PK-ICR: Persona-Knowledge Interactive Multi-Context Retrieval for Grounded Dialogue

Minsik Oh* **Joosung Lee** **Jiwei Li** **Guoyin Wang**

Abstract

Identifying relevant persona or knowledge for conversational systems is critical to grounded dialogue response generation. However, each grounding has been mostly researched in isolation with more practical multi-context dialogue tasks introduced in recent works. We define Persona and Knowledge Dual Context Identification as the task to identify persona and knowledge jointly for a given dialogue, which could be of elevated importance in complex multi-context dialogue settings. We develop a novel grounding retrieval method that utilizes all contexts of dialogue simultaneously. Our method requires less computational power via utilizing neural QA retrieval models. We further introduce our novel null-positive rank test which measures ranking performance on semantically dissimilar samples (i.e. hard negatives) in relation to data augmentation.

1 Introduction

Effective conversation agents require external context as grounding information to enhance response generation. There has been much progress on each persona (Majumder et al., 2020; Joshi et al., 2017; Shuster et al., 2018; Wu et al., 2019) and knowledge (Li et al., 2022; Dinan et al., 2018; Zhao et al., 2020; Liu et al., 2021) grounded dialogue systems respectably. However, the combination of both and more unique contexts has not been studied, with limited interest in industry persona-based QA systems (Byron et al., 2017; Ky and Joshi, 2021).

Feng et al. (2020); Dinan et al. (2018); Moghe et al. (2018) have shown the importance of directly optimizing knowledge extraction in dialogue, while Zhang et al. (2018a); Gu et al. (2021); Liu et al. (2020) have shown the importance of directly optimizing for concrete persona. We further argue that in practical settings, it is more realistic to

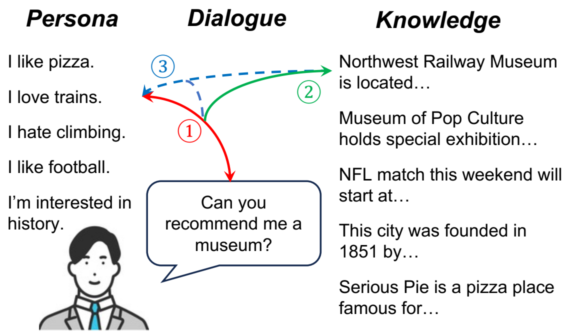

Figure 1: Illustration of dialogue component interactions regarding PK-ICR. In (1), dialogue is augmented with each persona to allow necessary interactions between persona and knowledge. (2) performs knowledge retrieval with (1). (3) performs precise persona scoring.

assume the utility of multiple contexts, with an explicit use-case being travel assistance agent (Jang et al., 2021).

Following the Knowledge Identification task in DIALKI (Wu et al., 2021), we define *Persona and Knowledge Dual Context Identification* as the task to identify persona and knowledge jointly for a given dialogue. The task is similar to persona-based QA task in industry (Byron et al., 2017; Ky and Joshi, 2021), of creating a search engine based on persona, with exception of our study being in an interactive dialogue setting. We emphasize the specific interactions (Fig. 1) between persona, knowledge, and dialogue. We aim to formalize the nature of component-wise interactions via this research, resulting in enhanced multi-context retrieval methodology.

This separation of grounding retrieval tasks could be a particular benefit for multi-context dialogue, in which we can study complex context-wise interactions first, then apply the identified behavior as a sub-component of end-to-end systems. As a starting point, we re-purpose existing tasks and find that Question Answering (QA) is a good candidate (Adolphs et al., 2021). This provides the

*Please correspond to minsik@mskresearch.org.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16383–16395
December 6-10, 2023

benefit of reduced computation and streamlined architecture by reusing powerful retrieval models previously developed for diverse tasks.

We develop a framework that exploits this relation, of which an interesting aspect is combining persona and dialogue[1] as a form of component augmentation. This may be of further utility in complex systems as each pertains to attributes and actions of the human respectively. Interestingly, our suggested augmentation method creates positive and hard negative samples which could be applied to enhance retrieval (Appendix G). We introduce a novel evaluation methodology of the *Null-positive Rank Test* (NRT) to quantify this trait.

Our contributions are summarized as follows.[2]

1. **Persona and knowledge dual context retrieval methodology.** We enhance specific interactions between all components to successfully retrieve dialogue contexts. We achieve SOTA performance for both persona and knowledge retrieval.

2. **Framework for cross-task adaptation of dialogue context interactions.** We introduce a framework to benefit from existing performant retrieval models for complex dialogue grounding retrieval. Our zero-shot inference allows reduced computation (Table C1) and streamlined architecture.

3. **Evaluating the hard-negative trait of Persona-augmented Dialogue.** We augment dialogue with persona to form an enhanced retrieval input, in which we observe hard negative traits. We introduce a novel test to isolate this trait, applicable to scenarios where semantically dissimilar samples are produced via data augmentation.

2 Related Works

"Knowledge-enhanced text generation" (Zhu et al., 2022; Yu et al., 2022) incorporates internal or external grounding contexts in tackling generative tasks such as dialogue or Q & A. Our research significantly contributes to the development of sophisticated "knowledge selection" for external knowledge systems. Our work is the first to model how to effectively select multiple distinct types of grounding contexts (persona & knowledge) for dialogue response generation.

To develop dialogue systems that rely on external knowledge information, open-domain dialogue datasets of Wizard of Wikipedia (Dinan et al., 2018) and PersonaChat (Zhang et al., 2018b) are most commonly studied. They consist of conversations that are grounded by Wikipedia knowledge and persona profile information, respectively. More recent datasets consist of conversations grounded by both persona & knowledge (Jang et al., 2021) or blended scenarios (Shuster et al., 2020; Smith et al., 2020). In line with the prior works, we treat persona and knowledge as different groundings with distinct characteristics and investigate semantic relations.

Integrating either persona or knowledge bases with dialogue agents in isolation has been actively studied. Zhang et al. (2018a); Majumder et al. (2020); Xu et al. (2020); Rashkin et al. (2019) for Persona, and Dinan et al. (2018); Zhao et al. (2020); Liu et al. (2021); Li et al. (2020); Ghazvininejad et al. (2017) for knowledge. Persona-only method prohibits elaboration with detailed knowledge. In contrast, relevant knowledge might depend on the persona of the user. We address the limitations by studying all dialogue component interactions.

Knowledge Identification (Wu et al., 2021) task has been defined in recent papers stemming from knowledge-grounded dialogue. Our work aligns with the view in Wu et al. (2021) that context identification is a separately important task in an interactive dialogue setting, with similarities to open question answering (Min et al., 2019; Chen et al., 2017) and industry persona-based QA systems (Byron et al., 2017; Ky and Joshi, 2021). Our research expands upon the Knowledge Identification task to specify persona & knowledge as dual contexts to be jointly retrieved from the dialogue.

3 Methodology

We maximize interactions between all components of a conversation turn for effective retrieval of dialogue groundings. Knowledge retrieval is a top-1 ranking task (Section 3.1), and persona retrieval is a point-wise scoring task with 1 or 0 true persona label (Section 3.2). We solve knowledge retrieval in a zero-shot manner, while we introduce *null-positive rank test* to investigate the hard-negative traits of Persona-augmented Dialogue (Section 3.3). [3]

3.1 Knowledge Retrieval

We introduce a novel adaptation of dialogue components as QA prompts (example in Fig. A1). This

[1] Persona-augmented Dialogue

[2] Code for our experiments is available : https://github.com/minsik-ai/PK-ICR

[3] We note that each sequence - knowledge and persona retrieval - may be further optimized independently.

form is selected to infer relations between all inputs of the grounded dialogue and to replicate short question and descriptive answer pairs.

$$E : \{Q_i, A_j\} = \{P_i + D, K_j\} \quad (1)$$

E is input to our model. Q_i, A_j, P_i, K_j are specific QA candidates and persona & knowledge pairs. D is the dialogue for the pairs.

We then find the best knowledge for all pairs of i and j in a permutative manner and record the knowledge of the most aligned pair.

$$\mathit{best}_i, \mathit{best}_j = \underset{i \in 1...n, j \in 1...m}{\arg\max} M_q\{P_i + D, K_j\} \quad (2)$$

$$\mathit{true}_j = \mathit{best}_j \quad (3)$$

best_i, best_j are indices from best-scoring persona / knowledge pair. true_j is the index of predicted knowledge K. best_i is discarded. M_q is QA retrieval model for pair likelihood score. n / m is persona / knowledge count respectively.

3.2 Persona Retrieval

Continuing from Section 3.1, we fine-tune the QA retrieval model using augmented persona and predicted true knowledge pairs only.[4] We report that Persona-augmented Dialogue exhibits hard negative attributes (Section 3.3).

$$E' : \{Q_i, A_{\mathit{true}}\} = \{P_i + D, K_{\mathit{true}_j}\} \quad (4)$$

$$M_q \xrightarrow{E'_{\mathit{train}}} M_f \quad (5)$$

E' is input to our model similar to E, only difference being fixed true knowledge. E'_{train} is data from a separate training set formulated in the same manner as E' with labeled true knowledge. M_f is the fine-tuned model (Appendix D).

Next, we infer selected data pairs with M_f to obtain the persona likelihood score. We avoid retrieving unrelated persona via a threshold.

$$p_i = M_f\{P_i + D, K_{\mathit{true}_j}\} \quad (6)$$

$$\mathit{true}_i = \arg\max_i \begin{cases} p_i, & \text{if } p_i \geq p_{\mathit{thres}} \\ 0, & \text{otherwise} \end{cases} \quad \text{for } i \in 1...n. \quad (7)$$

[4]This results in additional reduced computation of $O(nm)$ to $O(n)$ for both training and inference. In effect, this decreases negative pairs from 3M to 0.3M with 10x speedup.

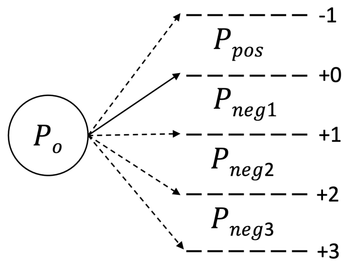

Figure 2: Null-positive Rank Test (NRT). P_o, P_{pos}, P_{neg_i} denote null-positive sample, positive and negative personas respectively. We omit augmentation $+D$ in the figure for brevity. $r_{min} = -1$ and $r_{max} = +3$ in this figure. Arrows are possible positions for P_o. Numbers on the rightmost side are the null-positive adjusted rank values, being 0 right below P_{pos} (example in Table A1).

p_i is the likelihood score for P_i. p_{thres} is the likelihood score threshold to remove persona that doesn't correspond to the dialogue turn. true_i is the index of the predicted true persona.

Finally, the retrieved grounding information is:

$$R : \{D, P, K\} = \{D, P_{\mathit{true}_i}, K_{\mathit{true}_j}\} \quad (8)$$

R is the retrieved true persona & knowledge pair for the given dialogue turn.

3.3 Null-positive Rank Test

We stress that fine-tuning model M_q with Persona-augmented Dialogue ($P_i + D$) to create model M_f is a specific choice. This is because the QA setup cannot be utilized without adjustments, due to the model scoring output skewing higher (Fig. E1). To analyze without inflated scores, we first interpret Persona-augmented Dialogue as hard negative sampling (Appendix G), in which the augmentation produces non-trivially hard-to-distinguish samples.[5]

To evaluate the above observation, we present a novel methodology of **null-positive rank test** to quantify the inherent difficulty of ranking $P_i + D$ samples. Inspired by ranking metrics such as MRR, MAP, and NDCG, we utilize rank of a specific sample to compute model performance. This allows us to isolate the discriminative performance of the model corresponding to samples of interest, regardless of score output (Fig. 2, example in Table A1).

We designate null-positive[6] (P_o) sample as a baseline for the model. We measure the following:

[5]Note that positive samples are also created (Table A1). This augmentation is compatible with persona-only tasks.

[6]"Null-positive" term corresponds to the fact that the ideal model should have no preference on how to score the likeli-

Model Type	Accuracy (%)
Baseline	65.06
BERT-base	11.78
Proto-gen (Saha et al., 2022)	85.18
D & K_j	79.26
P_i & K_j	84.62
$P_i + D$ & K_j	**94.69 (+29.63)**

Table 1: Knowledge retrieval accuracy (cross-encoder) per asymmetric QA prompt. Zero-shot. Similar results for bi-encoder, which compares vectors (Table H1).

Can the model rank null-positive sample correctly in relation to non-trivially dissimilar augmented samples? The "non-triviality" metric which computes the average distance of null-positive sample's rank from the ideal rank[7] is as follows:

$$\neg T = \frac{\sum_{r=r_{min}}^{r_{max}} n_r * |r|}{\sum_{r=r_{min}}^{r_{max}} n_r} \tag{9}$$

Variants of the metric are in Appendix B. $\neg T$ is non-triviality metric, with lower values of the metric meaning the model ranks better. n_r is the number of P_o samples with adjusted rank r (Fig. 2). We report "non-triviality" for each model M_q, M_f.

4 Experiment Setup

We utilize the Call For Customized Conversation (Jang et al., 2021) dataset for which each conversation is built with both the user's persona and Wikipedia knowledge grounding information. We utilize multiple neural QA models trained on MS MARCO dataset (Nguyen et al., 2016). More details in Appendix D.

5 Results

5.1 Knowledge Retrieval

We experiment with ablations of dialogue / persona / knowledge interactions and find permutative evaluation of eq. 1 form yields best performance. Table 1 shows strong performance increase for our prompt input from dialogue-only model, confirming that all components of dialogue is important.[8]

hood of the null-positive sample, except that it should rank right below all positive sample(s). Another name considered was neutral rank test. In a real scenario, there could be multiple positive samples. Naturally, our metric weighs short rank distances less.

[7]Right below positive sample, above all negative samples.

[8]We verify that our cross-task adaptation of Q&A is significantly stronger than NLI, STS, and NSP (Table H1).

Model Type	Accuracy (%)
Baseline	86.86
BERT-base	71.82
Proto-gen (Saha et al., 2022)	87.75
D & P_i	86.78
P_i & K_{true_j}	86.75
$P_i + D$ & K_{true_j}	83.83
P_i & K_{true_j} (fine-tuned)	89.12
$P_i + D$ & K_{true_j} (fine-tuned)	**91.57 (+4.71)**

Table 2: Persona retrieval accuracy (cross-encoder) per asymmetric QA prompt. Zero-shot unless fine-tuned. Similar results for bi-encoder (Table H2).

Type	0-Acc (%)	$\neg T^2$	$\neg T$	$\neg T_+$	$\neg T_-$
Z.S.	79.30	1.84	1.02	1.04	0.62
Ours	**86.81**	**1.66**	**0.97**	**0.96**	**0.56**

Table 3: Null-positive rank test results for $P_i + D$ & K_{true_j} cross-encoder models. Ours model is the fine-tuned variant, and Z.S. is Zero-Shot model. We report persona retrieval accuracy when $p_{thres} = 0$ (0-Acc) and variants of non-triviality (eq. 9, 12, 10, 11). Smaller non-triviality means superior ranking capability. Similar results for bi-encoder (Table F1).

5.2 Persona Retrieval

Table 2 shows that fine-tuned $P_i + D$ model has the best performance. However, we observe low performance for the non-fine-tuned $P_i + D$ model. This is due to QA relationship of dialogue to true knowledge affecting the likelihood score (Fig. E1). Thus fine-tuning the model is a necessity to harness cross-domain adaptation.

5.3 Null-positive Rank Test

To verify our observation of the effectiveness of $P_i + D$, we perform null-positive rank test (Section 3.3). The performance of the model has increased in top-1 rank setting (0 threshold, 0-Acc)[9] and all variants of non-triviality have improved for both models (Table 3, F1). We analyze sample count per rank (Fig. 3, F1).

6 Discussions and Conclusion

We introduce persona-knowledge dual context retrieval method PK-ICR in this paper. We perform QA-informed prompt-augmentations of data that

[9]This is performance on persona retrieval free from score-related effect (Fig. E1).

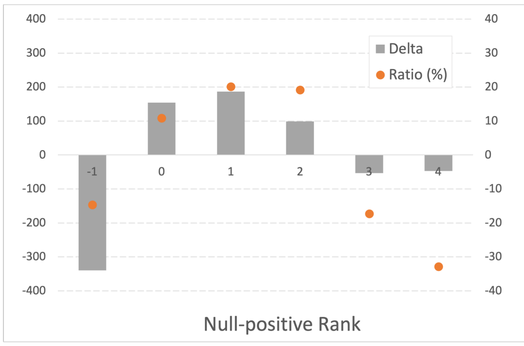

Figure 3: Analysis of null-positive rank data for $P_i + D$ & K_{true_j} cross-encoder model. Delta value is the change between Zero-shot model and Ours model in terms of sample count (left axis). Ratio value is delta value divided by sample count for Zero-shot, in % (right axis). We report movements of delta in correct directions for rank -1, 0 and ranks with long distance 3, 4. Similar results for bi-encoder (Fig. F1).

successfully exploit the interactions between multiple dialogue components. We perform zero-shot top-1 knowledge retrieval and precise persona scoring. We present a novel evaluation method of null-positive rank test as to isolate the hard-negative effect of Persona-augmented Dialogue. We obtain SOTA results on both retrieval tasks of the Call For Customized Conversation benchmark and report the alignment of the non-triviality metric with threshold-free performance. With our research, we hope to stimulate readers to model dialogue context as an interactive whole of multiple components.

As the NLP community aims to tackle more complex dialogue systems, our methods may be further enhanced by sophisticated grounding contexts and interactions present in dialogue. Considering all components of dialogue, being persona, knowledge, and dialogue in a travel agent scenario, is crucial to obtain each grounding context required for accurate responses. We suggest two future directions for dialogue systems.

1. One possible future direction is incorporating different forms of grounding such as persona/knowledge summaries, web searches, Wikipedia documents, and extracted sentences (Wu et al., 2021) for multi-context interactions.

2. Another would be extending our methodology to more sophisticated dialogue settings such as long-term memory (Bae et al., 2022) or mutual persona (Liu et al., 2020).

Our persona-aware dialogue augmentation (① in Figure 1) is a form of modeling the human behavior in regards to Turing test (Li et al., 2016; Vinyals and Le, 2015). We isolate effectiveness of our augmentation with successful Null-Positive Rank Test (Section 3.3). We suggest two future directions regarding persona-dialogue augmentation.

1. One direction to pursue would be emphasizing different interactions, such as augmenting persona and knowledge first, or complex forms such as providing negative context in the augmentation themselves. Whether human-like behavioral modeling works better, or computationally tractable methods could be developed, is another interesting question.

2. Another direction to pursue would be advanced prompting, where specifying persona or knowledge via pre/post-fix prompts (i.e. "I'm thinking 'I like football'") provides more indicator information regarding each grounding and may improve performance.

Finally, our novel Null-positive Rank Test (NRT, Section 3.3) is widely applicable to information retrieval and ranking models. We only trained the models with positive and negative dialogue augmentations, but interestingly, we report that the model improves ranking correctness for "neutral" (non-augmented) dialogue. This evaluation method could be directly used to compute whether the model obtains sophisticated ranking capabilities. This would be especially relevant in the context of data augmentation or other ranking tasks, including personalized search / QA. We recommend performing our test for future works.

7 Limitations

Our cross-task adaptation of dialogue grounding retrieval to QA task is limited in terms of the target task and our prompt construction. In addition, retrieval models informed by inductive bias for multi-context scenarios could further improve our methodology.

We specifically study multi-context interactions and retrieval in dialogues, which is a relevant and novel problem for advancing broadly capable dialogue systems. As an extension to our research, future work could also report on modeling downstream generation tasks based on grounding interactions.

Acknowledgement

We sincerely thank the ARR reviewers for their valuable insights.

References

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021. Reason first, then respond: Modular generation for knowledge-infused dialogue.

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.

Donna K. Byron, Alexander Pikovsky, David G. Radley, and Timothy P. Winkler. 2017. Answering questions via a persona-based natural language processing (nlp) system.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, William B. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *AAAI*, abs/1702.01932.

Jia-Chen Gu, Zhenhua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1126–1136, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proc. of SIGIR*.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2021. Call for customized conversation: Customized conversation grounding persona and knowledge.

Chaitanya K. Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Piseth Ky and Anila Joshi. 2021. [link].

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model.

Lin-Xiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *NeurIPS*, abs/2008.12918.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021a. Pretrained transformers for text ranking: Bert and beyond.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, Online. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *Conference on Empirical Methods in Natural Language Processing*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval.

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.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. *arXiv preprint arXiv:1809.08205*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Sougata Saha, Souvik Das, and Rohini Srihari. 2022. Proto-gen: An end-to-end neural generator for persona and knowledge grounded response generation. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 9–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Bowen Wu, Mengyuan Li, Zongsheng Wang, Yifu Chen, Derek Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2019. Guiding variational response generator to exploit persona.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval.

Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.*, 54(11s).

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.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. *ICLR*, abs/2002.10348.

Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Lin, Meng Jiang, and Wenhao Yu. 2022. Knowledge-augmented methods for natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 12–20.

A Samples

A.1 QA Cross-Task Adaptation Prompt Construction

Question : "{I want to visit Seven Wonders of the Ancient World.} {Wow, what is this?}" Answer : "{The Great Pyramid of Giza ... of the Seven Wonders of the Ancient World, ...}"

Figure A1: Resulting QA cross-task adaptation prompt of persona & knowledge pair (eq. 1). Question form is "{persona} {dialogue}" while answer is "{knowledge}".

A.2 Persona-augmented Dialogue

Persona-augmented Dialogue	Notation	Adj. Rank
I like mountains, where to go for a hike?	$P_{pos} + D$	-1
where to go for a hike?	$P_o = D$	0
I like rock music, where to go for a hike?	$P_{neg1} + D$	+1
I don't like pizza, where to go for a hike?	$P_{neg2} + D$	+2
I don't like scary movies, where to go for a hike?	$P_{neg3} + D$	+3

Table A1: We display ideal rank order for Persona-augmented dialogue ($P_i + D$) along with null-positive sample P_o (underlined). The rank is adjusted to be 0 for the ideal null-positive rank. This table corresponds to notations in Fig. 2.

B Null-Positive Rank Test Variants

We introduce variants of non-triviality $\neg T$ metric (eq. 9). Smaller numbers are better for all variants.

- $\neg T_+$ to only observe positive rank displacements.

$$\neg T_+ = \frac{\sum_{r=0}^{r_{max}} n_r * |r|}{\sum_{r=0}^{r_{max}} n_r} \tag{10}$$

- $\neg T_-$ to only observe negative rank displacements.

$$\neg T_- = \frac{\sum_{r=r_{min}}^{0} n_r * |r|}{\sum_{r=r_{min}}^{0} n_r} \tag{11}$$

- $\neg T^2$ similar to how Mean Squared Error relates to Mean Absolute Error.

$$\neg T^2 = \frac{\sum_{r=r_{min}}^{r_{max}} n_r * r^2}{\sum_{r=r_{min}}^{r_{max}} n_r} \tag{12}$$

- $\neg T_{weighted}$ to provide constant weights for each rank.

$$\neg T_{weighted} = \frac{\sum_{r=r_{min}}^{r_{max}} w_r * n_r * |r|}{\sum_{r=r_{min}}^{r_{max}} w_r * n_r} \tag{13}$$

C Computational Efficiency

Saha et al. (2022) utilizes 2 BART models for input and persona & knowledge groundings.[10] In contrast, PK-ICR utilizes 1 MiniLM (Wang et al., 2020) cross-encoder retrieval model for computing similarity scores. Saha et al. (2022) concatenates all groundings as one input, while PK-ICR groups them into pairs. Our zero-shot knowledge retrieval (Section 3.1) doesn't require any training, while Saha et al. (2022) trains for a maximum of 15 epochs.

Metrics	Saha et al. (2022)	Ours
Train Samples	1.1M	633K (Section 3.2)
Model Params	210M	33M

Table C1: Computational effort required for the methods.

D Experiment Setup

We utilize Call For Customized Conversation (Jang et al., 2021) dataset for evaluation and fine-tuning, which has 10 knowledge candidates and 5 persona candidates per dialogue. We utilize 12 layer MiniLM (Wang et al., 2020) (33M params) cross-encoder trained on MS MARCO[11] (Nguyen et al., 2016) from Sentence-BERT library (Reimers and Gurevych, 2019) and DistillBERT (66M params) TAS-B (Hofstätter et al., 2021) bi-encoder model trained on same data[12]. This data is for semantic search, with trained models evaluating short questions and long passages together. In addition, we report baseline performance on DPR (Karpukhin et al., 2020) models[13] (110M params) trained on NQ (Kwiatkowski et al., 2019) dataset[14], with a dummy short segment of "Title", and treating Knowledge as long answer segment. For persona search (eq. 5, 7), we fine-tune for 2 epochs, 32 batch size, and sigmoid activation function with Binary Cross Entropy (cross-encoder) / Cosine Similarity (bi-encoder) Loss with $p_{thres} = 0.5$. We list the official evaluation results on the test data. For MobileBERT (25M params) and BERT-base (110M params), we evaluate with Next Sentence Prediction task. We experiment with DistillRoBERTa (82M params) STS[15] and NLI[16] cross-encoder models. We work with RTX 3090 NVIDIA GPU.

[10] 2 BART encoders and 1 decoder for generation (may positively affect retrieval).
[11] MRR@10 on MS MARCO Dev Set: 39.02
[12] MRR@10 on MS MARCO Dev Set: 34.43
[13] Bi-encoder model with separate question and answer encoders.
[14] NQ test set Accuracy@20: 78.4, Accuracy@100: 85.4
[15] STSbenchmark test performance: 87.92
[16] Accuracy on MNLI mismatched set: 83.98, we compare with *entailment* score.

E Persona Retrieval Threshold Experiments

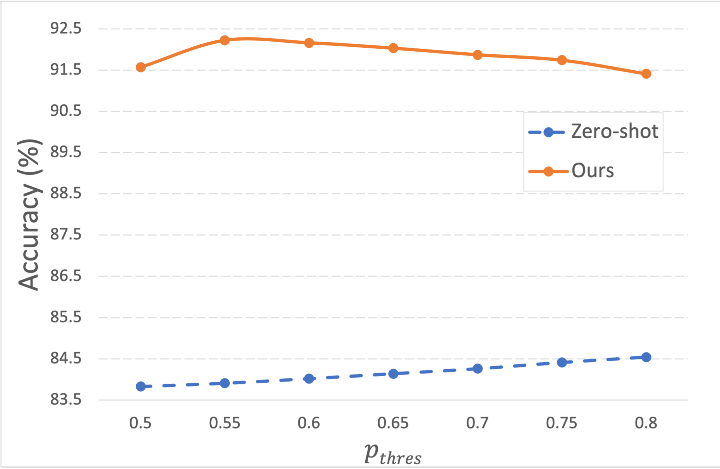

Figure E1: Persona threshold ablation experiments with P_i & K_{true_j} cross-encoder model. We report persona accuracy. p_{thres} is defined in eq. 7. Dotted line correspond to Zero-shot model, and solid line is our best model. We find visible peak at 0.55 with our best model while Zero-shot model performance keeps increasing > 0.8.

F Null-Positive Rank Test for Bi-Encoder

Type	0-Acc (%)	$\neg T^2$	$\neg T$	$\neg T_+$	$\neg T_-$
Z.S.	77.90	4.27	1.59	1.79	0.55
Ours	**85.53**	**1.95**	**0.99**	**0.98**	**0.50**

Table F1: Null-positive rank test results for $P_i + D$ & K_{true_j} bi-encoder models. Ours model is the fine-tuned variant, and Z.S. is Zero-Shot model. We report persona retrieval accuracy when $p_{thres} = 0$ (0-Acc) and variants of non-triviality (eq. 9, 12, 10, 11). Smaller non-triviality means superior ranking capability.

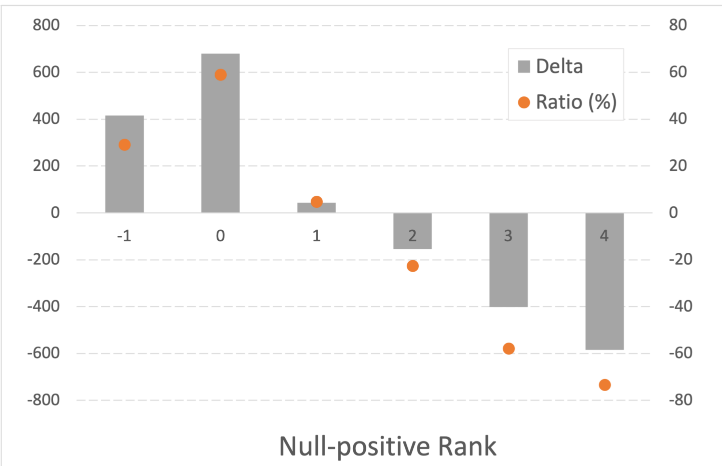

Figure F1: Analysis of null-positive rank data for $P_i + D$ & K_{true_j} bi-encoder model. Delta value is the change between Zero-shot model and Ours model in terms of sample count (left axis). Ratio value is delta value divided by sample count for Zero-shot, in % (right axis). We report large movements of delta in correct directions for rank 0 and ranks with long distance 3, 4.

G Background: Ranking and Hard Negative Sampling

Text ranking is a task to generate an ordered list of texts in response to a query (Lin et al., 2021a). It is a core task of information retrieval (IR) where you obtain a ranked list of samples ordered by estimated relevance to the query. We introduce widely accepted neural approaches, 'cross-encoder' and 'bi-encoder'. We will also describe 'hard negative sampling', a data-centric approach to improve retrieval models.

For cross-encoder (Nogueira et al., 2019), query and a single sample are concatenated by '[SEP]' token as an input to the model, resulting in a relevance score (FFN output of the '[CLS]' token representation) for the specific sample. We note that this setup is similar to sentence-wise classification settings presented in Devlin et al. (2018). Then, the samples are ordered by relevance score to produce the final ranked list.

For bi-encoder (Reimers and Gurevych, 2019), we generate dense vectors (sentence embeddings) per each query and each sample. This is obtained via '[CLS]' token representation of a specially fine-tuned model[17], with a single query or sample input. The representations are then compared as pairs via cosine-similarity or dot-product to compute relevance scores. While original bi-encoder setup computes sentence-wise similarity (STS), we utilize models fine-tuned on QA data (Appendix D).

Hard negative sampling (also known as hard negative mining) is a technique to obtain specific samples (hard negatives) that are difficult to distinguish from positive samples, yet have a different label. The hard negative samples are then incorporated during model fine-tuning to improve model capabilities. For example, in the context of ranking, non-relevant texts scoring high by how many keywords match (Xiong et al., 2020) may be considered hard negatives. Xiong et al. (2020); Luan et al. (2021); Lin et al. (2021b) have demonstrated that hard negative sampling improves ranking models considerably.

H Detailed Results

More experiments are listed here. Our bi-encoder and cross-task experiments confirm our findings in Section 5. Explanation of the models in Appendix G.

Model Type	Accuracy (%)
Baseline	65.06
MobileBERT	9.49
BERT-base	11.78
Proto-gen (Saha et al., 2022)	85.18
NLI (cross-encoder)	17.96
STS (cross-encoder)	51.33
D & K_j (cross-encoder)	79.26
P_i & K_j (DPR)	75.54
P_i & K_j (bi-encoder)	80.73
P_i & K_j (cross-encoder)	84.62
$P_i + D$ & K_j (DPR)	83.98
$P_i + D$ & K_j (bi-encoder)	92.67
$P_i + D$ & K_j (cross-encoder)	**94.69**

Table H1: Knowledge retrieval results, all models are zero-shot. We report top-1 knowledge retrieval accuracy per asymmetric QA prompt. D, K, P each refer to dialogue, knowledge and persona.

[17] Supervised learning via Siamese Network.

Model Type	Accuracy (%)
Baseline	86.86
MobileBERT	86.86
BERT-base	71.82
Proto-gen (Saha et al., 2022)	87.75
D & P_i (cross-encoder)	86.78
P_i & K_{true_j} (DPR)	75.54
P_i & K_{true_j} (bi-encoder)	78.64
P_i & K_{true_j} (cross-encoder)	86.75
P_i & K_{true_j} (cross-encoder, fine-tuned)	89.12
$P_i + D$ & K_{true_j} (DPR)	74.76
$P_i + D$ & K_{true_j} (bi-encoder)	77.90
$P_i + D$ & K_{true_j} (cross-encoder)	83.83
$P_i + D$ & K_{true_j} (bi-encoder, fine-tuned)	85.55
$P_i + D$ & K_{true_j} (cross-encoder, fine-tuned)	**91.57**

Table H2: Persona retrieval results, models are zero-shot unless fine-tuned. We report persona retrieval accuracy per asymmetric QA prompt. We do not fine-tune DPR model due to implementation limitations. D, K, P each refer to dialogue, knowledge and persona.

I Retrieval Output Samples

We list some of the retrieved outputs with our best model in Table I1.

dialogue D	persona P_{true}	knowledge K_{true}
I think I've been there before but I don't remember the name of this place.	I am fond of Modernist architechure.	The Casa de les Punxes or Casa Terradas is a building designed by the Modernista architect Josep Puig I Cadafalch. Located in the intersection between the streets of Rosselló, Bruc and the Avinguda Diagonal in the Barcelona Eixample area.
How much this railway line costed in those times?	I love railway.	Because of the difficult physical conditions of the route and state of technology, the construction was renowned as an international engineering achievement, one that cost US$8 million and the lives of an estimated 5,000 to 10,000 workers.
Who built this rail line?	I love railway.	The line was built by the United States and the principal incentive was the vast increase in passenger and freight traffic from eastern USA to California following the 1849 California Gold Rush.
What's the highest point in the Mulanje Massif?	I like to climbing up the elevations on my neighborhood to take a look around.	Sapitwa Peak, the highest point on the massif at 3,002 m, is the highest point in Malawi.
Who was the first explorer to find this mountain?	I have fantasies of being a Livingstone type explorer.	The first European to report seeing the Massif was David Livingstone in 1859, but archeological investigation reveals evidence of human visits to the Massif from the Stone Age onwards.
Now I remember, can you tell me some characteristics of this channel?	N / A	And may be the oldest canal in England that is still in use. It is usually thought to have been built around AD 120 by the Romans, but there is no consensus among authors.

Table I1: persona, knowledge and dialogue retrieved examples from our best model.