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

Anonymous ACL submission

Abstract

Dialogue context identification is a critical component of grounded dialogue response generation, with Persona and Knowledge being major areas of study. However, each context has been studied in isolation with more practical multi-context tasks only recently introduced. We define Persona and Knowledge Dual Context Identification as the task to identify Persona and Knowledge jointly for a given dialogue, which would be of fundamental importance in complex multi-context Dialogue settings. We develop a novel multi-context retrieval method that utilizes all contexts of dialogue simultaneously while also requiring limited training via zero-shot inference due to compatibility with neural Q & A models. Techniques we develop for enhanced retrieval are *cross-domain formulation, component augmentations, permutative evaluation, and selective fine-tuning*. We further analyze the hard-negative behavior of combining Persona and Dialogue via our novel null-positive rank test. We achieve SOTA with 90%+ performance for both tasks of Persona retrieval & Knowledge retrieval on the Call For Customized Conversation benchmark. We provide code for our training, retrieval, and test in zip file with submission.

1 Introduction

Dialogue context identification (Wu et al., 2021; Feng et al., 2020) is a crucial component of grounded dialogue generation. There has been much progress on each Persona and Knowledge grounded dialogue systems respectably, however combination of both and more unique contexts has not been extensively studied. In practical settings, it is more realistic to assume utility of multiple components, with an explicit use-case being travel assistance agent (Jang et al., 2021).

One important aspect of multi-context configuration is that Persona and Knowledge pairs should be retrieved from given Dialogue. Following Grounding prediction tasks in (Jang et al., 2021), we define *Persona and Knowledge Dual Context Identification* as the task to identify Persona and Knowledge jointly for a given dialogue. We emphasize that there are specific interactions (Figure 1) that happen between Persona, Knowledge and Dialogue, thus they cannot be predicted separately from partial components.

We aim to formalize the nature of component-wise interactions via this research, resulting in enhanced Persona and Knowledge dual context retrieval methodology that enhances the downstream task of dialogue generation. This separation of tasks is of a particularly fundamental 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 domains and find that Question and Answering is a good candidate (Adolphs et al., 2021).

We develop a framework that exploits this relation, of which a particularly interesting method is combining Persona and Dialogue[1] as a form of component augmentation. Combining Persona and Dialogue as a single element in complex systems might be of further utility as each pertains to attributes and actions of the human respectively. Interestingly, our suggested combination seems to induce non-triviality that corresponds to hard negatives that could be applied to enhance retrieval. We introduce a novel evaluation methodology of the *Null-positive Rank Test* (NRT) to quantify this trait. We further enhance the effect of augmentation via selective fine-tuning and permutative evaluation.

Our contributions are summarized as follows.

1. **Persona and Knowledge dual context retrieval methodology.** We enhance specific interactions between all components to successfully retrieve dialogue components. We achieve SOTA performance for both Persona and Knowledge retrieval. Notably, no model fine-tuning is required

[1] Persona-Dialogue Augmentation

for zero-shot top-1 Knowledge retrieval method.

2. **Framework to enhance cross-domain formulation of dialogue context interactions.** (Adolphs et al., 2021) showed that Q & A formulation is relevant for Knowledge grounded dialogue in E2E setting. We develop a cross-domain retrieval framework for multi-context dialogue that can repurpose existing retrieval models.

3. **Quantifying non-triviality induced via Persona-Dialogue Augmentation.** We augment Dialogue with Persona to form an enhanced input to our retrieval method, in which we observe hard negative traits. We present a novel test methodology to isolate the capabilities of models on induced non-triviality.

2 Related Works

Integrating Persona or Knowledge bases with dialogue agents in isolation have been actively studied. For Persona integration, datasets and systems include PersonaChat (Zhang et al., 2018) and many others (Majumder et al., 2020; Joshi et al., 2017; Shuster et al., 2018; Wu et al., 2019; Xu et al., 2020; Rashkin et al., 2019). Datasets for Knowledge integration are (Dinan et al., 2018; Zhou et al., 2018). Persona-only method is limited in that lack of Knowledge context prohibits the agent from elaborating with specific detailed information. In contrast, the shortcoming of the Knowledge-only approach is that relevant Knowledge itself might depend on Persona of the user. We address the limitations of previous studies via studying interactions between all components of dialogue.

Knowledge Identification (Wu et al., 2021; Feng et al., 2020) task has been defined and studied in recent papers stemming from evaluation of Knowledge grounded dialogue. However, it does not take user's Persona into account and mostly focuses on span-prediction from long-form passages. Our research expands upon Knowledge Identification task to specify Persona & Knowledge as dual context to be jointly retrieved from dialogue.

3 Methodology

We provide a brief overview of methodology in Appendix A. A glossary for the equations is provided in Appendix B.

3.1 Knowledge Retrieval

We introduce a novel formulation of Persona, Knowledge and Dialogue as Q & A input (Figure 5). This form is specifically selected to infer relations between all inputs of the grounded dialogue during answer likelihood calculation and to replicate short question and descriptive answer pairs often found in the Q & A setting. Sample available in Appendix F.

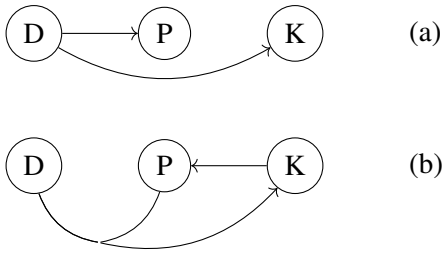

Figure 1: Comparison of baseline retrieval method with PK-ICR. In baseline system (a), Persona and Knowledge are retrieved separately via Dialogue-only input. PK-ICR (b) introduces an intermediate step to encode Dialogue together with Persona. This allows for necessary interaction between Persona and Knowledge.

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

We then perform permutative Persona-Knowledge evaluation (Figure 4) on all pairs of augmented Persona and Knowledge E. We find the best Knowledge via computing all pairs and recording Knowledge of most aligned pair. This is to make sure we find the best Knowledge that aligns with the Dialogue and Persona of the human.

$$true_j = \arg\max_j M_q\{P_i + D, K_j\} \quad \text{for } i \in 1...n, j \in 1...m. \quad (2)$$

3.2 Persona Retrieval

Continuing from Section 3.1, we fine-tune the Q & A retrieval model M_q using augmented Persona and predicted true Knowledge K_{true_j} pairs only, without incorrect Knowledge pairs.[2] We further remark that Persona-augmented Dialogue exhibits Hard Negative characteristics, with detailed discussion in Section 3.3.

$$E' : \{Q_i, A_{true}\} = \{P_i + D, K_{true_j}\} \quad (3)$$

$$M_q \xrightarrow{E'_{train}} M_f \quad (4)$$

[2]This results in 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.

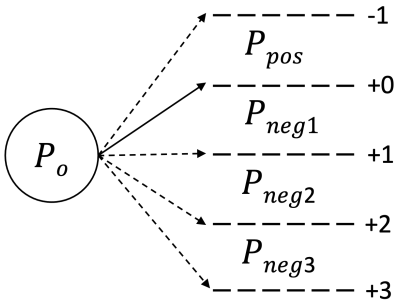

Figure 2: A sample ranking procedure during Null-positive Rank Test (NRT). P_o, P_{pos}, P_{neg_i} denote null-positive sample, positive and negative Personas respectively. We omit dialogue augmentation $+D$ in the figure for brevity. $r_{min} = -1$ and $r_{max} = +3$ in this figure. Note that the likelihood for Persona is ordered from top to bottom. Arrows are possible positions for P_o. Numbers on the right side of Personas are the null-positive rank values, which are configured to be 0 when right below P_{pos}. Corresponding samples in Table 4.

Finally, we infer E' data pairs with model M_f to obtain Persona likelihood score. We utilize a threshold p_{thres} to avoid retrieving unrelated Persona. Certain Dialogue has no Persona assigned to it, which we can replicate with the threshold.

$$p_i = M_f\{P_i + D, K_{true_j}\} \tag{5}$$

$$true_i = \arg\max_i \begin{cases} p_i, & \text{if } p_i \geq p_{thres} \\ 0, & \text{otherwise} \end{cases} \quad \text{for } i \in 1...n. \tag{6}$$

Retrieved Persona and Knowledge for given Dialogue D is as follows, notated by R :

$$R : \{D, P, K\} = \{D, P_{true_i}, K_{true_j}\} \tag{7}$$

3.3 Null-positive Rank Test

We stress that fine-tuning our model with Persona-augmented Dialogue ($P_i + D$) to create model M_f is a specific choice, closely associated with our retrieval setup with Q & A formulation. Appendix G discusses how score will be skewed higher without fine-tuning. Furthermore, we interpret Persona-augmented Dialogue as a form of Hard Negative sampling, in which augmentation of Persona with Dialogue creates non-trivial Questions that require enhanced model capability.[3] We discuss hard negative observation in detail in Appendix E.

[3]We note that this augmentation can also be utilized with Persona-only tasks.

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

Table 1: Knowledge retrieval results. We report top-1 Knowledge retrieval accuracy per asymmetric Q & A input. D, K, P each refer to Dialogue, Knowledge and Persona. Pairwise means all possible permutations are ranked (Figure 4).

As to strengthen hard negative observation, we present a novel methodology of **null-positive rank test** to quantify the inherent difficulty of ranking $P_i + D$ samples. We designate a null-positive[4] (P_o) sample as a representative baseline for the model. The inquiry is the following: *Where does the null-positive sample rank when contrasted with nontrivial candidates?* This method allows us to isolate the discriminative performance of the model corresponding to samples of interest, regardless of score output.

The "non-triviality" metric which computes average distance of null-positive (P_o) rank from ideal rank of 0 is as follows, notated by $\neg T$:

$$\neg T = \frac{\sum_{r=r_{min}}^{r_{max}} n_r * |r|}{\sum_{r=r_{min}}^{r_{max}} n_r} \tag{8}$$

A lower value of $\neg T$ means the model can distinguish hard negative samples better. Glossary for symbols are in Appendix B. We provide variants of the metric and detailed descriptions regarding null-positive selection and ideal rank in Appendix C.

4 Experiment Setup

We utilize Call For Customized Conversation (Jang et al., 2021) dataset and cross-encoder neural Q & A model trained on MS MARCO dataset (Nguyen et al., 2016). More details in Appendix D.

5 Results

5.1 Knowledge Retrieval

We experiment with various ablations of Dialogue / Persona / Knowledge interactions and find per-

[4]"Null-positive" term corresponds to the fact that the ideal model should have no preference on how to score the likelihood of the null-positive sample, except that it should rank right below all positive sample(s). Another name considered was neutral rank test.

Model Type	Accuracy (%)
Baseline	86.86
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 results. We report Persona retrieval accuracy per asymmetric Q & A input. D, K, P each refer to Dialogue, Knowledge and Persona. Available candidates of P_i are compared with fixed K_{true} (Figure 4).

Model Type	0-Acc (%)	$\neg T$	$\neg T_+$	$\neg T_-$
Zero-shot	79.30	1.02	1.04	0.62
Ours	**86.81**	**0.97**	**0.96**	**0.56**

Table 3: Null-positive rank test results for $P_i + D$ & K_{true_j} models. Ours model is the fine-tuned variant. We report Persona retrieval accuracy when $p_{thres} = 0$ (0-Acc), overall / positive / negative non-triviality ($\neg T$, $\neg T_+$, $\neg T_-$) (eq. 8). Smaller non-triviality means the model ranks the set of augmented Persona $P_i + D$ easier.

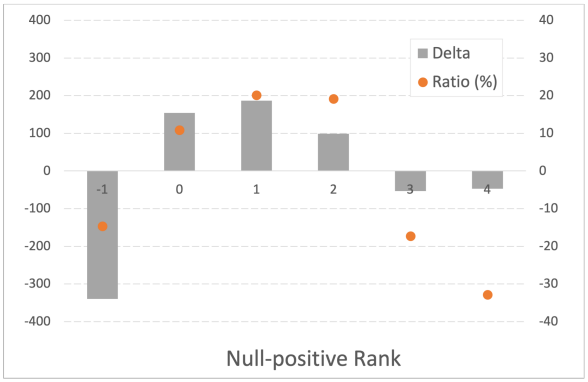

Figure 3: Analysis of null-positive rank data for $P_i + D$ & K_{true_j} models. Delta value refer to change between Ours model and Zero-shot model in terms of sample count (left axis). Ratio value is delta value divided by sample count for Zero-shot, in % (right axis). For rank 0, delta > 0 which is an improvement. Additionally, our non-triviality results capture improvements in crucial rank positions, being delta < 0 observed for rank -1 and ranks with long distance 3, 4.

mutative evaluation of eq. 1 form yields best performance for selecting top-1 Knowledge. Table 1 shows strong performance increase compared to dialogue-only model which confirms that considering all components of dialogue is important. Additionally, we verify our cross-domain formulation in Appendix F.

5.2 Persona Retrieval

Fine-tuning of $P_i + D$ model yields performance increase, as shown in Table 2. However, we observe low performance for $P_i + D$ in comparison to P_i and the baseline. We suspect that this is due to lack of score normalization, in that Q & A relationship of Dialogue to true Knowledge may affect likelihood score. Thus fine-tuning the model is a necessity to utilize Q & A formulation properly. Further experimental results are in Appendix G.

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). Table 3 show that performance of the model has increased in top-1 rank setting(0 threshold, 0-Acc)[5]. We further discover that overall / positive / negative non-triviality has all improved. We examine sample count per rank in Figure 3. Further discussions are available in Appendix C.

6 Conclusion

We introduce high-performing Persona-Knowledge dual context retrieval method PK-ICR in this paper. We perform Q & A informed augmentations of data that successfully exploit the interactions between Persona, Dialogue and Knowledge. We perform zero-shot top-1 Knowledge retrieval and precise Persona scoring. We present novel evaluation method of null-positive rank test as to isolate hard-negative effect of Persona-Dialogue augmentation.

We hope to stimulate readers to model dialogue context as an interactive whole of multiple components, rather than considering Persona and Knowledge individually. As NLP community aim to tackle more and more complex dialogue systems, our effort to materialize multi-context identification as a pre-requisite task and to develop specific methods informed by both dialogue and information retrieval research may be further enhanced with application to more complex dialogue retrieval tasks. Thus, we emphasize the importance of this work as the first milestone of **multi-context retrieval for dialogue systems.**

[5]This is performance on persona retrieval free from score-related effect in Appendix G.

Acknowledgement

We sincerely thank the ARR September reviewers for their valuable time spent reviewing the paper. We want to highlight the meta reviewer's comments which were very helpful in our preparing the revision.

References

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

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.

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.

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.

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.

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.

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.

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.

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.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.

A Methodology Overview

Our methodology aims to maximize interactions between all components of a conversation turn (Dialogue, Persona and Knowledge). Knowledge retrieval is a top-1 ranking task (Section 3.1), while we treat Persona retrieval as 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 effectiveness of our Persona retrieval method, along with hard-negative characteristics of Persona-augmented Dialogue (Section 3.3). We tackle both Knowledge and Persona retrieval in sequential manner.[6]

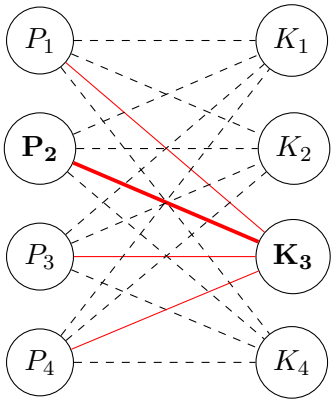

Figure 4: Persona-Knowledge permutations computed in search for best Persona & Knowledge. Best Persona & Knowledge pair is P_2 and K_3. Candidate pairs for Persona search (eq. 3) are marked in red.

B Glossary for Equations

- E is the pair for inference with the retrieval model.
- Q_i, A_j are specific Q & A candidate pairs.
- P_i, K_j is specific Persona and Knowledge pairs, and D is the dialogue corresponding to the pairs.
- M_q is Q & A retrieval model that returns relevancy score, and M_f is the fine-tuned model.
- n / m is the number of Persona / Knowledge respectively.
- $true_i, true_j$ is index of predicted true Persona P and Knowledge K.
- E' is input to the Q & A model similar to E, only difference being fixed true Knowledge K_{true_j}. We note E'_{train} separately because it is data from separate training set formulated in same manner as E' with labeled true Knowledge.
- p_{thres} is the likelihood score threshold utilized to remove Persona that doesn't correspond to the dialogue turn. Ablating this threshold allows us to examine the precise scoring of Persona.
- $\neg T$ is non-triviality metric that represent metric for the model's capability of distinguishing provided samples. Lower is better.
- P_o is null-positive sample utilized with null-positive rank test.
- r is relative rank of P_0 sample, which should be between r_{min} and r_{max}.
- r_{min}, r_{max} are minimum and maximum rank possible for the sample set, in which case worst performance of the model is observed. Typically $r_{min} < 0$ and $r_{max} > 0$. For our experiments, $r_{min} = -1$ and $r_{max} = 4$.
- n_r is count of P_o samples that have rank r.

[6] We note that each step - Knowledge and Persona retrieval - may be further optimized independently.

C Null-Positive Rank Test

C.1 Null-Positive Definition

For our null-positive rank test, we define P_o as Dialogue-only sample D. This sample is relevant to Q & A but acts as a baseline sample for our Persona ranking method. Thus, we rank D & K_{true} against $P_i + D$ & K_{true} as to compute how well the model discriminates hard negatives corresponding to Persona-augmented Dialogue. We normalize the rank of P_o by starting with -1, as P_o should rank right below $P_{pos} + D$ in an ideal Persona distribution where P_{pos} is sufficiently distinct from all P_{neg_i} candidates[7](Figure 2, Table 4).

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

Table 4: We display ideal rank order for Persona-Augmented Dialogue ($P_i + D$) along with null-positive sample P_o (underlined). This table corresponds to notations in Figure 2. If a model is weak, it would not be able to rank null-positive sample correctly against other augmented samples.

C.2 Null-positive Variants

We introduce variants of non-triviality $\neg T$ metric (eq. 8). $\neg T_+$, $\neg T_-$ are useful for further validation in the case where there is an unequal number of positives and negatives. $\neg T_{weighted}$ is useful when a certain rank is more important to avoid. Smaller numbers are better for all variants.

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

$$\neg T_+ = \frac{\sum_{r=0}^{r_{max}} n_r * |r|}{\sum_{r=0}^{r_{max}} n_r} \tag{9}$$

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

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

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

$$\neg T_{weighted} = \frac{\sum_{r=r_{min}}^{r_{max}} w_r * n_r * |r|}{\sum_{r=r_{min}}^{r_{max}} w_r * n_r} \tag{11}$$

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 integrate neural Question and Answering retrieval model from Sentence-BERT library (Reimers and Gurevych, 2019) as starting model M_q. Specifically, we utilize 12 layer MiniLM (Wang et al., 2020) (33M params) based cross-encoder trained on MS MARCO[8] (Nguyen et al., 2016). This model fits very well with our formulation

[7]However, in real-world scenario it may not be the case, i.e. there could be a Persona with "I like hills" for Table 4. This may explain the increase in sample count observed for short rank distances of 1, 2 in Figure 3. Short rank distance is expectedly weighted less as in eq. 8.

[8]MRR@10 on MS MARCO Dev Set: 39.02

since its purpose is for semantic search, with model evaluating short questions and long passages together. For Persona search (eq. 4, 6), we fine-tune for 2 epochs, 32 batch size, and sigmoid activation function with Binary Cross Entropy loss and provide a threshold of 0.5 in our best configuration. We list the official evaluation results on the test data, with TF-IDF baseline provided by dataset authors. We also compare performance with DistillRoBERTa (82M params) based STS[9] and NLI[10] CrossEncoder models. We work with RTX 3090 NVIDIA GPU.

[9] STSbenchmark test performance: 87.92

[10] Accuracy on MNLI mismatched set: 83.98

E Hard Negative Persona Samples

Specifically, while false Persona are not closely related to Dialogue and true Knowledge, they are not exactly contradictory or abnormal even with context (i.e. "I have fantasy about fort" vs "I would like to work with military", see for samples.) Thus, they are high-quality negative samples appropriate for selective fine-tuning of the Persona model.

Corresponding positive and hard negative pair samples are listed in Table 5.

Positive	Hard Negative
Q : I would like to visit France. I know this place, but I don't remember the name of this place. A : The Château de Verteuil is a historic building in Charente, France.	Q : I have English relatives. I know this place, but I don't remember the name of this place. A : The Château de Verteuil is a historic building in Charente, France.
Q : I have the fantasy about fort. What are the attractions in the park? A : The Great Lines Heritage Park, consists of Fort Amherst, Chatham Lines, the Field of Fire (later known as the Great Lines), Inner Lines, Medway Park (sports centre) together with the Lower Lines.	Q : I would like work with military. What are the attractions in the park? A : The Great Lines Heritage Park, consists of Fort Amherst, Chatham Lines, the Field of Fire (later known as the Great Lines), Inner Lines, Medway Park (sports centre) together with the Lower Lines.
Q : I would like to visit the Gallery 30 again. I think I've been there before but I don't remember the name of this place. A : Gallery 30 is an American fine art and craft gallery.	Q : I Like Gettysburg. I think I've been there before but I don't remember the name of this place. A : Gallery 30 is an American fine art and craft gallery.
Q : I love Historic architecture. What is the influence of historical architecture in this gallery? A : The historic 19th Century building that originally housed Gallery 30 is located at 30 York Street, also known as US Route 30 and the historic Lincoln Highway, in Gettysburg, Pennsylvania. The brick building bears an official plaque that verifies it as a Civil War building that was standing during the Battle of Gettysburg in 1863.	Q : I am interested in History. What is the influence of historical architecture in this gallery? A : The historic 19th Century building that originally housed Gallery 30 is located at 30 York Street, also known as US Route 30 and the historic Lincoln Highway, in Gettysburg, Pennsylvania. The brick building bears an official plaque that verifies it as a Civil War building that was standing during the Battle of Gettysburg in 1863.
Q : I have visited Glenridding village five years back. Which village is located to the west of Greenside Mine? A : The mine was west of Glenridding village, which is by the southern end of Ullswater in the parish of Patterdale.	Q : I love Atomic Weapons Research Establishment. Which village is located to the west of Greenside Mine? A : The mine was west of Glenridding village, which is by the southern end of Ullswater in the parish of Patterdale.

Table 5: Positive and hard negative samples constructed from Persona augmented Dialogue as Question (Q). Answer (A) is fixed as true Knowledge.

F Cross-domain Formulation Experiments

We compare other possible formulations of Dialogue and Knowledge via NLI, STS and Q&A retrieval models in Table 6. The models are described in Section 4. For NLIv1 we compare the Knowledge using $1 - contradictory$ score, while for NLIv2 we compare using $entailment$ score. In Table 6, we find expected result in that Q&A model is best with 28 point higher accuracy compared to STS model.

Model Type	Accuracy (%)
D & K_j, NLIv1	9.08
D & K_j, NLIv2	17.96
D & K_j, STS	51.33
D & K_j, Q&A	**79.26 (+27.93)**

Table 6: Knowledge Zero-shot formulation test. We perform top-1 Knowledge selection task via modelling asymmetric interactions of Dialogue and Knowledge as inter-sentence relations available in NLI, STS and Q&A tasks. Zero-shot inference is performed with the cross-encoder models trained for the tasks, as described in Section 5.1.

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 5: Resulting Q&A formulation of Persona & Knowledge pair (eq. 1). Question form is "{Persona} {Dialogue}" while answer is "{Knowledge}".

G Persona Retrieval Threshold Experiments

We describe Persona score normalization from re-training on augmented Persona in this section. Due to the existence of Dialogue (query) D in augmented Persona $P_i + D$, M_q inference score may be higher than actual likelihood of Persona P_i corresponding to given Knowledge K_{true_j}. Thus, we find that fine-tuning the model with binary score on augmented Persona & true Knowledge ($P_i + D$ & K_{true_j}) pairs assists in obtaining the normalized likelihood of Persona given context. The experimental result of Persona accuracy is provided per p_{thres} ablation (Equation 6) in Figure 6. We find that fine-tuned model has increased performance across all thresholds, including 0.0 where the output has top-1 characteristics. We also find that the score increases in tandem with Persona threshold for non-fine-tuned case, in contrast to visible peak at 0.55 for fine-tuned case.

H Retrieval Output Samples

We list some of the retrieved outputs with our best model in Table 7.

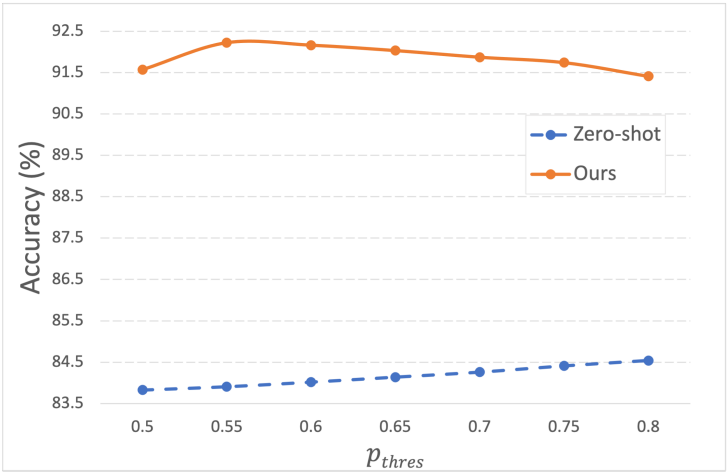

Figure 6: Persona threshold ablation experiments with P_i & K_{true_j} model. We report Persona accuracy. p_{thres} is defined in Appendix B. Dotted line correspond to Zero-shot model, and solid line is our best model. We find visible peak at 0.55 with our best model while Zero-shot model performance keeps increasing > 0.8.

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 7: Persona, Knowledge and Dialogue retrieved examples from our best model.