	**21.00**	<u>16.15</u>	<u>21.92</u>	<u>21.23</u>	<u>16.31</u>
COMET-EXP	**21.23**	19.82	15.44	20.95	19.90	15.38	20.33	20.02	15.18	20.00	20.27	15.37
+ NLI-remove	20.72	19.96	15.27	21.12	20.40	<u>15.56</u>	21.66	20.77	15.88	21.77	20.91	16.01
+ CAFFEINE	20.97	<u>20.06</u>	15.32	**21.63**	**20.73**	**15.86**	**21.97**	**21.10**	**16.18**	**22.26**	**21.32**	**16.37**
$k = 30$												
GOLD	20.88	19.65	<u>15.45</u>	21.09	20.18	15.56	20.50	19.89	15.09	20.41	20.47	15.46
+ NLI-remove	20.65	19.85	15.14	21.09	20.43	15.68	21.50	20.83	15.98	21.59	20.93	16.07
+ CAFFEINE	20.89	**20.13**	15.43	<u>21.42</u>	<u>20.77</u>	<u>15.96</u>	<u>21.73</u>	<u>21.07</u>	<u>16.19</u>	<u>22.01</u>	<u>21.29</u>	<u>16.32</u>
COMET-EXP	**21.40**	19.89	**15.52**	21.06	20.10	15.40	20.38	20.03	15.27	20.06	20.50	15.59
+ NLI-remove	20.60	19.86	15.20	21.02	20.33	15.57	21.35	20.62	15.83	21.71	20.97	16.05
+ CAFFEINE	<u>20.96</u>	<u>20.11</u>	15.37	**21.73**	**20.85**	**15.99**	**22.17**	**21.24**	**16.27**	**22.57**	**21.58**	**16.54**

Table 4: Performance in response generation. **Bold** and <u>underline</u> show the best and second-highest in each column.

Settings	Session 2			Session 3			Session 4			Session 5		
	B-1	R-1	R-L	B-1	R-1	R-L	B-1	R-1	R-L	B-1	R-1	R-L
No Memory	20.75	19.38	15.16	20.42	19.53	15.09	19.88	19.56	14.98	19.87	20.16	15.33
GOLD	<u>21.19</u>	19.86	**15.50**	21.24	20.16	15.47	20.57	19.94	15.16	20.49	20.53	15.55
+ DNLI-remove	20.87	20.07	15.31	21.15	20.50	15.70	21.37	20.82	15.97	21.52	20.95	16.08
+ DNLI-recent	20.92	20.09	15.36	21.16	20.58	15.80	21.36	20.83	16.01	21.60	21.08	16.20
+ CAFFEINE	20.94	**20.15**	15.41	21.33	**20.69**	<u>15.89</u>	<u>21.54</u>	<u>21.01</u>	<u>16.17</u>	<u>21.75</u>	21.18	<u>16.30</u>
COMET-EXP	**21.23**	19.82	<u>15.44</u>	20.95	19.90	15.38	20.33	20.02	15.18	20.00	20.27	15.37
+ DNLI-remove	20.81	20.01	15.26	21.13	20.46	15.73	21.53	20.96	16.12	21.66	21.06	16.18
+ DNLI-recent	20.92	<u>20.10</u>	15.41	<u>21.35</u>	**20.69**	**15.91**	21.51	20.93	16.11	21.72	<u>21.20</u>	**16.34**
+ CAFFEINE	20.89	<u>20.10</u>	15.40	**21.37**	<u>20.62</u>	15.81	**21.82**	**21.06**	**16.19**	**22.07**	**21.21**	<u>16.30</u>

Table 5: Performance in response generation with DNLI model. **Bold** and <u>underline</u> show the best and second-highest in each column.

Prompt
You will be provided with two contradictory persona sentences, along with their source personas and the dialogue fragments from which these persona sentences were derived. Your task is to resolve the contradiction between the two persona sentences based on the dialogue fragments and the source persona of each contradictory persona. You can use these two strategies: **[Resolution]**: If the two personas are based on the same event but change over time (possibly due to a temporal difference or other events in between), adjust and aggregate them into one new persona sentence. **[Disambiguation]**: If the contradiction between them isn't due to time changes or they are derived from unrelated events in the first place, utilize dialogue fragments to rewrite (clarify/specify) each persona. First, provide a rationale for your choice (**[Resolution]** or **[Disambiguation]**). Then, based on this rationale, generate refined persona sentence(s). If the two personas are not contradictory, generate **[NO_CONFLICT]**.

Example 1:
Persona 1: I am a programmer.
Dialogue fragment of Persona 1:
A: As a computer programmer, I spend a lot of time writing and debugging code. It's rewarding to see my work contribute to the development of functional and efficient software.
Source Persona: I am a programmer.
Persona 2: I get fire.
Dialogue fragment of Persona 2:
A: I got fired. It was quite unexpected, and I'm still processing everything.
B: I'm really sorry to hear that. Do you want to talk about what happened?
Source Persona: I don't have a job right now.
Rationale: There is a temporal connection between the two personas. Persona 1 is about being a programmer, whereas Persona 2 is about having been fired. Both personas can exist over time with Persona 2 occurring after Persona 1.
[Resolution]: I am a programmer who has recently been fired.

Example 2:
Persona 1: I feel happy.
Dialogue fragment of Persona 1:
A: My favorite baseball team won this season! I'm so excited!
B: Wow, really? That's great news!
Source Persona: My team won the league.
Persona 2: I have been feeling quite stressed.
Dialogue fragment of Persona 2:
A: Work has been overwhelming with tight deadlines and high expectations. It's been hard to find time for relaxation and self-care.
B: With stress piling up, it's important to find ways to unwind and destress. Maybe I'll try meditation or yoga to help manage my anxiety.
Source Persona: I work very hard.
Rationale: The two personas do not reflect changes over time but rather different emotional states in response to separate circumstances; one, a moment of happiness due to a favorite team winning, and the other, underlying stress caused by work pressures.
[Disambiguation]:
- **Persona 1:** I feel happy when my favorite baseball team wins.
- **Persona 2:** I am a person dealing with work-related stress and looking for ways to manage anxiety.

Example 3:
Persona 1: I am a vegetarian.
Dialogue fragment of Persona 1:
A: I've been a vegetarian for about 5 years now. It started for health reasons, but now it's also about sustainability and animal welfare.
Source Persona: I am a vegetarian.
Persona 2: I enjoy reading fiction books.
Dialogue fragments of Persona 2:
A: One of my favorite pastimes is reading fiction. I love getting lost in different worlds and lives through the pages of a good novel.
Source Persona: I enjoy reading fiction books.
Rationale: The two persona sentences do not contradict each other as they pertain to different aspects of the speaker's identity. One persona is about dietary preference (being a vegetarian), and the other is about a hobby or interest (enjoying reading fiction books). There is no inherent conflict between being a vegetarian and enjoying reading fiction, so the two persona sentences can coexist without the need for resolution or disambiguation.
[NO_CONFLICT]

Example 4: ...

Table 6: The prompt for CAFFEINE (Five-shot setting, Examples 4 and 5 are omitted in this table). The "preservation" strategy is represented as **[NO_CONFLICT]** in our prompt.

Prompt
You will be generating the next turn of a given dialogue context between Speaker A and Speaker B. Alongside the dialogue context, you'll be given persona statements about both speakers. Your response should be 1-2 sentences, utilizing the persona statements as guidance to create an appropriate reply. Generate appropriate answers using given persona statements as memory.   What is the most appropriate next utterance (3 sentences max)?   **Persona Statements of A:** {A's personas within the top-k retrieved personas from long-term memory}   **Persona Statements of B:** {B's personas within the top-k retrieved personas from long-term memory}   **Dialogue:** {dialogue context}   **Response:** {generation result}

Table 7: The prompt for response generation (zero-shot).

Example 1 - [Resolution]
Before refinement:
`B's Persona 1`: I feel happy.
`B's Persona 2`: I feel sad.
Contextual Backgrounds (relevant dialogue contexts):
In this example, both personas are derived from the same dialogue context.
A: It's really bad from the first of the year until tax deadline day but it keeps me pretty busy the rest of the year as well. I will need to think it over. It does get lonely sometimes.
B: I can only imagine. That was another thing about being a vet that was hard. People expected you to be on call all the time, so I didn't have a chance for any sort of social life. That's gotten a lot better since I quit. I've found a good group of friends to hang out with at least a few times a week at a local cafe.
CAFFEINE's output:
`Explanation:` The two persona sentences reflect contradictory emotional states. Persona 1 expresses happiness due to having a good group of friends to hang out with, whereas Persona 2 expresses sadness and loneliness from lacking a social life. Both personas are based on the same events and experiences, indicating a change in emotional state over time.
`After refinement:` I used to feel sad and lonely when I was a vet, but now I feel happy because I have a good group of friends to hang out with at a cafe every week.

Example 2 - [Disambiguation]
Before refinement:
`B's Persona 1`: I want to play with the dogs.
`B's Persona 2`: I want to eat them.
Contextual Backgrounds (relevant dialogue contexts):
A: Yeah definitely, that's great to hear you're enjoying your new vegan diet. Oh, cute! I love labs. Yes, my puppies keep me busy as well. The training phase is definitely a lot of work. How old are your labs now?
B: Almost 4. Still act like puppies running around like crazy. I think I wouldn't get any exercise if it weren't for them though, so that is a good thing for me.
`- - - - -`
A: That sounds delicious!!! I agree, a recipe exchange would be great. True, vegan cheese does help with that. Have you found any favorite brands for vegan cheese? I found a couple that I like, and some that I didn't enjoy as much.
B: I have tried a couple, none that I've enjoyed so far. the next on my list to try is cashew cheese, which I have heard is superior, so I am excited to try it. I hear it makes great nachos, and I miss nachos so much!
CAFFEINE's output:
`Explanation:` The two personas do not reflect changes over time but rather different topics of discussion. Persona 1 is about the speaker's dogs and their age, while Persona 1 is about discussing vegan cheese and favorite brands.
`After refinement:`
`B's Persona 1`: My 4-year-old labs (labrador) keep me active and it's a good thing for me.
`B's Persona 2`: I enjoy trying different brands of vegan cheese, although I haven't found a favorite yet.

Table 8: More examples of persona refinement done by CAFFEINE. Color magenta highlights the contextual cues from the relevant dialogue contexts.

Example 1 - [Resolution]
Before refinement:
`A's Persona 1:` I feel excited.
`A's Persona 2:` I get tired.
Contextual Backgrounds (relevant dialogue contexts):
In this example, both personas are derived from the same dialogue context.
B: I am sure you feel so small and at peace up there. When will your mission make its trip? I imagine training for space is long and strenuous.
A: It absolutely is! The trip is scheduled for around this time next year. I'm going to have to get on a pretty strict diet and exercise routine to make sure that I am fit enough for the trip.
CAFFEINE's output:
`Explanation:` The two persona sentences contradict each other in terms of the emotional state of the speaker. One persona expresses excitement, while the other expresses tiredness. However, both personas are based on the same event or circumstance of training for a space mission. Thus, there is a temporal aspect to this contradiction.
`After refinement:` I am currently feeling tired due to the strenuous training for my upcoming space mission, but I am also excited about the opportunity.
Example 2 - [Disambiguation]
Before refinement:
`A's Persona 1:` I feel relaxed.
`A's Persona 2:` I feel tired.
Contextual Backgrounds (relevant dialogue contexts):
B: That sounds like so much fun. I have never been fishing in my life. Is it difficult to learn?
A: There's a bit of a knack to the rod, but once you get that down, it's pretty simple. It's a lot of just sitting around and waiting. It's very peaceful and relaxing.
- - - - -
B: We are always on the lookout for a new restaurant. Have you tried a great place recently?
A: I have been so busy with work, that I haven't! I was hoping you might have a recommendation. We definitely need to go somewhere with great crab legs, or else a nice sushi joint. Somewhere with delicious cocktails as well!
CAFFEINE's output:
`Explanation:` The two persona sentences do not reflect changes over time but rather different emotional states and interests. Persona 1 expresses feeling happy, while Persona 2 expresses being bored. The contradiction lies in the difference in emotional experiences and preferences.
`After refinement:`
`A's Persona 1:` I feel relaxed when I go fishing.
`A's Persona 2:` I feel tired because I spend a lot of time at work.

Table 9: More examples of persona refinement done by CAFFEINE. Color magenta highlights the contextual cues from the relevant dialogue contexts.

Example 1 - [Preservation]
Before refinement:
B's Persona 1: I love punk music.
B's Persona 2: I enjoy romantic comedies and would like to watch some cop shows.
Contextual Backgrounds (relevant dialogue contexts):
A: I am great. Residing in Arizona, skydiving is awesome, Barnes & Noble is my favorite, you?
B: I like movies over books, love punk music and just love to cruise around!
- - - - -
A: I read the Tori Spelling novel you told me about!
B: Awesome! What did you think of it?
A: I thought it was really good. Thanks for recommending it. You also said you watch a lot of movies. Can you recommend a good movie to me too?
B: Well, I've been watching a lot of romantic comedies lately. We should watch a couple the next time we have a movie night! I also wanted to ask, are there any cop shows you'd recommend? I'd like to try watching them too.
CAFFEINE's output:
Explanation: The two persona sentences do not contradict each other as they pertain to different areas of interest. Persona 1 expresses a love for punk music, while Persona 2 expresses a liking for romantic comedies and an interest in watching cop shows. There is no inherent conflict between loving punk music and enjoying romantic comedies or cop shows, so the two persona sentences can coexist without the need for resolution or disambiguation.
Example 2 - [Preservation]
Before refinement:
B's Persona 1: I got a new car.
B's Persona 2: I like to drink water.
Contextual Backgrounds (relevant dialogue contexts):
A: How was the drive up to San Diego? Did you bring your kids or are you flying solo? I hear the beaches are beautiful there.
B: The drive was long, but the scenery was gorgeous! Took a lot longer to make the trip up than I expected because with the kids I had to make more stops to stretch our legs, feed them, and give them a little time to burn off some energy. I think I always took for granted the luxury of traveling sans kids in the past, but honestly I don't mind the company! It was the first time traveling with my kids in my new car. How's it going for you?
- - - - -
A: I have not been to San Diego, but I have always wanted to go. You will have to tell me all about it.
B: I will. I will be on the lookout for the best iced tea! Haha. A: You know it is one of my favorites to drink. Besides tea I only really drink water
B: I love water, too. I wish I could get my kids to drink more water.
CAFFEINE's output:
Explanation: The two personas do not contradict each other as they pertain to different aspects of the speaker's life. Persona 1 is about getting a new car and taking a trip to San Diego, while Persona 2 is about the speaker's preference for drinking water. The two persona sentences can coexist without the need for resolution or disambiguation.

Table 10: Examples of "preservation" addressing the sub-optimal performance of NLI models that solely rely on the persona sentences without contextual backgrounds. Color magenta highlights the contextual cues.