You Truly Understand What I Need
: Intellectual and Friendly Dialogue Agents grounding Knowledge and Persona

Jungwoo Lim[1], Myunghoon Kang[1*], Yuna Hur[1*], Seungwon Jung[1*], Jinsung Kim[1*],
Yoonna Jang[1], Dongyub Lee[3], Hyesung Ji[2], Donghoon Shin[2],
Seungryong Kim[1§] and Heuiseok Lim[1§]
[1]Korea University, [2]Dialogue Tech Division, NCSOFT, [3]Naver Corporation

{wjddn803,chaos8527,yj72722,redlion0929,jin62304,seungryong_kim,limhseok}@korea.ac.kr,

{hyesung84,dhshin}@ncsoft.com, dongyub.lee@navercorp.com

Abstract

To build a conversational agent that interacts fluently with humans, previous studies blend knowledge or personal profile into the pre-trained language model. However, the model that considers knowledge and persona at the same time is still limited, leading to hallucination and a passive way of using personas. We propose an effective dialogue agent that grounds external knowledge and persona simultaneously. The agent selects the proper knowledge and persona to use for generating the answers with our candidate scoring implemented with a poly-encoder. Then, our model generates the utterance with lesser hallucination and more engagingness utilizing retrieval augmented generation with knowledge-persona enhanced query. We conduct experiments on the persona-knowledge chat and achieve state-of-the-art performance in grounding and generation tasks on the automatic metrics. Moreover, we validate the answers from the models regarding hallucination and engagingness through human evaluation and qualitative results. We show our retriever's effectiveness in extracting relevant documents compared to the other previous retrievers, along with the comparison of multiple candidate scoring methods. Code is available at https://github.com/dlawjddn803/INFO

1 Introduction

To build an ultimate conversational agent that interacts with humans fluently, previous studies provide generative neural network-based models (Sordoni et al., 2015; Vinyals and Le, 2015). Although the answers generated from those models are plausible, they lack informativeness and engagingness resulting in bland responses compared to humans (Li et al., 2016; Gao et al.,

---

* Equal Contributors
§ Corresponding author

Dialogue	
Human: Is it in England?	
Machine: No, it is actually in Scotland where you are going.	
Human: Where in Scotland?	

Human's Persona
I will travel through North Ayrshire.
I am going to Scotland.
I like history.
I am interested in architecture.
I love to garden.

Ground Truth Knowledge
Eglinton Castle was a large Gothic castellated mansion in Kilwinning, North Ayrshire, Scotland..

Predicted Answers	
BART$_{base}$	It is in Scotland, which is a place you love.
BART$_{large}$	It is in Scotland. in Scotland. in Scotland. in

Ground Truth Response
It is in North Ayrshire so you could visit when you travel through.

Table 1: Example of the generated answers from a typical generative model, i.e., BART. We can find that BART$_{base}$ uses different persona sentence which has not appeared human's personal profiles resulting in hallucinated answer. Also, BART$_{large}$ generates less engaging answers by making use of the knowledge only to answer the question. Both generated responses are in the situation of hallucination and are less engaging.

2018). However, for knowledgeable and attractive conversation, people usually provide informative replies by considering the background of the person whom they are talking to. Towards a human-like manner of dialogue, Ghazvininejad et al. (2018) and Dinan et al. (2018) introduce the knowledge-grounded conversation for the knowledgeable and informative responses, whereas Zhang et al. (2018a) suggest the persona-grounded dialogue for the personalized responses to the users.

To improve the machine's answer with the external knowledge base, one injects the factual knowledge into the parameters of the language model (Raffel et al., 2020; Roberts et al., 2020). Despite the models' capability of utilizing external knowledge implicitly, they produce "hallucinations" in the responses (Marcus, 2020). The hallucination

in the dialogue involves the situation where the generated output contradicts the reference knowledge. Also, it includes the situation when the generated output cannot be confirmed from the knowledge source (Ji et al., 2022). To mitigate these hallucinated answers, hybrid models employing parametric memory with non-parametric (i.e., retrieval-based) memory are introduced to directly access external memories, leading the source to be inspected and interpreted (Karpukhin et al., 2020; Petroni et al., 2020; Lewis et al., 2020b).

On the other hand, Zhang et al. (2018a) suggest persona-chat dialogues with the corresponding personal profiles of each interlocutor to avoid general and monotonous answers from the machine. Though See et al. (2019); Liu et al. (2020) show comparable quality in generating personalized conversation, the generated utterances merely confirm each interlocutor's persona resulting in a passive manner of speaking such as "I have four children". In addition, the incoherent topics of the dialogues lead to shallow levels of conversation between the interlocutors. To elaborate on this chit-chat conversation supported by external knowledge, Jang et al. (2022) presents a novel persona-knowledge chat with a generative model that considers persona information and world knowledge altogether. Despite obtaining the knowledge and persona when generating the answers, the generative models' responses still exhibit both hallucination and lesser engagingness as in Table 1.

In this paper, we propose INFO (Intellectual and Friendly dialOg agents) that responds with external knowledge and persona simultaneously. Owing to the enhanced capturing relevancy between the context and each candidate set, the knowledge selector and persona selector for the grounding task are implemented with the poly-encoder. To alleviate hallucinated responses from the model, we adopt retrieval-augmented generation (RAG) (Lewis et al., 2020b) by utilizing non-parametric memory and parametric generator in addition to the enhanced input query. By injecting predicted sources as input to the retrieved-augmented generator, our model maintains consistency between grounding and generation while training. Therefore, our model generates more knowledgeable and engaging answers in an active manner with less hallucination.

We show that INFO achieves the highest scores on both grounding and generation tasks in empirical experiments. Also, we compare diverse candidate scoring modules including bi-encoder, cross-encoder, and poly-encoder and demonstrate their effect on generation. We additionally conduct experiments to show the effectiveness of the retriever module compared to sparse and dense retrievers. The qualitative results and human evaluation are also presented to validate our model's capability to generate human-like answers.

Our contributions are as follows:

- We propose the model that grounds persona information and external knowledge with lesser hallucination and adequate utilization of persona in an active manner simultaneously.

- Our approach suggests that the generated responses from the model are interpretable regarding what the model refers to while generating.

- We show that INFO achieves the SoTA performance in all of the automatic metrics and demonstrate its comparable quality with human evaluation and qualitative analysis.

2 Related Works

2.1 Knowledge Grounded Conversation

To let the neural network models ground external knowledge and generate informative answers, Ghazvininejad et al. (2018) suggests a data-driven neural conversational agent that provides knowledgeable answers. Also, Dinan et al. (2018) introduces open-domain dialogue where the two speakers are talking with Wikipedia knowledge. To inject the external knowledge into the pre-trained language model efficiently, Raffel et al. (2020); Roberts et al. (2020) success in equipping the knowledge into the parameters and show comparable performance in open-domain question and answering tasks. However, the approach is not capable of expand or revise their inherent knowledge and provides hallucination (Marcus, 2020). To overcome the limitations, Lewis et al. (2020b) combines a pre-trained parametric model and non-parametric memory for the open-domain question and answering to reduce hallucination. Since their non-parametric memory can be updated without extra pre-training, revising knowledge is more efficient. Furthermore, it is found that a retrieval-augmented
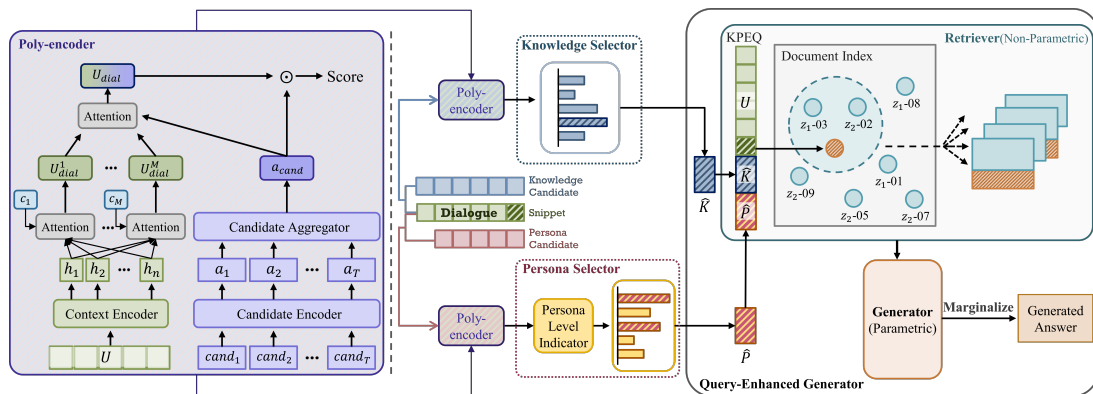
Figure 1: Overview of our method. U is the input comprises dialogue history and knowledge snippet, and $cand$ denotes each candidate from the grounding tasks. The grounding score is obtained through the dot product operation with the representation of input context U_{dial} and candidate a_t. The predicted sources convert into the knowledge-persona enhanced query (KPEQ) with dialogue history and KPEQ is fed into the retrieval-augmented generator to generate the responses.

generator also reduces hallucination in knowledge-grounded conversation as well (Shuster et al., 2021), and a similar approach recently achieves outstanding performance in knowledge-grounded conversation (Paranjape et al., 2021).

2.2 Persona Grounded Conversation

In order to alleviate bland and general answers with consistent personality, Zhang et al. (2018a) constructs a persona-chat dataset. In the dataset, the two interlocutors chat with the persona profile sentences. Along with this dataset, Zhang et al. (2018a) introduces the model with a profile memory network by considering the dialogue history to perform attention over the persona. They enlarge the persona-chat dataset with Reddit corpus, and pre-trained the model with these dataset. After that, they fine-tune pre-trained model on the persona-chat (Mazare et al., 2018). Also, Liu et al. (2020) trains a receiver to reinforce the mutual persona understanding between interlocutors, and Wolf et al. (2019) utilize pre-trained models (Radford et al., 2019) to build personalized dialogue agents.

2.3 Encoders for Sentence Scoring

There exist diverse encoder structures for sentence scoring. Bi-encoder scores the relevance between sentences by feeding context and candidates into separate encoders. An example of bi-encoders are memory networks (Zhang et al., 2018a), transformer memory networks (Dinan et al., 2018), LSTM (Lowe et al., 2015). Since bi-encoder calculates with cached encoded sentence representations, it is relatively fast in computation. However, the bi-encoder has a limitation of capturing mutual information between context and candidates. Cross-encoder, on the other hand, scores by aligning context and candidates in one sequence. A type of cross-encoders is a sequential matching network that is based on deep matching networks (Yang et al., 2018) and gated self-attention (Zhang et al., 2018b). Although using a cross-encoder can achieve rich interaction between the sentences within the encoder, the problem of slow processing still remains. To exploit both benefits of each model, poly-encoder adopts attention mechanism into the bi-encoder architecture and shows satisfactory performances as cross-encoder with fast inference time (Humeau et al., 2019). For the enhanced representation of grounding knowledge and persona, we employ a poly-encoder as a selector for each grounding task.

3 Method

To generate more knowledgeable and engaging dialogue, we introduce our conversational model that grounds external knowledge and persona information as in Figure 1. We first encode the input with the pre-trained language model, and then choose the proper knowledge and persona from the given candidates for each selector. We employ poly-encoder (Humeau et al., 2019) as knowledge selector and persona selector to exploit its enhanced capability of capturing relevance between candidate set and context (i.e., dialogue history). Then, the predicted persona and knowledge are aligned into one sequence

1055

to the dialogue history for consistency between grounding and generation. The sequence is defined as a knowledge-persona enhanced query (KPEQ), then it feeds into the retriever-augmented generator (RAG). The generator then extracts the relevant paragraphs to refer from the knowledge index to reduce hallucination.

3.1 Input Construction

The given dialogue is notated as $\{(u_1^{hm}, u_1^{mc}), ...(u_o^{hm}, u_o^{mc})\}$, where o is the number of rounds. u^{hm} and u^{mc} indicate the utterances of human and machines, respectively. We first take o-th round dialogue history, except for the final machine's reply u_o^{mc}, for the initial input for the model. We define the clue of the dialogue as knowledge snippet cl_k to inform the machine of which topic the user is interested in. The knowledge snippet is the name of the landmark that the user encounters, which is given topic from the dialogue. We then align the dialogue history and knowledge snippet into the one sequence for the model input as $U = \{u_1^{hm}, u_1^{mc}, ...u_o^{hm}, cl_k\}$.

3.2 Model Components

3.2.1 Poly-Encoder Based Candidate Scoring

For knowledge and persona grounding tasks, we suggest poly-encoder-based candidate scoring to leverage the capability of capturing the semantic similarities between the context input and the candidates. It is employed to select proper sources to be used when generating the utterance. When the context input U comes in, we compute the grounding scores of each candidate utilizing the embeddings of context input and encoded candidates in the poly-encoder. The grounding score is used to select the most suitable source(s) in the knowledge selector and persona selector, which will be introduced in the following Section 3.2.2 and 3.2.3.

In poly-encoder architecture (Humeau et al., 2019), candidates are fed into the candidate encoder and denoted as $\{a_1, ..., a_T\}$ where T is the number of candidates in the set. Each candidate embedding a_t is the first output of the candidate encoder, which is represented by the transformer model. After encoding candidates, the context input (i.e., dialogue history) is embedded with a separate context encoder. Unlike the candidate encoder, the context encoder embeds the dialogue into multiple vectors through M context codes

$\{c_1, ...c_M\}$, which are learned for capturing diverse aspects of a given context rather than using one embedding. Each context code is used to extract U_{dial}^m by attending over all the previous layer's output as follows.

$$U_{dial}^m = \sum_j w_j^{c_m} h_j \qquad (1)$$

Note that the $h_1, ..., h_n$ is the output of the pre-trained language model and n is the number of tokens in the input. The weights are computed as $(w_1^{c_m}, ..., w_n^{c_m}) = \text{softmax}(c_m \cdot h_1, ..., c_m \cdot h_n)$.

Then, the final attention proceeds between the global features of the input and a given candidate. In other words, the final dialogue feature U_{dial} is obtained by aggregating each dialogue feature U_{dial}^m, while gaining richer interactions with context codes as in Equation 2.

$$U_{dial} = \sum_m w_m U_{dial}^m, \qquad (2)$$

where $w_1, ..., w_M$ can be obtained from $\text{softmax}(a_t \cdot U_{dial}^1, ..., a_t \cdot U_{dial}^M)$.

The final predicted candidate is chosen based on the highest score that is acquired from the dot product operation as $(U_{dial} \cdot a_t)$.

3.2.2 Knowledge Selector (KS)

We build a knowledge selector for the knowledge grounding task, employing poly-encoder-based candidate scoring. When the grounding scores are produced from the candidate scoring module, the label with the highest score is selected as the predicted knowledge.

The knowledge loss $\mathcal{L}_{KG}$ for the knowledge grounding task is computed with cross-entropy loss (Brier et al., 1950) as in Equation 3.

$$\mathcal{L}_{KG} = -\sum_j kl_j \cdot \log \hat{kl}_j, \qquad (3)$$

kl_j is the ground-truth label from the knowledge candidates of the j-th example.

3.2.3 Persona Selector (PS)

We also implement a persona selector for the persona grounding task. Since multiple personas can be chosen to generate the responses, consideration of one or more persona sentences are needed. Similar to the knowledge selector, we assign the grounding score to each persona

candidate with the candidate scoring module as in Equation 1 and 2.

When the scores of each candidate are computed from the candidate scoring module, then the persona level indicator classifies which the number of the persona should be selected with the [CLS] token of the model input U. After predicting the level of persona-engagingness, we pick persona sentences to be grounded according to the number predicted. For example, if the persona level indicator predicts 2, then top-2 persona sentences are chosen in the persona grounding task. The selected persona sentence(s) are marked as 1 otherwise, 0. We use binary cross-entropy loss for persona grounding as in Equation 4.

$$\mathcal{L}_{PG} = \\ -\sum_j pl_j \cdot \log \hat{pl}_j + (1 - pl_j) \cdot \log(1 - \hat{pl}_j) \quad (4)$$

Note that pl_j is the ground-truth label from the knowledge candidates of the j-th example.

3.2.4 Query-Enhanced Generator

Following the works of Lewis et al. (2020b), we exploit the retrieval augmented generation's capability to reduce hallucination and access the memory directly. For a consistent way of training while solving grounding and generation tasks, we reconstruct the query that feeds into the retriever. When the knowledge and persona are predicted from each selector, we aggregate them with dialogue history into one sequence. Then, the final query is denoted as $\text{KPEQ} = \{U; \hat{P}; \hat{K}\}$ and defined as a knowledge-persona enhanced query. $\hat{P}$ and $\hat{K}$ are predicted persona and knowledge from each candidate set, respectively.

The retriever r_η aims to search top-K latent paragraphs with the KPEQ. We utilize a pre-trained dense passage retriever (DPR) (Karpukhin et al., 2020) trained on natural question dataset (Kwiatkowski et al., 2019) which has parametric memory and bi-encoder architecture to retrieve a latent document embedding following Lewis et al. (2020b) :

$$r_\eta(z|\text{KPEQ}) \propto exp(\mathbf{d}(z)^\top \mathbf{q}(\text{KPEQ})), \quad (5)$$

where $\mathbf{d}(\cdot)$ is an embedding from a document encoder and $\mathbf{q}(\cdot)$ is a representation from query encoder, both implemented with BERT_{base}. z denotes the list of document.

With the relevant paragraphs from the retriever, we employ RAG-Token architecture as the generator to borrow its strength of predicting each target token based on top-K different paragraphs. Since RAG-Sequence, which has a different architecture to RAG-Token, uses the same document from the retriever to predict each token as depicted in Equation 6, the result may opt to depend on the retrieved document (Lewis et al., 2020a). The two different versions of RAGs (Lewis et al., 2020b) are as follows:

$$S_{\text{RS}}(y|x) \approx \\ \sum_{z \in \text{top-k}(p(\cdot|x))} r_\eta(z|x) \prod_i^N g_\theta(y_i|x, z, y_{1:i-1}) \quad (6)$$

$$S_{\text{RT}}(y|x) \approx \\ \prod_i^N \sum_{z \in \text{top-k}(p(\cdot|x))} r_\eta(z|x) g_\theta(y_i|x, z, y_{1:i-1}), \quad (7)$$

where S_{RS} indicates our method with RAG-Sequence architecture and S_{RT} denotes ours with the RAG-Token model. x is a token of KPEQ and y_i is a single token from the ground truth responses. Also, z is a retrieved paragraph from the retriever and N is the maximum sequence length.

The S_{RT} generator $g(\cdot)$ marginalizes the loss from different paragraphs when generating answers. In detail, the generator outputs a distribution for the next token for each document before marginalizing as in Equation 7 where η denotes the parameter of the retriever, and θ indicates the parameter of the generator. After that, the generator repeats the process with the following output token. Finally, the S_{RT} aims to generate the next token following an auto-regressive manner with a standard beam search. In other words, the model minimizes the negative marginal log-likelihood for each input/output pair (KPEQ_j, y_j). The language model loss is formulated as :

$$\mathcal{L}_S = -\sum_j \log p(y_j|\text{KPEQ}_j) \quad (8)$$

3.3 Final Objectives

We then train the full model in the multi-tasking manner. The full objectives of the model is indicated as Equation 9.

$$\mathcal{L} = \lambda_{KG}\mathcal{L}_{KG} + \lambda_{PG}\mathcal{L}_{PG} + \lambda_S\mathcal{L}_S \quad (9)$$

Models	Generation						Grounding (Acc.)	
	chrF++	BLEU	R-1	R-2	R-L	BERTScore	Persona	Knowledge
GPT2$_{small}$	28.73	11.43	36.58	19.44	32.62	88.56	67.44	69.59
GPT2$_{medium}$	30.12	12.31	38.29	21.17	34.12	88.92	67.44	72.42
BART$_{base}$	29.77	11.99	36.24	19.73	32.13	88.35	67.45	72.18
BART$_{large}$	30.69	11.91	36.57	19.83	32.05	88.10	67.44	71.01
INFO (S_{RS})	51.33	29.36	53.36	40.36	51.16	92.00	**82.70**	**99.24**
INFO (S_{RT})	**53.29**	**31.46**	**58.26**	**42.35**	**53.06**	**92.29**	80.87	99.22

Table 2: Main results on the official validation set. S_{RS} denotes our method with RAG-Sequence architecture and S_{RT} indicates the model with RAG-Token model as generator. The models are evaluated by generation metrics, including chrF++, BLEU, ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), and BERTScore.

We control the proportion of each task and we set λ_{KG}, λ_{PG}, and λ_S as 1:1:5 for the experiments, respectively. We find the value of each λ with manual search.

4 Experiments

4.1 Experiment Details

Dataset FoCus (Jang et al., 2022) is the dataset for customized dialogue benchmark, where each conversation is directly grounded with knowledge and persona. The dataset includes knowledge-aware dialogue with personal profiles between humans and machines. There are 12,484 dialogues about 5,152 knowledge sources from Wikipedia and 32,855 persona sentences. To validate the knowledge grounding capability and customized dialogue generation, we evaluate our method with the official FoCus validation set for the effectiveness of experiments since the result from the official test set can be tested only through the leaderboard[*].

Experimental Setup For each candidate scoring module, we implement poly-encoder (Humeau et al., 2019) with BERT$_{large}$, and the number of context codes is 16. For the dialogue generation, we implement our method with Hugging Face (Wolf et al., 2020) and use `facebook/rag-token-nq` as the backbone model. We use the same architecture of retriever and generator from RAG along with the decoding and leverage our knowledge index for non-parametric query-document ranking with FAISS library (Johnson et al., 2019). The knowledge index consists of the paragraphs from the given Wikipedia knowledge entitled with the name of the given landmark. We set learning rate as 6.25e-6 with AdamW (Kingma and Ba, 2014)

for the optimization. The batch size is set as 32, and the number of dialogue history is 1. The whole model was trained for three epochs on RTX A6000 GPU and took 8 hours per one epoch.

Baselines We implement the baselines from previous study (Jang et al., 2022) and we conduct experiments with GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020a) as well. For a fair comparison, we demonstrate the results on GPT-2$_{small}$, which has 12 layers, and BART$_{base}$, which has 6 encoders and 6 decoder layers. Also, GPT-2$_{medium}$ contains 24 layers of the decoder, and BART$_{large}$ possesses 12 layers for each encoder and decoder.

4.2 Automatic Evaluation

We show the main results on the FoCus dataset with automatic metrics in grounding and generation tasks. The official metrics for the benchmark are chrF++ (Popović, 2017), BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004). To consider the semantic similarity score for each token between candidate and reference sentences using contextual representation, we additionally adopt BERTscore (Zhang* et al., 2020). For grounding task, we used accuracy for both knowledge and persona grounding, and F1 score for the persona grounding.

In Table 2, it is found that our method shows substantial improvements in all the metrics from generation to grounding compared to the baselines. Especially, the performances of INFO increase over 18% at least regarding the generation metrics except for BERTScore. Furthermore, our model achieves remarkable success in persona and knowledge accuracy. Unlike the performance in other generation metrics, S_{RS} demonstrates better persona accuracy than S_{RT}. This result might be

Model	Generation						Grounding		
	chrF++	BLEU	R-1	R-2	R-L	BERTScore	Persona (Acc.)	Persona (F1)	Knowledge (Acc.)
S_{RT} Bi-encoder	51.83	29.51	56.35	40.80	51.37	91.86	**88.10**	38.20	99.18
Cross-encoder	49.90	27.18	53.57	38.25	49.29	91.52	87.09	35.32	**99.49**
Poly-encoder	**53.29**	**31.46**	**58.26**	**42.35**	**53.06**	**92.29**	80.87	**39.56**	99.22

Table 3: Performances comparison between the encoding modules for grounding tasks

attributed to the architecture of the generator, which is more applicable to sentence classification tasks such as persona grounding. The official test result is also demonstrated in Appendix A, but BERTscore is missing due to the unreleased ground truth.

4.3 Human Evaluation

We conduct a human evaluation to validate the responses from our model through Amazon Mturk services[†]. The assessment criteria are fluency, adequacy, provenance, engagingness, and hallucination. In specific, provenance is the level of utilization of the ground truth knowledge into the responses, whereas engagingness means how much the answers are persona-related. Also, hallucination indicates whether the answer contradicts the persona and knowledge or cannot be verified from the source content. We randomly chose 50 dialogues from the official test set, and three workers were allocated to evaluate each dialogue generated by our model and baselines. We asked the workers to rank the answers according to each criterion following Cho and May (2020). Rank is scaled from 1 to 5, and the lower number is mapped to the better quality except for hallucination. The agreement between the annotators is calculated with Fleiss' Kappa coefficient and is 0.4185 indicating fair agreement. The relations between the annotators hardly exist since we collect the results from the Amazon Mturk workers.

As in Table 4, INFO surpasses BART$_{base}$, BART$_{large}$, GPT-2$_{small}$ and GPT-2$_{medium}$ in all of the criteria. INFO achieves the highest rank in adequacy, fluency, and provenance and generates a more human-like response than other generative models. Also, the workers ranked our model the lowest when they were asked to rank the responses in the most hallucinated order. Thus, it can be found that INFO generates more engaging and fewer hallucination utterances with respect to the human. The distribution of the rank per each criterion is illustrated in Appendix B.

Models	Avg. Rank				
	Ad. ↓	Fl. ↓	Prov. ↓	Eng. ↓	Hall. ↑
GPT-2$_{small}$	3.57	3.41	3.58	3.46	2.49
GPT-2$_{medium}$	3.11	3.10	3.04	3.25	3.02
BART$_{base}$	3.43	3.29	3.47	3.22	2.45
BART$_{large}$	3.31	3.63	3.29	3.44	2.69
INFO (Ours)	**1.57**	**1.57**	**1.62**	**1.63**	**4.35**

Table 4: Human evaluation. The value in the table is the average rank of the each model's response. The abbreviation Ad. Fl. Prov. Eng. and Hall denote adequacy, fluency, provenance, engaginess, and hallucination, respectively.

5 Results and Analysis

5.1 Variants on Candidate Scoring Module

To validate the poly-encoder as a candidate scoring module, we apply diverse candidate scoring modules, including the bi-encoder and cross-encoder. From the results in Table 3, we can find that the poly-encoder outperforms in the generation task. In the grounding task, S_{RT} with cross-encoder scoring shows improved accuracy on grounding persona and knowledge. The result seems to be S_{RT} with bi-encoder and cross-encoder are better than that with poly-encoder. However, the F1 score of INFO is higher than the two candidate scoring modules implying that low accuracy in persona is due to the tendency of active use on the persona in poly-encoder while the other two models opt to predict not to use persona sentence. The results suggest that the high accuracy of persona not always guarantees the engagingness in the dialogue.

5.2 Comparison on other Retrievers

We show that INFO is effective in retrieving knowledge compared to other sparse and dense retrievers. We retrieve the knowledge from our knowledge index built with Wikipedia paragraphs. We utilize TF-IDF (Joachims, 1996), and deep passage retrieval (DPR) (Karpukhin et al., 2020). In the case of TF-IDF, we set the sum of query

and knowledge tokens less than or equal to 512, which is the maximum sequence length of DPR and INFO. We use `bert-base-uncased` as the tokenizer. For DPR, we extract less than 40 knowledge using TF-IDF due to memory limitations. We first retrieve the five paragraphs related to the query that comprises knowledge snippet, dialogue history, predicted knowledge candidate, and selected persona sentences. In Table 5, we find that the retriever we used outperforms compared to the TF-IDF and DPR in all the metrics, including BERTscore. The results imply that INFO's retriever is suitable for extracting similar paragraphs rather than other retrievers.

Model	chrF++	BLEU	R-1	R-2	R-L	BERTScore
TF-IDF	19.91	3.52	13.91	9.96	12.43	51.54
DPR	20.57	3.86	12.44	6.55	10.20	47.48
INFO	**26.36**	**7.40**	**15.48**	**12.18**	**14.32**	**53.14**

Table 5: Comparison with other retrievers

5.3 Effect of Selectors on Generation

We measure each selector module's effect on the generation task by changing the query which feds into the retriever on a validation set. The experimental results are shown in Table 6, where GT_K, GT_P represents ground truth knowledge and persona. Although the query that comprises the ground truth source shows the highest scores, INFO demonstrates comparable results on the generation task. From the result where the performance increase of INFO + GT_P is larger than that of INFO + GT_K about 2.8%p, we can identify that our persona selector still has more space to achieve its maximum level.

Query	chrF++	BLEU	R-1	R-2	R-L	BERTScore
INFO (RT)	53.29	31.46	58.26	42.35	53.06	92.29
+GT_K	53.35	31.56	58.31	42.55	53.18	92.29
+GT_P	56.19	34.39	61.61	45.46	56.01	92.79
+GT_K+GT_P	**56.40**	**34.60**	**61.88**	**45.64**	**56.16**	**92.84**

Table 6: Comparison between the generation performances based on the variants of query with ground truth knowledge and persona. Note that all the performance is evaluated with the official validation set.

5.4 Qualitative Analysis

In Table 7, an example from the predicted results is illustrated. In the case of BART$_{large}$, and GPT-2$_{medium}$, the responses only reflect the ground

Given Landmark

Finding Nemo Submarine Voyage

Dialogue

Human: What area of the park is this ride in?
Machine: This ride is located in the Tomorrowland area of Disneyland.
Human: Has this ride always been about Finding Nemo?

Human's Persona

I've never been to California.
My favorite cartoon is Finding Nemo.
I would like to visit Disneyland.
My favorite color is yellow.
I enjoy swimming.

Ground Truth Knowledge (Grounding)

Based on the characters and settings of the 2003 Disney · Pixar, Finding Nemo, it is a re-theming of the classic Submarine Voyage attraction that operated from 1959 to 1998.

Retrieved Knowledge (Generation)

The original Submarine Voyage was built in 1959 as part of the then new Tomorrowland...

In 2008, Finding Nemo Submarine Voyage received an award for outstanding achievement from the Themed Entertainment Association.

Finding Nemo Submarine Voyage is ... which opened on June 11, 2007. ... it is a re-theming of the classic Submarine Voyage attraction that operated from 1959 to 1998.

On July 15, 2005, two days before the 50th Anniversary of Disneyland, the Finding Nemo Submarine Voyage was officially announced at the new Turtle Talk...

The attraction reuses the eight original 1959 Submarine Voyage through Liquid Space attraction ...

Predicted Answers

BART$_{base}$	Yes, that's right. You're a fan of the "Fantasy" film, so I.
BART$_{large}$	Yes, the ride is based on the characters and settings of the 2003 Disney · Pixar film
GPT-2$_{small}$	No, it was originally a way to show that you love Finding Nemo.
GPT-2$_{medium}$	Yes, it has operated from 1959 to 1998.
INFO (Ours)	No, this attraction is actually a re-theme of the classic submarine voyage attraction that operated from 1959 to 1998. The attraction is based on the characters and settings of the 2003 Disney Pixar film Finding Nemo, which is your favorite cartoon.

Ground Truth Response

No, your favorite cartoon is a new addition to this ride. The current Finding Nemo ride is a re-theming of the classic "Submarine Voyage" attraction that operated here from 1959 to 1998.

Table 7: Qualitative result. All the predicted results in grounding task are from our model, INFO and it predicts the correct answers in both tasks. We add other baselines' responses for comparative analysis.

truth knowledge resulting in less engaged answers without any persona-related phrases. Although BART$_{base}$ seems to employ a persona sentence in the form of the phrase "You're fan of the Fantasy film", its used sentence does not appear in human's personal profiles. This result also indicates that the utterance is hard to identify its provenance on the knowledge source. Moreover, GPT-2$_{small}$ generates the utterance that contradicts the ground truth knowledge. From the result, we can find that the generated responses from the baselines show hallucinations on both persona and knowledge. Unlike other baselines, our model blends ground truth knowledge and persona sentence into the

response with less hallucination and engagingness. In addition, the retrieved knowledge source that our model refers to provides interpretability and provenance of the responses to the users. More examples are also depicted in Appendix C.

6 Conclusions

In this paper, we presented a conversational agent that generates responses grounding the user's persona and external knowledge. We utilized poly-encoder-based candidate scoring for each grounding task. We additionally implement persona level indicator to consider multiple persona selections for delicate persona grounding. With predicted sources, we construct a knowledge-persona enhanced query to retrieve latent paragraphs, and they are used to generate informative and engaging responses by marginalizing loss for each token. We show that our method achieves the state-of-the-art (SoTA) score in both grounding and generation tasks in the persona-knowledge conversation dataset. We also demonstrate that the responses from INFO show less hallucination and more engagingness through human evaluation and qualitative analysis. We also compare the grounding modules and retrievers to show INFO's effectiveness.

7 Limitations

The proposed model INFO has limitations. Given the INFO's settings, the model cannot deal with real-world application, which means the absence of ground truth knowledge or persona candidates in the grounding task. We also conducted the human evaluation to evaluate the capability of the proposed model's mitigating hallucination in dialogue generation. However, the number of cases is relatively small for evaluating the capability of mitigating hallucination. Finally, INFO demands high GPU computation resources, since it marginalizes loss at the token level.

We plan to improve the INFO for future work. We will train and evaluate the INFO in open-domain settings as well as real-world settings for the applicable conversational agents. Moreover, we will conduct human evaluations with more cases. Especially, we will enhance the way of quantitative measurement for the model's hallucinated answers. Last but not least, we will improve the generator of INFO with more computationally efficient components.

8 Acknowledgement

References

Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. *ACL 2018*, page 2.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin,

Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10803–10812.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Thorsten Joachims. 1996. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

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

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*. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Gary Marcus. 2020. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2021. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. In *International Conference on Learning Representations*.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and William B Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

Oriol Vinyals and Quoc V Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 245–254.

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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752.

A   Automatic Evaluation on Official Test Set

Models	Generation					Grounding (Acc.)	
	chrF++	BLEU	R-1	R-2	R-L	Persona	Knowledge
GPT2$_{small}$	28.83	11.60	36.28	19.56	32.42	67.83	70.95
GPT2$_{medium}$	30.34	12.58	38.35	21.16	34.34	67.64	72.46
BART$_{base}$	29.80	12.15	36.26	19.73	32.06	67.66	72.02
BART$_{large}$	30.63	11.86	36.36	19.42	31.73	67.62	70.53
INFO (RS)	52.81	29.41	56.37	40.41	51.16	**82.74**	98.88
INFO (RT)	**54.61**	**32.33**	**58.27**	**42.39**	**53.09**	80.83	**99.10**

Table 8: Main results on the official test set. RT indicates the model with RAG-Token model as generator. The models are evaluated by generation metrics, including chrF++, BLEU, ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L). The accuracy for persona grounding task and knowledge grounding task are also noted. Since BERTscore is not the official generation metric, we cannot evaluate the result on the metric as the ground truth of the test is not yet disclosed.

B   Human Evaluation Distribution on Each Criteria

(a) Adequacy
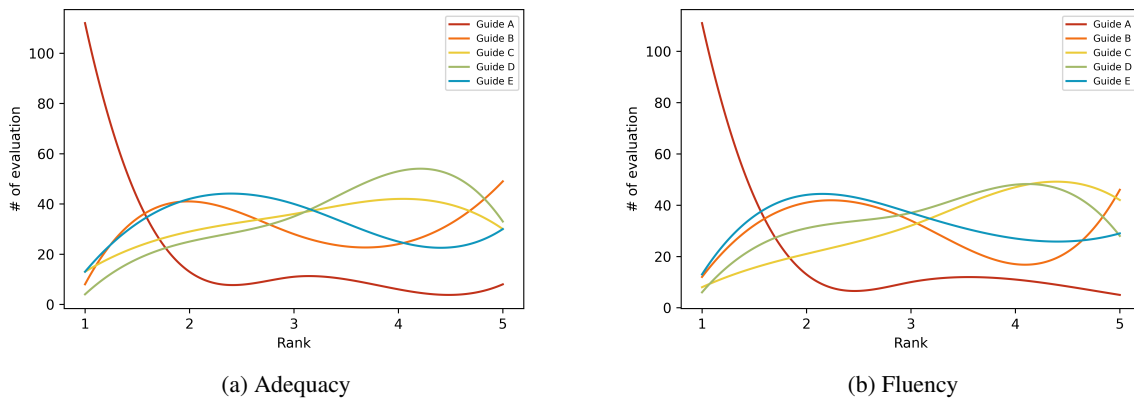
(b) Fluency

Figure 2: The distribution of the rank on the adequacy and fluency criteria. Guide A to E indicates INFO, BART$_{base}$, BART$_{large}$, GPT-2$_{small}$, and GPT-2$_{medium}$, in the order.

(a) Provenance
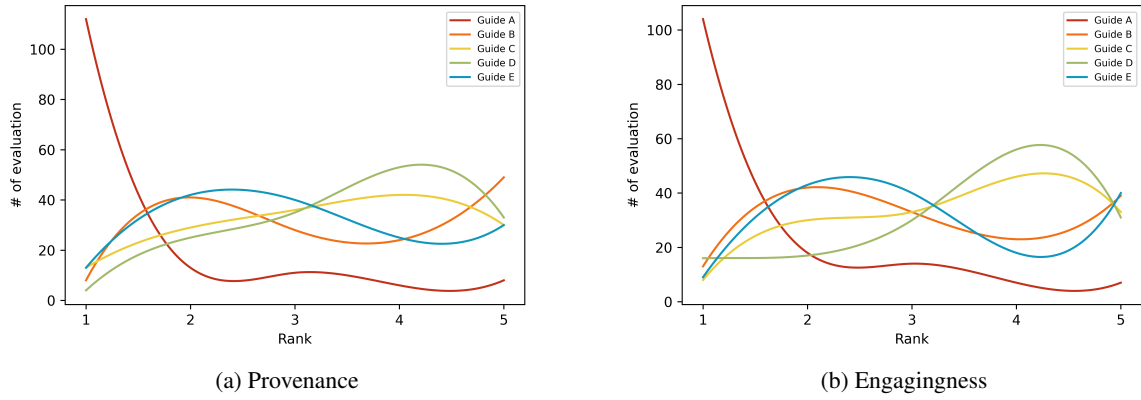
(b) Engagingness

Figure 3: The distribution of the rank on the provenance and engagingness criteria. Guide A to E indicates INFO, $BART_{base}$, $BART_{large}$, GPT-2_{small}, and GPT-2_{medium}, in the order.
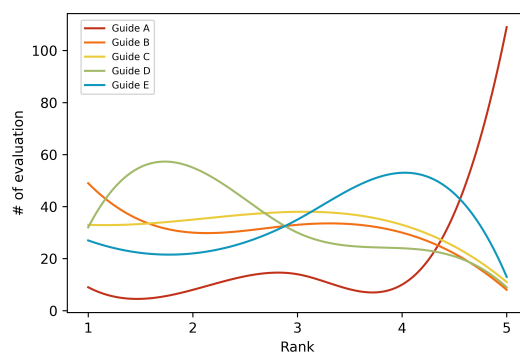
Figure 4: The distribution of the rank on the less hallucination criterion. Note that the highest rank (1) means the most hallucinated. Guide A to E indicates INFO, $BART_{base}$, $BART_{large}$, GPT-2_{small}, and GPT-2_{medium}, in the order.

C Qualitative Results

Given Landmark

Nocton Hall

Dialogue

Human: I know this place, but I don't remember the name of this place.

Human's Persona

I have a hall in my house.
I have worked with military hospital.
I would like to go to England.
I have heard about Frederick John Robinson.
I have respect to American Officers.

Ground Truth Knowledge (Grounding)

Nocton Hall is a historic Grade II listed building in the village of Nocton, in Lincolnshire, England.

Retrieved Knowledge (Generation)

"Hogesdon" is first recorded in the Domesday Book, meaning an Anglo-Saxon farm (or "fortified enclosure") belonging to Hoch, or Hocq...

By 1677, the church was in such poor structural condition that the Dean of Chichester convened a consistory court with three parishioners and asked them why repairs had not been carried out...

he south transept had historically served as the mausoleum for the Cavendish family, who resided at nearby Latimer...

Following the diminished influence of the ancient Priory, predecessor of the two parishes of St Bartholomew, disputes began to arise over rights to tithes and taxes payable by lay residents who claimed allegiance with the nearby and anciently associated parish of St Botolph Aldersgate - an unintended consequence and legacy of King Kenry VIII's religious reforms.

a wall-mounted stone pulpit, a stone altar, a series of tall, pointed windows high in the walls, an unadorned stone font and a short wooden steeple-little more than an extended belfry-extending from the nave roof...

Predicted Answers

BART$_{base}$	This is the Nocton Hall, a historic Grade II listed building in the village of Nocton,
BART$_{large}$	Nocton Hall is a historic Grade II listed building in the village of Nocton, in Lincoln
GPT-2$_{small}$	This is Nocton Hall, a historic Grade II listed building in the village of Nocton, in
GPT-2$_{medium}$	This is the Nocton Hall, a historic Grade II listed building in the village of Nocton,
INFO (Ours)	This will be a nice vision for you. Nocton Hall is a historic Grade II listed building in England that you would like to visit.

Ground Truth Response

You can see Nocton Hall in the village of Nocton, in Lincolnshire of England, the country you want to visit.

Given Landmark

Maiden Castle, Dorset

Dialogue

Human: Wow, this is amazing! What is this?
Machine: It is Maiden Castle in Dorset. I thought you would like it since you are interested in historic forts.
Human: Who owns the site today?

Human's Persona

I like Britain.
I have been to Dorset.
I am interested in historic forts.
I hope to work for English Heritage.
I would like to visit an old fort.

Ground Truth Knowledge (Grounding)

Today the site is protected as a Scheduled Ancient Monument and is maintained by English Heritage.

Retrieved Knowledge (Generation)

Portland Castle is an artillery fort constructed by Henry VIII on the Isle of Portland, Dorset, between 1539 and 1541...

this version of events, or even that the hill fort was attacked by the Romans...

Between 1985 and 1986 further excavations under Niall Sharples were prompted by the hill fort's deteriorating condition, partly caused by the large number of visitors to the site...

a Tudor rose and the initials E.R. (Elizabeth Regina), has been preserved and can be seen in the inner bailey of the castle mounted on a replica carriage...

Constructed on a territorial boundary in about 600 BC, the first hill fort at Maiden Castle was a 6.4-hectare (16-acre) area surrounded by a single ditch...

Predicted Answers

BART$_{base}$	The site is maintained by English Heritage, the country you are from.
BART$_{large}$	Today the site is owned by English Heritage.....
GPT-2$_{small}$	Today the site is protected as a Scheduled Ancient Monument and is maintained by English Heritage.
GPT-2$_{medium}$	Today the site is maintained by English Heritage.
INFO (Ours)	Today the site is owned by English Heritage. You may wish to research this further since you hope to work for English Heritage.

Ground Truth Response

It is owned by English Heritage; a company you hope to work for.

Table 9: Qualitative results. All the predicted results in grounding task are from our model, INFO and it predicts the correct answers in both tasks. We add other baselines' responses for comparative analysis.