$\mathcal{K}$-PERM: Personalized Response Generation Using Dynamic Knowledge Retrieval and Persona-Adaptive Queries

Kanak Raj[*1], **Kaushik Roy**[*2], **Vamshi Bonagiri**[3], **Priyanshul Govil**[3], **Krishnaprasad Thirunarayan**[4], **Raxit Goswami**[1], **Manas Gaur**[1,3]

[1]HealthCareNLP LLC
[2]Artificial Intelligence Institute, University of South Carolina
[3]University of Maryland, Baltimore County
[4]Wright State University
kanak8278@gmail.com, kaushikr@email.sc.edu, {vamshib2,pgovil1,manas}@umbc.edu, t.k.prasad@wright.edu, raxit.G@healthcarenlp.com

Abstract

Personalizing conversational agents can enhance the quality of conversations and increase user engagement. However, they often lack external knowledge to appropriately tend to a user's persona. This is crucial for practical applications like mental health support, nutrition planning, culturally sensitive conversations, or reducing toxic behavior in conversational agents. To enhance the relevance and comprehensiveness of personalized responses, we propose using a two-step approach that involves (1) selectively integrating user personas and (2) contextualizing the response by supplementing information from a background knowledge source. We develop $\mathcal{K}$**-PERM** (**K**nowledge-guided **PE**rsonalization with **R**eward **M**odulation), a dynamic conversational agent that combines these elements. $\mathcal{K}$-PERM achieves state-of-the-art performance on the popular FoCus dataset, containing real-world personalized conversations concerning global landmarks. We show that using responses from $\mathcal{K}$-PERM can improve performance in state-of-the-art LLMs (GPT 3.5) by 10.5%, highlighting the impact of $\mathcal{K}$-PERM for personalizing chatbots. Our code is released to the public for further explorations: https://github.com/kanak8278/DialogKPERM

Introduction

Recent trends in Large Language Models (LLMs) have demonstrated remarkable abilities in conversational AI (Jo et al. 2023)(Shin, Hsieh, and Kim 2023). However, personalization is a potential area for LLMs that requires improvement (Zhang et al. 2023). Personalization in conversational AI can go beyond chit-chat conversations and aid in user engagement by understanding their personas better and providing accurate responses (Bender et al. 2021)(Joshi, Mi, and Faltings 2017).

Prior research on personalization has primarily focused on casual conversations, emphasizing details such as a user's preferences. The lack of external knowledge hinders a model's ability to adapt to different personas (Deshpande et al. 2023). Therefore, recent shifts in chatbot personalization utilize both persona information and knowledge (Qian et al. 2021) (Liu et al. 2023). However, identifying a suitable context aligned with user preferences during a conversation remains a significant challenge for current LLMs.

While using various prompting methods may allow a user to steer LLMs toward desired behavior, they only work at an utterance level. This may not be feasible for longer conversations, as the context often shifts (Shuster et al. 2021). Therefore, we require the chatbot to learn how to retrieve appropriate content based on a user's query and assess whether a response requires contextualization (retrieval of meta-information) and personalization (selecting appropriate persona, if necessary).

To address this issue, we propose using **K**nowledge-guided **PE**rsonalization of response generation with **R**eward **M**odulation ($\mathcal{K}$**-PERM**). $\mathcal{K}$**-PERM** uses dynamic knowledge retrieval along with personalization to improve a machine's adaptive capabilities to different personas and contexts. Our Personalized Response Generation task involves two major components:

1. **Understanding Conversation Context:** We use Dense Passage Retrieval (DPR) (Karpukhin et al. 2020) to select the most pertinent information from a larger text corpus containing real-world information.
2. **Incorporating Appropriate Personas:** We introduce a selector module capable of choosing a persona that aligns with the user query. We model persona selection as a multiple-choice question-answering task, which includes an option to opt for "no-persona" in generic cases.

To the best of our knowledge, our study is the first to dynamically and flexibly incorporate knowledge, persona, and conversation history into a learning strategy that can be used to train or guide LLMs to be personalized. Our model-agnostic framework can be utilized to train conversational agents for personalized response generation in open-domain settings. $\mathcal{K}$-PERM outperforms prior attempts in personalization despite being a 24 times smaller model. We also augment GPT 3.5 with $\mathcal{K}$-PERM's responses and show a considerable improvement of 10.5% when compared to GPT3.5 in

[*]These authors contributed equally. The first, third, and fourth authors interned at KAI^2 lab @ UMBC.

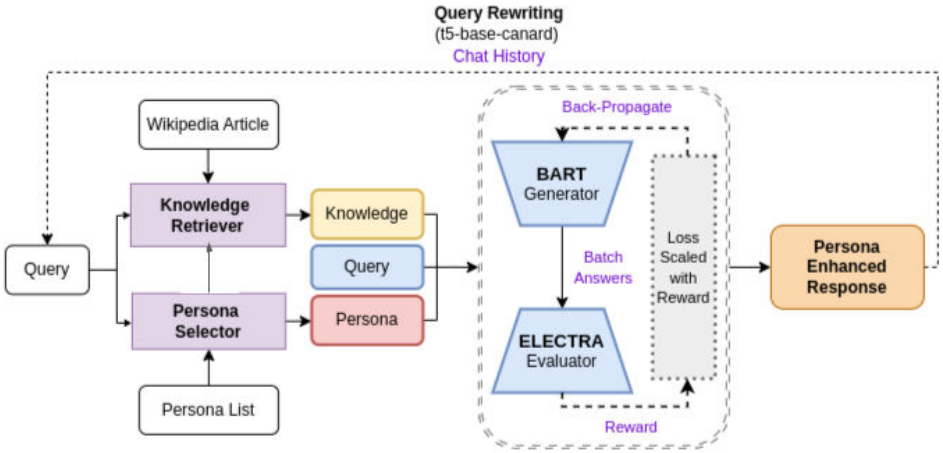

Figure 1: $\mathcal{K}$-PERM Model Architecture Overview. The model architecture comprises a Persona Selector and Knowledge Extractor, which leverages the history and question prompt to identify pertinent persona and knowledge.

a zero-shot setting. Our results demonstrate that knowledge-guided learning can help guide conversational agents toward better personalization.

Methodology

Let $U^t = \{u_1^t, u_2^t, \ldots, u_H^t\}$ be the history of H user utterances on a topic t. Each utterance $u_i^t \in U^t$ is a question and response pair denoted as (q_i^t, r_i^t). The q_i^t are questions by the user, specific to the topic t, and the r_i^t are responses that are either tailored according to a set of n user personas $P = \{p_1, p_2, \ldots, p_n\}$ or generic (no-persona). Our goal is to model the probability of the latest user response r_H^t given a set of K passages, denoted by $\mathcal{Z}_K^{U^t}$, relevant to the utterance history U^t (the passages are drawn from an external knowledge source, e.g., a document store). Thus, our goal is to learn the probability distribution

$$\mathbb{P}_\theta(r_H^t \mid \mathcal{Z}_K^{U^t}) \tag{1}$$

where θ are the parameters of the probability distribution.

$\mathbb{P}_\theta$ can be any auto-regressive language model capable of generating sentences token-wise. Since the user is likely to have responded according to their set of personas, we train a persona selector module P_{select}, that takes as input the user's utterance history U^t, and the set of passages $\mathcal{Z}_K^{U^t}$, and outputs one or more personas from P (denoted as P') for customizing responses using $\mathbb{P}_\theta$. Therefore, Equation (1) is modified as

$$\begin{aligned} P' = P_{\text{select}}(\mathcal{Z}_K^{U^t}, P), \quad P' \subseteq P, \\ \mathbb{P}_\theta(r_H^t \mid \mathcal{Z}_K^{U^t}, P') \end{aligned} \tag{2}$$

$\mathcal{K}$-PERM

Figure 1 explains the entire model architecture of $\mathcal{K}$-PERM. Utilizing the conversation history U^t, we access a document store, retrieve pertinent passages, and rank them based on their relevance. This allows us to leverage the retrieved information to select compatible user personas and generate personalized responses. Our method personalizes based on the personas in P'. If $P' = \emptyset$, generic responses are generated.

Knowledge Retriever

For dynamically retrieving passages based on U^t, we built upon a process called DPR – P_{select} in Equation (2). DPR uses semantic similarity search to retrieve passages from a vectorized database. This allows us to go beyond an *exact-match* by retrieving passages that can answer reasoning-type questions (*what is? what if? what could be?*) by automatically adapting to the input queries in U^t.

We improve DPR in two ways. First, we utilize a Sentence-BERT model for performing a **retrieve-rank process using a paired cross-encoder and bi-encoder**. A cross-encoder retrieves a set of passages given the last query $q_H^t \in U^t$, and subsequently, the bi-encoder ranks and selects the top-K passages to result in $\mathcal{Z}_K^{U^t}$. We create dense encodings of the passages in $z_j \in \mathcal{Z}_K^{U^t}$ using the MPNet model from the Sentence-BERT transformer family (Reimers and Gurevych 2019). Likewise, the encoding for q_H^t is represented as z_H. We fine-tune MPNet on our dataset before using it to obtain dense encodings (refer Appendix D) (Song et al. 2020).

Persona Selector (P_{select})

We model persona selection as a commonsense inference task, conditioned upon the query knowledge $\mathcal{Z}_K^{U^t}$ retrieved using the information in q_H^t and the set of user-personas P, formally written as $P' = P_{\text{select}}(\mathcal{Z}_K^{U^t}, P)$ as shown in Equation (2). The dataset contains the ground truth for the user's personas corresponding to the responses in the utterance history U^t. Using this, we train the P_{select} model as a multi-label classifier and sample the top-2 classes from the resulting logits ($|P'| = 2$). For the base P_{select} model, we empirically observed XLNET as the best performer (Yang et al. 2019).

Response Generation through Reward Modulation

The response is generated by pairing a BART(Base) generator with an ELECTRA(Base) evaluator that measures the similarity between the generated response and the ground truth (Clark et al. 2019). We introduce a balancing reward function (R_i) modulating generative capabilities (e.g., coherence) of the BART model and high fidelity to the ground truth responses (in terms of matched words).

Reward Function

Consider the ground truth response for a query q_i^t to be r_i^t, which consists of n tokens, where each token is indexed using t_i. Let r_k^t be the k^{th} response in the generated response list comprising m tokens, where each token is indexed using t_j. We generate BERT encodings for each word in the response vectors, both for the n words in r_i^t and the m words in r_k^t. The reward (R_i) is

$$R_i = \alpha \cdot \text{BLEU}\left(r_i^t, r_k^t\right) + (1-\alpha) \sum_{(t_i \in r_i^t, t_j \in r_i^k)} \max_{t_i} \text{WMD}\left(t_i, t_j\right) \quad (3)$$

where WMD denotes the Word Mover Distance (Kusner et al. 2015) and $\alpha[*] \in [0,1]$ balances between a well-generated response by BART (given by the WMD distance) and closeness to the ground truth responses (provided by the BLEU score).

Persona-tailored Reward Function

In addition to producing syntactically sound responses and responses that are close to the ground truth, the responses need to be tailored to user-personas, i.e., the output of the P_{select} model. Thus, we modify Equation (3) as follows:

$$R_i = \alpha\,\text{BLEU}\left(r_i^t, r_k^t\right) + \beta \sum_{(t_i \in r_i^t, t_j \in r_i^k)} \max_{t_i} \text{WMD}\left(t_i, t_j\right) + \gamma \cdot \text{Loss}(P_{\text{select}}, P_{\text{gt}}), \quad \alpha+\beta+\gamma = 1 \quad (4)$$

where gt is the ground truth, and $\text{Loss}(P_{\text{select}}, P_{\text{gt}})$ refers to the loss function, i.e., the training error in the persona selected and the ground truth persona (if present).

Experiments

We utilize the FoCus dataset developed by Jang et al. (2022) for our experiments, as it contains customized answers built with persona and Wikipedia knowledge instead of just persona ((Zhang et al. 2018)). See Appendix A for dataset details. Our experiments use BART as the language model $(\mathbb{P}_\theta)$. Fine-tuning details are described in Appendix C.

Evaluation Criteria

We used Rouge–1/2/L/L-Sum and BLEU scores to evaluate $\mathcal{K}$-PERM. In addition, we use two transformer-based metrics for evaluating natural language generation: BERTScore (BF1: BERTScore F1-score) and NUBIA, which measure semantic relations, contractions, irrelevancy, and logical agreement (Zhang et al. 2020; Kane et al. 2020). Semantic relations evaluate whether the generated text is relevant and coherent and maintains the intended meaning and context of the input query. Measuring these characteristics for our proposed model is important to ensure LLMs' correct merging of knowledge and persona.

Baselines

We compare our model with two baselines – **(1) GODEL**: A large pre-trained Transformer-based encoder-decoder model for goal-directed dialogues similar to FoCus (Peng et al. 2022). We used the pre-trained GODEL model and enhanced it with personalization by incorporating a persona selected by our persona selector model; **(2) $\mathbf{BART_{FoCus}}$**: We utilized the BART model provided with the FoCus dataset (Jang et al. 2022). We fine-tuned this model ($\text{BART}_{\text{FoCus}}$) using our training and validation sets for a fair comparison. Results are reported on our test set, although it was not made available by the authors of the FoCus dataset.

Results and Discussion

Table 1 compares three models: $\text{BART}_{\text{FoCus}}$ (406 Million Parameters), GODEL (6 Billion parameters), and $\mathcal{K}$-PERM (250 Million parameters). The results show that the pre-trained version of GODEL, without personalization, performed worse than $\text{BART}_{\text{FoCus}}$. However, when GODEL incorporated our persona selector model, it achieved a syntactic similarity closer to $\text{BART}_{\text{FoCus}}$, suggesting that the persona-based approach improved GODEL's syntactic quality—in terms of semantic similarity measured by BERTScore, GODEL with persona outperformed $\text{BART}_{\text{FoCus}}$, indicating that GODEL, when utilizing personas, generated responses that were more semantically similar to the desired outputs. $\mathcal{K}$-PERM significantly outperformed GODEL in terms of syntactic generation quality and semantic similarity. Ablation studies (see Appendix Table 2) on $\mathcal{K}$-PERM showed that using the ground persona and our persona selector resulted in the highest quality generation both syntactically and semantically. However, there were cases where $\mathcal{K}$-PERM ignored the persona for specific queries requiring personalization, resulting in errors compared to using the ground persona. Despite this limitation, the knowledge retriever used in $\mathcal{K}$-PERM demonstrated competitive performance, relying on information from Wikipedia articles rather than handcrafted knowledge in the FoCus dataset.

Another set of experiments with $\mathcal{K}$-PERM using NUBIA as the metric (see Table 2) showed that using personas yielded better semantic relations, logical agreement, lower contradiction, and lower irrelevancy compared to providing all personas simultaneously. When using retrieved knowledge, $\mathcal{K}$-PERM achieved a higher NUBIA score compared to using handcrafted knowledge in $\text{BART}_{\text{FoCus}}$. In a blind assessment of responses obtained from 5 different systems, generated across 90 user queries with varying numbers of personas, $\mathcal{K}$-PERM consistently outperformed the competition. It achieved the top position in 54 query cases, showcas-

Models	BLEU	R1	R2	RL	BF1
ptGODEL	5.21	25.82	9.87	21.2	43.86
ptGODEL (P_{select})	6.18	30.02	13.77	26.11	44.34
$\text{BART}_{\text{FoCus}}$	6.24	30.98	14.22	26.81	43.85
$\mathcal{K}$-PERM (without P, GK)	2.14	24.68	9.46	21.89	44.15
$\mathcal{K}$-PERM (GK only)	11.2	31.14	15.49	26.73	43.26
$\mathcal{K}$-PERM (GP only)	2.4	27.47	12.03	22.37	44.35
$\mathcal{K}$-PERM (All P+$\mathcal{Z}_k$)	11.22	35.19	19.27	31.18	43.45
$\mathcal{K}$-PERM (GP+$\mathcal{Z}_k$)	**14.72**	**43.09**	**25.43**	**37.95**	**47.36**
$\mathcal{K}$-PERM (P_{select}+$\mathcal{Z}_k$)	<u>12.01</u>	<u>37.53</u>	<u>23.96</u>	<u>33.47</u>	<u>46.06</u>

Table 1: Performance of $\mathcal{K}$-PERM on FoCUS dataset. SP:P_{select}, GP: Ground Persona, $\mathcal{Z}_k$: Retrieved Knowledge, P: Persona, GK: Ground Truth Knowledge. Bold-face: Best and Underlined is 2nd best. pt: pre-trained.

ing its remarkable performance in contrast to GODEL and GPT 3.5, as depicted in Appendix Figure 3.

Finally, we use $\mathcal{K}$-PERM to augment a state-of-the-art LLM, GPT3.5. Our results in Figure 2 show that augmenting GPT3.5 with $\mathcal{K}$-PERM improves its performance significantly (10.5%), highlighting the advantage of $\mathcal{K}$-PERM in personalization.

Conclusion

We created $\mathcal{K}$-PERM, a practical and comprehensive technique for generating personalized responses, demonstrating its superior performance compared to baseline methods. Notably, $\mathcal{K}$-PERM responses closely match human-curated responses on the FoCus dataset. Despite being a simpler language model than ChatGPT, $\mathcal{K}$-PERM ranked second in aligning generated responses with the ground truth in FoCus. The reward modulation setting in $\mathcal{K}$-PERM successfully achieves the desired response generation, enabling it to guide large language models like GPT 3.5 to produce personalized results. Our approach can extend to other domains and large language models using relevant goal-oriented and personalized datasets.

Limitations

- We evaluate $\mathcal{K}$-PERM only on the FoCus dataset. To the best of our knowledge, only one dataset incorporates persona, context, and queries together. However, we believe that with the growing interests and advancements, more such datasets will be constructed, and the solutions will evolve and be generalized.

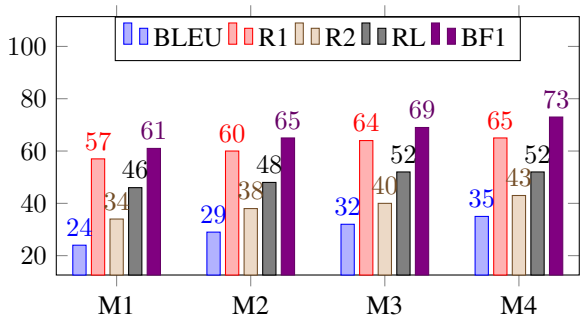

Figure 2: $\mathcal{K}$-PERM improves personalization in GPT 3.5 via zero-shot prompting. This experiment aimed to assess the performance improvement of GPT 3.5 when combined with $\mathcal{K}$-PERM. M1 is GPT 3.5 and M2, M3, and M4 represents zero-shot prompting of GPT 3.5 using responses from $\mathcal{K}$-PERM with (All P+Z_k), (GP+Z_k), and (P_{select}+Z_k) respectively. Score were rounded off for visibility.

- Our methodology did not thoroughly experiment with all the state-of-the-art LLMs such as Llama and Mistral (Jiang et al. 2023). However, as our methods are model-agnostic, applying our work to these models should yield comparable results.

Appendix

A Dataset

We utilize the publicly available FoCus dataset, which consists of passages describing landmarks (Jang et al. 2022). The dataset includes 13,484 dialogs for training and validation, with an average of 5.6 rounds per dialog and approximately 7,715 Wikipedia landmarks. The dialogs contain a total of 75,971 utterances, with an average length of 24.0 per utterance. We split the dataset into three sets: train (10,284 samples), validation (1,600 samples), and test (1,600 samples), comprising 57,928, 9,008, and 9,035 utterances, respectively. The training set includes 36,472 knowledge-based utterances and 21,456 utterances with both persona and knowledge, coming from 4,918 landmarks. Validation and test datasets consist of 5,664 and 5,707 knowledge-based utterances, respectively. Additionally, the validation set includes 3,344 utterances featuring both persona and knowledge, while the test set comprises 3,328 such utterances. The validation set encompasses 1,414 Wikipedia landmarks, whereas the test set involves 1,383 landmarks. The dataset references "ground persona" and "ground knowledge," representing the ground truth persona and passage selected by crowd workers. Additionally, all the questions in the dataset were rewritten using T5-CANARD, a query-rewriting model (Qian and Oard 2021) (Table 4 shows why question rewriting was needed).

B Related Work

PersonaChat and PersonaChat 2.0 are chit-chat conversation datasets used to train conversational agents (Liu, Peng, and Ni 2022; Wu et al. 2020). They incorporate encoder-decoder

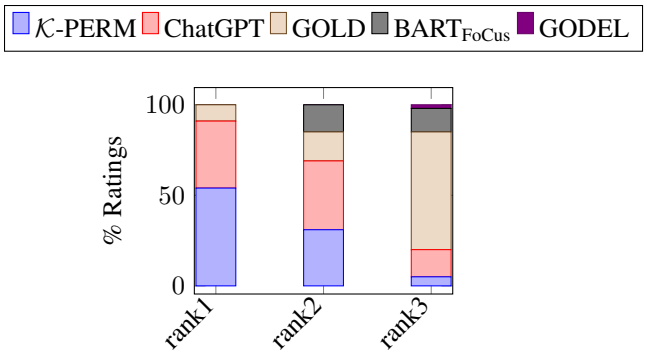

Figure 3: $\mathcal{K}$-PERM was preferred 32% more than ChatGPT by the annotators based on a blind evaluation of 90 queries taken randomly from the FoCus dataset. GOLD is the ground truth in FoCus Dataset.

models, reinforcement learning, few-shot learning, and hierarchical attention mechanisms to improve personalization by considering persona and conversational history (Li et al. 2016; Qiu and Zhang 2021; Young et al. 2022). In previous personalization work, external knowledge was not a significant focus until Mazumder et al. introduced retrieval as a sub-task in PersonaChat to enhance personalization (Majumder et al. 2021). However, their method was not specifically evaluated on retrieval performance, using ROC Stories as a proxy knowledge source. This approach did not adequately capture personalization, especially for goal-oriented or information-seeking dialogues. $\mathcal{K}$-PERM addresses this gap by introducing retrieval augmented generation as a practical baseline to improve personalization (Gaur et al. 2022; Lewis et al. 2020). Unlike previous approaches that fine-tuned models on PersonaChat, $\mathcal{K}$-PERM uses base generative models for evaluation and achieves desired behaviors through reinforcement learning as the evaluator (Lipton et al. 2018; Huang et al. 2023). Additionally, $\mathcal{K}$-PERM introduces unique persona selection, making it the first realistic response generation model for real-time goal-oriented or information-seeking dialogues.

C Human Evaluation

We conducted a blind evaluation of 90 responses generated from 5 systems. As the task was trivial, the authors instructed the annotators verbally. The task was made available on a participant pool management system which credits students completing the task. We experimented with five systems: $\mathcal{K}$-PERM, BART$_{\text{FoCus}}$, ChatGPT, GODEL, and Ground Truth. Our model, $\mathcal{K}$-PERM, consistently stood out as a top performer by securing first place in 11 instances and maintaining a strong second position in 68 instances. This underscores its exceptional performance. The Top 3 results and percentage of times any specific model is selected are depicted in Figure 3. Table 3 shows a sample from our human evaluation. More examples are present https://shorturl.at/uP023.

D Fine-Tuning MPNet for DPR

We used contrastive fine-tuning as described in (Song et al. 2020). Next, we employed a combination of locality-sensitive hashing (LSH) and Facebook AI Semantic Search (FAISS), which uses **Maximum Inner Product Search (MIPS)** (Johnson, Douze, and Jégou 2019) **to efficiently obtain** $\mathcal{Z}_{\mathbf{K}}^{\mathbf{U^t}}$. We evaluated the knowledge retriever's efficiency using BERTScore, which is commonly used to compare retriever-augmented generations(Lim et al. 2023).

We varied K between 5 and 20 and observed high similarity (BERTScore) with ground truth passages (from the FoCus dataset) at $K = 10$. Therefore, we agreed that the appropriate number of retrieved passages would be 10. We also experimented with different Sentence Transformer models for dense encodings other than MPNet and standard retrievers such as TF-IDF and BM25 (refer to Appendix Table 4).

E Training Process

We employed BART, an auto-regressive encoder-decoder model, for generating personalized responses by incorporating special tags like <*question*>, <*knowledge*>, <*history*>, and <*persona*> during fine-tuning. Training took approximately 16 hours on a single NVIDIA T4 GPU, and response generation utilized beam search with a beam size of 5 for stability over nucleus sampling.(Chen and Yang 2021; Shaham and Levy 2022).

Acknowledgements

We acknowledge partial support from the UMBC Faculty Fellowship to Dr. Manas Gaur. We also acknowledge High-Performance Computing Support from HealthcareNLP Softech LLC and NSF MRI Award #1920079. Any opinions, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, UMBC, or HealthcareNLP Softech LLC.

References

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Chen, J.; and Yang, D. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Models	NUBIA	SR	Contraction (%)	Irrelevancy (%)	LA (%)
$BART_{\text{FoCus}}$	0.10	1.601	28.72	72.71	16.75
$\mathcal{K}$-PERM (GK only)	0.24	2.20	11.78	74.4	13.5
$\mathcal{K}$-PERM (All P + $\mathcal{Z}_k$)	0.22	1.90	13.18	70.1	17.5
$\mathcal{K}$-PERM (GP + $\mathcal{Z}_k$)	**0.34**	**2.80**	**11.87**	**66.38**	**21.75**
$\mathcal{K}$-PERM (SP + $\mathcal{Z}_k$)	0.31	2.42	11.85	69.61	19.44

Table 2: Performance of $\mathcal{K}$-PERM using NUBIA Score. Results are compared with baseline $\text{BART}_{\text{FoCus}}$. Bold-faced are best (GP), and Underlined are second-best (P_{select}). LA: Logical Agreement, SR: Semantic Relatedness.

Wikipedia Topic	/wiki/Mount_Monadnock
Query	Where is the mountain?
Personalized Response Generation: Example 1	
Ground Truth	Mount Monadnock is a mountain in the towns of Jaffrey and Dublin, New Hampshire.
$\text{BART}_{\text{FoCus}}$	Mount Monadnock is at New Hamsphire.
GODEL	It's near the Rivers and the Masses.
ChatGPT	Based on the context and your persona, you have already visited Mount Monadnock! It is located in the towns of Jaffrey and Dublin, New Hampshire.
$\mathcal{K}$-PERM*	Mount Monadnock is a **3,165-ft (965 m)** mountain in the towns of Jaffrey and Dublin, New Hampshire, a place you have visited.
Personalized Response Generation: Example 1	
Wikipedia Topic	/wiki/Assiniboine_Park_Zoo
Query	Where is this place?
ChatGPT,	"The query "Where is this place?" is referring to the Assiniboine Park Zoo in Winnipeg, Manitoba, Canada. If you like Canada, you should check out this zoo when possible! It's been a staple of the city since 1904!
$\mathcal{K}$-PERM†	Assiniboine Park Zoo is an **80-acre (32 ha)** zoo in Manitoba, Canada that was established in 1904.

Table 3: Human Assessment: $\mathcal{K}$-PERM did better than ChatGPT in terms of personalized response generation. The personas "I have been to New Hampshire" (top) and "I like visiting Canada" (bottom) were used in the examples above. We boldface numeric values to show focus on the correctness of information delivery using a retrieval-augmented generation mechanism.

Model	**Original Query**				**Rewritten Query**			
# Passages	5	10	15	20	5	10	15	20
TFIDF	0.385	0.412	0.396	0.395	0.403	0.406	0.406	0.406
(TFIDF + ptECE)	0.422	0.434	0.430	0.427	0.441	0.452	0.457,	0.445
BM25	0.398	0.405	0.410	0.401	0.411	0.416	0.416	0.412
(BM25 + ptECE)	0.471	0.465	0.480	0.474	0.475	0.477	0.493	0.467
MPNet	0.542	0.558	0.613	0.621	0.540	0.560	0.608	0.604
$\text{MPNet}_{\text{Fine tuned}}$	**0.642**	**0.658**	**0.683**	**0.661**	**0.640**	**0.660**	**0.688**	**0.684**
ColBERT+X(Lawrie et al. 2022)	0.588	0.584	0.601	0.601	0.598	0.602	0.605	0.601
ColBERT (Khattab and Zaharia 2020)	0.560	0.591	0.603	0.60	0.572	0.602	0.602	0.598

Table 4: Evaluating knowledge retrievers over the indexed landmark Wikipedia articles in the FoCus dataset. Each score represents the maximum BERTScore computed between the retrieved passages and ground truth knowledge in FoCUS. ptECE: pre-trained ELECTRA Cross Encoder (Clark et al. 2020). Bold: Best, Underlined: 2nd best.

Gaur, M.; Gunaratna, K.; Srinivasan, V.; and Jin, H. 2022. Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10672–10680.

Huang, Q.; Zhang, Y.; Ko, T.; Liu, X.; Wu, B.; Wang, W.; and Tang, H. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12916–12923.

Jang, Y.; Lim, J.; Hur, Y.; Oh, D.; Son, S.; Lee, Y.; Shin,

D.; Kim, S.; and Lim, H. 2022. Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10803–10812.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jo, E.; Epstein, D. A.; Jung, H.; and Kim, Y.-H. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3): 535–547.

Joshi, C. K.; Mi, F.; and Faltings, B. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Kane, H.; Kocyigit, M. Y.; Abdalla, A.; Ajanoh, P.; and Coulibali, M. 2020. NUBIA: NeUral Based Interchangeability Assessor for Text Generation. arXiv:2004.14667.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and tau Yih, W. 2020. Dense Passage Retrieval for Open-Domain Question Answering. arXiv:2004.04906.

Khattab, O.; and Zaharia, M. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 39–48.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966. PMLR.

Lawrie, D.; Yang, E.; Oard, D. W.; and Mayfield, J. 2022. Multilingual ColBERT-X. *arXiv preprint arXiv:2209.01335*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1192–1202.

Lim, J.; Kang, M.; Hur, Y.; Jung, S.; Kim, J.; Jang, Y.; Lee, D.; Ji, H.; Shin, D.; Kim, S.; et al. 2023. You Truly Understand What I Need: Intellectual and Friendly Dialogue Agents grounding Knowledge and Persona. *arXiv preprint arXiv:2301.02401*.

Lipton, Z.; Li, X.; Gao, J.; Li, L.; Ahmed, F.; and Deng, L. 2018. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liu, S.; Cho, H. J.; Freedman, M.; Ma, X.; and May, J. 2023. RECAP: Retrieval-Enhanced Context-Aware Prefix Encoder for Personalized Dialogue Response Generation. *arXiv preprint arXiv:2306.07206*.

Liu, Z.; Peng, Y.; and Ni, S. 2022. Personalized Dialogue Generation Model Based on BERT and Hierarchical Copy Mechanism. *Journal of Computer and Communications*, 10(7): 35–52.

Majumder, B. P.; Berg-Kirkpatrick, T.; McAuley, J.; and Jhamtani, H. 2021. Unsupervised Enrichment of Persona-grounded Dialog with Background Stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 585–592.

Peng, B.; Galley, M.; He, P.; Brockett, C.; Liden, L.; Nouri, E.; Yu, Z.; Dolan, B.; and Gao, J. 2022. GODEL: Large-Scale Pre-training for Goal-Directed Dialog. arXiv.

Qian, H.; Dou, Z.; Zhu, Y.; Ma, Y.; and Wen, J.-R. 2021. Learning implicit user profile for personalized retrieval-based chatbot. In *proceedings of the 30th ACM international conference on Information & Knowledge Management*, 1467–1477.

Qian, X.; and Oard, D. W. 2021. Full-Collection Search with Passage and Document Evidence: Maryland at the TREC 2021 Conversational Assistance Track. *The Thirtieth Text REtrieval Conference*.

Qiu, S.; and Zhang, K. 2021. Learning personalized end-to-end task-oriented dialogue for fast and reliable adaptation. In *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, 62–66. IEEE.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Shaham, U.; and Levy, O. 2022. What Do You Get When You Cross Beam Search with Nucleus Sampling? arXiv:2107.09729.

Shin, D.; Hsieh, G.; and Kim, Y.-H. 2023. PlanFitting: Tailoring Personalized Exercise Plans with Large Language Models. *arXiv preprint arXiv:2309.12555*.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33: 16857–16867.

Wu, B.; Li, M.; Wang, Z.; Chen, Y.; Wong, D.; Feng, Q.; Huang, J.; and Wang, B. 2020. Guiding Variational Response Generator to Exploit Persona. arXiv:1911.02390.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Young, T.; Xing, F.; Pandelea, V.; Ni, J.; and Cambria, E. 2022. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11622–11629.

Zhang, K.; Zhao, F.; Kang, Y.; and Liu, X. 2023. Memory-Augmented LLM Personalization with Short- and Long-Term Memory Coordination. *arXiv preprint arXiv:2309.11696*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 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)*, 2204–2213.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.