RA2FD: Distilling Faithfulness into Efficient Dialogue Systems

Anonymous ACL submission

Abstract

Generating faithful and fast responses is crucial in the knowledge-grounded dialogue. Retrieval Augmented Generation (RAG) strategies are effective but are inference inefficient, while previous Retrieval Free Generations (RFG) are more efficient but sacrifice faithfulness. To solve this faithfulness-efficiency trade-off dilemma, we propose a novel retrieval-free model training scheme named Retrieval Augmented to Retrieval Free Distillation (RA2FD) to build a retrieval-free model that achieves higher faithfulness than the previous RFG method while maintaining inference efficiency. The core idea of RA2FD is to use a teacher-student framework to distill the faithfulness capacity of a teacher, which is an oracle RAG model that generates multiple knowledge-infused responses. The student retrieval-free model learns how to generate faithful responses from these teacher labels through sequence-level distillation and contrastive learning. Experiment results show that RA2FD let the faithfulness performance of an RFG model surpass the previous SOTA RFG baseline on three knowledge-grounded dialogue datasets by an average of 33% and even matching an RAG model's performance while significantly improving inference efficiency.

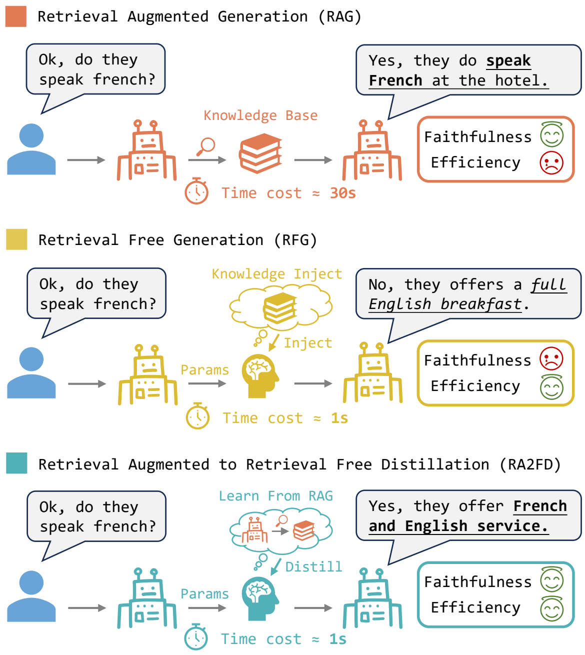

Figure 1: llustration of our contribution. The RAG system is faithful but is time-consuming during inference, while the RFG system offers faster reasoning speed but tends to hallucinate. Our method achieved a good trade-off between faithfulness and inference efficiency.

1 Introduction

The faithfulness of the system response is crucial when evaluating Language Models (LM) powered dialogue systems (Adiwardana et al., 2020). A faithful system means the system response is consistent with the appropriate knowledge. However, an unfaithful system will face the well-known 'hallucination' problem (Maynez et al., 2020).

One effective technique to improve faithfulness and reduce hallucination of the dialogue system is **Retrieval Augmented Generation (RAG)** (Jiang et al., 2023b; Zhao et al., 2023), which retrieves passages from a knowledge base to augment the response. However, the retrieval process takes several times longer than the generation process, leading to severe inference inefficiency (Thulke et al., 2023).

A straightforward approach to improve inference efficiency is to use the **Retrieval Free Generation (RFG)**, which discards the retrieval process and directly utilizes the knowledge injected in its parameters (Brown et al., 2020) to generate a response. This distinction makes it more challenging for the retrieval-free model to integrate correct knowledge into responses. To address this challenge, Xu et al. (2023) introduced a novel knowledge injection pretraining scheme. Xu et al. (2022) stored knowledge in multiple adapters, and Sun et al. (2023) proposed mixed contrastive learning to improve faithfulness.

However, previous RFG methods exclusively use human-labeled responses as the model training target, which are natural and fluent but contain limited knowledge tokens. These labels will lead the model to imitate the response style and generate a fluent response, yet ignore learning how to fuse necessary knowledge into the output. In the WoW dataset, the SOTA RFG's faithfulness is still 25% lower than RAG's (Sun et al., 2023). Thus, a key question arises: Can we simultaneously ensure the dialogue system's efficiency and faithfulness?

To address this problem, we propose a training scheme named Retrieval Augmented to Retrieval Free Distillation (RA2FD) that utilizes a teacher-student framework to build a retrieval-free model to achieve higher faithfulness and maintain inference efficiency. We distill the capacity to generate faithful responses from an oracle RAG teacher's response to an RFG student model through sequence-level distillation. The core idea is these responses are more knowledge-infused than human-labeled responses. Furthermore, to fully use the teacher's generation capacity, we let the teacher generate multiple knowledge-infused responses instead of only one. We then employ contrastive objectives to let the student model focus more on learning from a more faithful knowledge-infused response.

We conduct our experiments on three knowledge-grounded dialogue benchmarks. A task-oriented dialogue called DSTC9 (Kim et al., 2018) and two open-domain chatbots named WoW (Dinan et al., 2019) and FaithDial (Dziri et al., 2022). Our method achieves faithfulness improvements by an average of 33% to the previous SOTA RFG baseline. It also boosts inference speed by 50 and 2 times compared to the RAG methods on the DSTC9 and WoW/FaithDial datasets. In summary, we contribute to improving the faithfulness of the retrieval-free generation model from three aspects:

- We introduce a teacher-student framework to build a faithful and efficient RFG model. This model-agnostic framework can be directly applied to fine-tune large language models.
- In the framework, we use sequence-level distillation to distill the faithfulness capacity from multiple knowledge-infused responses generated by an oracle RAG teacher to an RFG student. We use contrastive objectives to ensure it learns from a more faithful response.
- Our method allows a retrieval-free model to achieve a new SOTA faithfulness performance on three knowledge-grounded dialogue benchmarks that match an RAG method while significantly improving inference efficiency.

2 Related Work

The open-domain chatbot (Huang et al., 2020) and task-oriented dialogue system (Zhang et al., 2020) that generates a response based on knowledge has received attention recently.

Unfaithfulness in LM Generation Unfaithfulness, which includes the hallucinations (Ji et al., 2023) phenomenon, is the response generated by an LM-based dialogue system that is inconsistent or unfaithful (Zhou et al., 2021; Filippova, 2020) to the appropriate knowledge. Training data is essential to the unfaithfulness problem in the LM-based dialogue system. Shen et al. (2021) filtered out untrustworthy samples from the training set, and Dziri et al. (2022) removed hallucinations in the Wizard of Wikipedia (WoW) dataset.

Retrieval Augmented Generation Open-domain chatbots use retrieve-based methods (Karpukhin et al., 2020; Eric et al., 2021; Li et al., 2022a; Shuster et al., 2021) to alleviate unfaithfulness of generation by integrating external knowledge (such as Wikipedia) (Zhao et al., 2023; Jiang et al., 2023b) as the input context. Meanwhile, for the task-oriented dialogue system that limits its external knowledge to a specific document or knowledge graph, the retriever (He et al., 2024; Rony et al., 2022; Kim et al., 2018) slightly diverges from the chatbot. However, though retrieval enriches the response information, such methods suffer from severe inference inefficiency (Thulke et al., 2023).

Retrieval Free Generation One way to overcome this drawback is to omit the retrieval process and use knowledge stored in parameters to generate responses. Sun et al. (2023) used mixed contrastive learning to enhance the knowledge elicitation process. Diao et al. (2023); Bang et al. (2023); Emelin et al. (2022) injected domain knowledge into the adapter while fixing the pre-trained language model (PLM). Instead of storing knowledge in multiple adapters, Xu et al. (2023); Li et al. (2022b) injected external knowledge into the PLM parameters. To better probe knowledge in PLMs, Liu et al. (2022) employed a multi-stage prompting approach in the open-domain chatbot. However, the faithfulness performance of existing retrieval-free methods is still far from satisfactory.

This paper introduces a novel teacher-student framework to build a retrieval-free dialogue model with higher faithfulness and inference efficiency. Previous retrieval-free methods exclusively use human-labeled responses as the training target, which are fluent but contain limited knowledge contexts. These training targets will lead the model to imitate the response style yet ignore learning how to fuse necessary knowledge context into the output. Unlike previous works, we distill the capacity to generate faithful responses from the retrieval-augmented model to the retrieval-free model by training a retrieval-free model using the multiple knowledge-infused responses generated by an oracle retrieval-augmented teacher, which are more knowledge-infused than human-labeled responses.

3 Methodology

Figure 2 presents the overview of our method. We first use an oracle retrieval-augmented teacher model to generate multiple knowledge-infused responses. Then, we distill the capacity to generate faithful responses from these responses to a retrieval-free student model through sequence-level distillation and contrastive learning.

To keep the notation consistent with our method, let $U_t = \{u_{t-w+1}, \cdots, u_{t-1}, u_t\}$ represent the dialogue history with a window size of w turns, and t is the index of each turn. u_t is the current user utterance. The knowledge-based dialogue system is designed to generate an informative response u_{t+1} using U_t and a knowledge snippet with n tokens $K = \{k_1, \cdots, k_n\}$.

3.1 Teacher Model Training

We employ an oracle Retrieval Augmented Generation (RAG) (Shuster et al., 2021) model as a teacher to improve the faithfulness of a Retrieval Free Generation (RFG) student model.

The oracle teacher model learns to predict the ground truth response u_{t+1} given the dialogue context U_t and the **ground truth knowledge** K. We let the loss of Maximum Likelihood Estimation (MLE) be the training loss of the teacher model.

$$\mathcal{L}_{\mathrm{MLE}} = -\sum_{i=1}^{|u_{t+1}|} \log p_\theta \left(w_i \mid w_{<i}, U_t, K\right), \quad (1)$$

where w_i is the i-th token of u_{t+1} and θ is the parameters of the teacher model. We perform teacher model inference on the training set to obtain the knowledge-infused teacher responses through auto-regressive response generation:

$$\begin{aligned} P\left(\hat{u}_{t+1}\right) &= p_\theta\left(\hat{u}_{t+1} \mid U_t, K\right) \\ &= \prod_{i=1}^{|\hat{u}_{t+1}|} p_\theta\left(\hat{w}_i \mid \hat{w}_{<i}, U_t, K\right), \end{aligned} \quad (2)$$

where $\hat{u}_{t+1}$ is the predicted response generated by the teacher model on the training set and $\hat{w}_i$ is the i-th token of $\hat{u}_{t+1}$.

3.2 Knowledge Injection

We build a retrieval-free dialogue system that begins with injecting knowledge into model parameters. The external knowledge of the DSTC9 dataset is the Frequently Asked Questions (FAQ) about the domains and the entities mentioned in the corpus. The external knowledge K can thus be further split into question $K_Q = \{q_1, \cdots, q_i\}$ and answer $K_A = \{a_1, \cdots, a_j\}$ with $K = \{K_Q, K_A\}$. We inject external knowledge by fine-tuning a language model on the FAQ corpus using an MLE loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{IN}} &= -\log p_\phi\left(K_A \mid K_Q\right) \\ &= -\sum_{t=1}^{j} \log p_\phi\left(a_t \mid a_{<t}, K_Q\right), \end{aligned} \quad (3)$$

where ϕ is the parameters of the retrieval-free generation model, and the model learns to predict the knowledge tokens for each step in a teacher-forcing (Williams and Zipser, 1989) paradigm.

For the Wizard of Wikipedia (WoW) and FaithDial datasets, we directly take the pre-trained language model as the Wikipedia knowledge-injected model since Wikipedia is a commonly used corpus in the language model's pre-training.

3.3 Sequence Level Distillation

Although the model can remember the external knowledge to some extent after knowledge injection, its faithfulness performance is still far from satisfactory.

To further enhance our retrieval-free generation model, we utilize the teacher-generated knowledge-infused label $\hat{u}_{t+1}$ as a training reference instead of using the ground truth label u_{t+1} (Kim and Rush, 2016). Concretely, we continue to fine-tune the student model's parameters ϕ by minimizing the following NLL loss:

$$\mathcal{L}_{\mathrm{NLL}} = -\sum_{i=1}^{|\hat{u}_{t+1}|} \log p_\phi\left(\hat{w}_i \mid \hat{w}_{<i}, U_t\right), \quad (4)$$

Unlike the teacher model, the retrieval-free student model does not require explicit knowledge input during model training and inference.

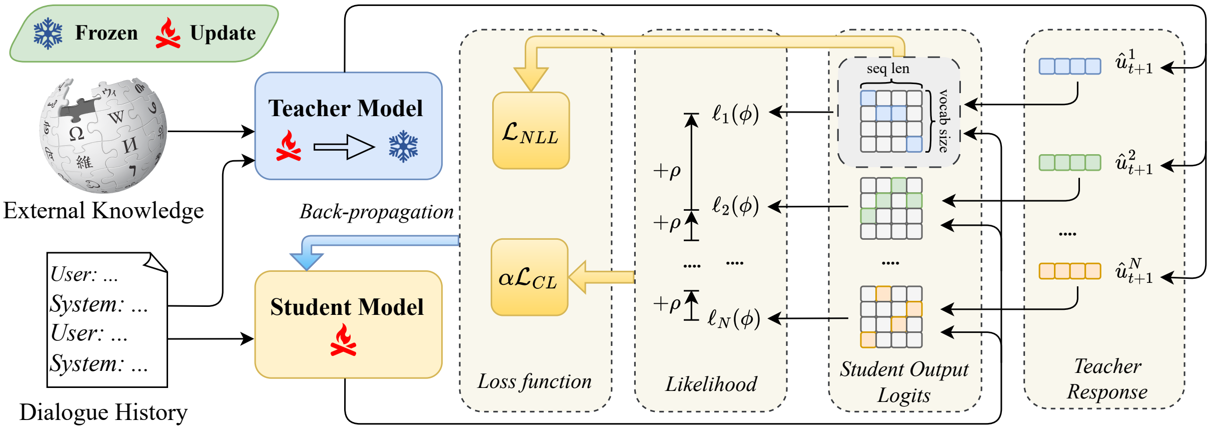

Figure 2: The oracle retrieval augmented teacher generates multiple knowledge-infused responses with ground truth knowledge input (Section 3.1). The student model first injects external knowledge into model parameters (Section 3.2), then performs distillation with NLL and CL loss using the teacher-generated labels (Section 3.3 and 3.4).

3.4 Multi-Label Contrastive Learning

Instead of letting the teacher model generate only one response, we let the teacher model generate multiple responses to fully utilize the teacher model's generation ability. Furthermore, we propose multi-label contrastive learning to improve the faithfulness of the student model.

Let $Y_T = \left\{\hat{u}^1_{t+1}, \hat{u}^2_{t+1}, \cdots, \hat{u}^M_{t+1}\right\}$ be M different labels the teacher model generates when performing beam-search on the training set. These M labels are ranked in descending order based on their total scores of fluency and faithfulness described in section 4.2. The prediction log-likelihood of teacher label $\hat{u}^i_{t+1}$ with L_i length is:

$$\ell_i(\phi) = \frac{1}{L_i} \sum_{j=1}^{|\hat{u}^i_{t+1}|} \log p_\phi \left(\hat{w}^i_j \mid \hat{w}^i_{<j}, U_t\right). \quad (5)$$

We encourage the model's prediction likelihood of a higher score label to be larger than the lower score label. To further enhance the student model to generate superior responses, we define a contrastive learning object for student model training:

$$\begin{aligned}\mathcal{L}_{\mathrm{CL}} = &\frac{1}{M} \sum_{m=1}^{M} \sum_{n=m+1}^{M} \\ &\max\left\{0, \rho - (\ell_m(\phi) - \ell_n(\phi))\right\},\end{aligned} \quad (6)$$

ρ is a pre-defined margin. The overall retrieval-free student model fine-tuning objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \alpha \mathcal{L}_{\mathrm{CL}}, \quad (7)$$

where the hyper-parameter $\alpha \in [0, 1]$ regulates the importance of these two components.

4 Experiments

4.1 Dataset

We conduct our experiments on the following three knowledge-grounded dialogue datasets:

DSTC9: A task-oriented conversation dataset in customer service scenarios. DSTC9 contains 9,167 conversations and 23,838 utterance pairs. These were newly collected (Kim et al., 2018) based on 9,139 knowledge candidates from FAQ web pages about the domains and entities in the original MultiWOZ2.1 (Eric et al., 2020) databases.

WoW: WoW is a commonly used open-domain knowledge-grounded dialogue based on Wikipedia. It involves two speakers, a knowledgeable wizard and an inquisitive apprentice, who start to discuss an initial topic. The dataset comprises 22,311 conversations with 201,999 turns. The test set includes 'Seen' and 'Unseen' to assess the model performance on familiar and new topics.

FaithDial: FaithDial (Dziri et al., 2022) is built based on the WoW dataset, which uses a data-centric method to revise the response in the original dataset to be more faithful and creative. The FaithDial contains 5,649 dialogues consisting of 50,761 utterances, and each dialogue uses the same knowledge candidate pool as the WoW dataset.

4.2 Evaluation Metrics

This paper uses automatic metrics to evaluate fluency and faithfulness. We also perform a turn-level human evaluation to investigate system responses generated by different methods.

Fluency: We employ widely used text generation measures, including **BLEU** (Papineni et al.,

DSTC9		Fluency			Faithfulness		Size
Method	Model	BLEU ↑	METEOR ↑	ROUGE-L ↑	KF1 ↑	BERTScore ↑	
RAG	BART	16.46	22.35	35.67	48.20	89.83	523M
	Llama 2	17.58	22.98	37.60	44.58	89.18	7.1B
	Mistral	17.79	23.00	37.72	44.64	89.18	7.1B
RFG	BART	15.77	21.66	35.21	34.22	87.42	406M
	Llama 2	17.03	22.58	37.12	35.64	87.66	7B
	Mistral	17.17	22.14	36.73	34.88	87.46	7B
RFG (RA2FD+)	BART	15.66	22.38	35.92	43.48	89.08	406M
	Llama 2	<u>18.24</u>	**23.82**	**39.03**	<u>46.99</u>	<u>89.69</u>	7B
	Mistral	**18.60**	<u>23.70</u>	<u>38.84</u>	**48.27**	**89.92**	7B

Table 1: Evaluation results of the RAG and RFG methods on the DSTC9 dataset. We highlight the best results with **boldface** and <u>underline</u> the second-best result. Our proposed RA2FD outperforms all RFG baselines by a substantial margin in all metrics, boosts the KF1 score of fine-tuned models (i.e., Method: RFG) by an average of 32.41%, and even outperforms the best-performing RAG-BART.

2002), **ROUGE** (Lin, 2004), and **METEOR** (Denkowski and Lavie, 2014), to evaluate the fluency of the model generations compared to ground-truth human responses.

Faithfulness: To assess the faithfulness of the generated response, we use **KnowledgeF1** (KF1) (Shuster et al., 2021). KF1 measures the uni-gram word overlap between the generated response and the external knowledge that the human relied on during data collection. We also use **BERTScore** to measure the semantic (Zhang* et al., 2020) similarity between a response and knowledge.

4.3 Baselines Details

I). Retrieval Free Generation (RFG): RFG generates system response by leveraging the implicit knowledge within its parameters

We use Llama 2-7b (Touvron et al., 2023), Mistral-7b (Jiang et al., 2023a), BART-Large (Lewis et al., 2020), KnowExpert (Xu et al., 2022), and the previous SOTA method in the WoW dataset, MixCL (Sun et al., 2023), as RFG baseline. Appendix A.2 shows the details of these methods.

II). Retrieval Augmented Generation (RAG): The RAG first retrieves knowledge snippiest from a knowledge base and incorporates it with the dialogue context to generate responses. We fine-tune Llama 2-7b, Mistral-7b, and BART-Large on three downstream tasks as our generation models.

The DSTC9 dataset's knowledge base consists of 12,039 FAQs about the domains and entities mentioned in the corpus. The knowledge candidates for each turn of the WoW and FaithDial datasets are about 70 Wikipedia abstracts relevant to the wizard and apprentice discussion topic.

III). We use a cross-encoder-based retriever (Thulke et al., 2023) detailed in Appendix A.3 to enhance different language models to generate system responses on DSTC9, WoW, and FaithDial benchmarks. We also compare RA2FD with RAG methods equipping different retrievers in Appendix A.4.

5 Experimental Results

5.1 Evaluation on DSTC9

Table 1 displays the automatic evaluation results on the DSTC9 dataset. Language models in the first section (Method: RAG) employ the same knowledge retriever with a capable retrieval accuracy: 68.75% on R@1, 90.25% on R@5, and 77.64% on MRR@5. From Table 1, we can deduce that:

I) The retriever (Method: RAG) boosts the KF1 score of all fine-tuned language model baselines (Method: RFG) by 31.30% on average and further enhances generation fluency across all metrics;

II) Compared to RA2FD's corresponding counterpart (e.g., RA2FD + Llama2 vs. Llama2 in RFG Method), all fine-tuned language models obtain a considerable faithfulness improvement of the response (32.41% higher KF1 score on average);

III) The KF1 performance of RA2FD is close to the leading RAG method (i.e., RA2FD + Mistral vs. BART in the RAG Method). Furthermore, RA2FD enables a pre-trained language model to excel in METEOR and ROUGE-L scores;

IV) RA2FD's retrieval-free architecture saves about 117M parameters compared to its corresponding counterpart in model size (e.g., RA2FD + Llama2 vs. Llama2 in the RAG Method).

WoW		Test seen				Test unseen				Size
		Fluency		Faithfulness		Fluency		Faithfulness		
Method	Model	B4 ↑	MT ↑	KF1 ↑	BT ↑	B4 ↑	MT ↑	KF1 ↑	BT ↑	
RAG	BART	6.28	9.68	29.88	85.54	6.01	9.62	30.83	85.54	531M
	Llama 2	6.32	11.02	30.02	85.56	6.10	10.81	30.97	85.56	7.1B
	Mistral	6.75	11.08	29.63	85.58	6.70	11.02	30.79	85.63	7.1B
RFG	BART	4.43	8.57	18.42	83.36	2.47	7.90	14.87	82.73	406M
	Llama 2	4.85	10.00	19.19	83.56	**3.86**	9.06	15.58	82.86	7B
	Mistral	5.04	10.11	18.56	83.35	3.46	9.12	14.65	82.50	7B
	MixCL	2.70	*20.50**	22.30	\	1.40	*18.00**	18.00	\	406M
	KnowExpert	3.19	8.05	13.68	82.38	2.06	7.14	11.45	81.75	117M
RFG (RA2FD+)	BART	**5.48**	8.74	26.53	84.76	1.73	7.27	18.04	82.89	406M
	Llama 2	5.18	9.75	**27.85**	**84.95**	3.57	**9.27**	**23.21**	**83.79**	7B
	Mistral	5.25	**10.30**	27.50	84.87	3.50	9.25	21.43	83.56	7B

Table 2: Evaluation results of the RAG and RFG methods on the WoW dataset. We highlight the best results with **boldface** and underline the second-best result. Our proposed RA2FD outperforms all RFG baselines in fluency and faithfulness of model response by a substantial margin and achieves a 24.89% higher KF1 score than the previous SOTA baseline, MixCL. The proposed RA2FD even approaches the best RAG-Llama 2.

DSTC9	BART	Llama2	Mistral
Base	15.77 / 34.22	17.03 / 35.64	17.17 / 34.88
$+\mathcal{L}_{\text{IN}}$	**15.91** / 36.60	17.69 / 38.28	17.17 / 37.50
$+\mathcal{L}_{\text{NLL}}$	15.57 / 41.44	17.16 / 43.85	18.15 / 45.71
$+\mathcal{L}_{\text{CL}}$	15.66 / **43.48**	**18.24 / 46.97**	**18.60 / 48.27**

Table 3: Study of model-agnostic of RA2FD on the DSTC9 dataset, indicated by the BLEU and KF1 scores. Revealing that the proposed method enhances the faithfulness of generated responses by an average of 32.41% while preserving fluency in task-oriented dialogues.

WoW	BART	Llama2	Mistral
Base	4.43 / 18.42	4.85 / 19.19	5.04 / 18.56
$+\mathcal{L}_{\text{NLL}}$	4.77 / 24.99	5.16 / 25.84	5.24 / 26.35
$+\mathcal{L}_{\text{CL}}$	**5.48 / 26.53**	**5.18 / 27.85**	**5.25 / 27.50**

Table 4: The ablation study conducted on the WoW dataset highlights the model-agnostic characteristics of RA2FD. It enhances the generation's faithfulness by an average of 45.77%. Our method also improves fluency in the response generated by the model.

5.2 Evaluation on WoW and FaithDial

Table 2 [1] and Table 5 show the automatic evaluation results of the WoW and FaithDial datasets. Three retrieval-augmented methods in the first section (Method: RAG) of both tables utilize the same retriever with the same retrieval accuracy of 25.02% on R@1 and 59.88% on R@5 due to the same knowledge pool used in FaithDial and WoW.

Note that although the retrieval accuracy of 25.02% on R@1 is much lower than the R@1 of 68.78% in the task-oriented dataset DSTC9, it is still proven to be outstanding (Kim et al., 2020). In comparison, the human-level accuracy is only 17.10% on R@1 in this open-domain retrieval task.

On the seen test set of the WoW and FaithDial, RA2FD surpasses all previous RFG baselines in fluency and faithfulness. Specifically, RA2FD + Llama2 achieves a 24.89% higher KF1 score than the previous SOTA baseline, MixCL, in the WoW dataset. It only falls slightly behind the KF1 score achieved by the top-performing RAG method (i.e., RA2FD + Llama2 vs. Llama2 in the RAG Method). Although the performance of RA2FD dipped in the unseen test set of WoW, it still achieved a SOTA result in faithfulness and fluency.

[1] Unlike the toolkit (Lavie and Agarwal, 2007) used in MixCL, we use the latest METEOR implementation (Denkowski and Lavie, 2014), which results in an incomparable result *.

5.3 Ablation Studies

Model-agnostic of RA2FD: Our proposed RA2FD is compatible with fine-tuning various pre-trained language models. Table 3 and Table 4 compare RA2FD with several of its ablative variants.

Base: A pre-trained language model fine-tuned on the DSTC9 or WoW dataset, equivalent to the baseline (i.e., BART, Llama 2, and Mistral) in the RFG part of Table 1 and Table 2.

FaithDial		Test seen				Test unseen				Size
		Fluency		Faithfulness		Fluency		Faithfulness		
Method	Model	B4 ↑	MT ↑	KF1 ↑	BT ↑	B4 ↑	MT ↑	KF1 ↑	BT ↑	
RAG	BART	6.86	12.46	24.30	84.63	7.40	13.02	24.75	84.52	531M
	Llama 2	7.11	13.31	25.26	84.80	7.39	13.59	25.76	84.71	7.1B
	Mistral	6.81	13.03	25.10	84.75	7.44	13.50	25.53	84.68	7.1B
RFG	BART	5.20	11.34	13.00	82.57	4.79	11.01	12.13	82.30	406M
	Llama 2	**6.10**	12.10	16.90	83.16	5.39	11.75	16.37	82.99	7B
	Mistral	5.78	11.67	15.60	82.93	**5.64**	11.71	15.89	82.87	7B
RFG (RA2FD+)	BART	5.55	11.58	18.45	83.47	4.81	10.73	14.63	82.59	406M
	Llama 2	5.78	**12.77**	**23.39**	**84.29**	5.62	12.09	20.02	83.49	7B
	Mistral	5.85	12.47	22.38	84.09	5.29	**12.40**	**20.83**	**83.61**	7B

Table 5: Evaluation results of RAG and RFG methods on the FaithDail dataset. We highlight the best results in **boldface** and underline the second-best result. The proposed RA2FD also effectively improves the faithfulness of a fine-tuned language model on the FaithDail dataset. The retriever's performance is identical to the WoW dataset due to the same knowledge pool used in FaithDial and WoW.

$+\mathcal{L}_{\text{IN}}$: A knowledge-injected version (Section 3.2) of the original pre-trained language model. We only inject task-oriented knowledge of the DSTC9 dataset and directly utilize the pre-trained Wikipedia knowledge in PLM for the WoW and FaithDial datasets. With the knowledge injected into model parameters, all basic PLM enhance their generation capabilities regarding fluency and faithfulness on the DSTC9 dataset.

$+\mathcal{L}_{\text{NLL}}$: We perform sequence-level distillation (Section 3.3) on the knowledge-injected model ($+\mathcal{L}_{\text{IN}}$) when the response generated by the teacher model is only one (i.e., $M = 1$). Compared with the original label (**Base**), using a teacher-generated knowledge-infused response based on a knowledge-injected model can improve the KF1 score by an average of 23.41% and 45.77% on the DSTC9 and WoW datasets, respectively. This part of the approach **contributes the most** to the overall performance gains in our proposed approach.

$+\mathcal{L}_{\text{CL}}$: Instead of generating only one response, we let the teacher output multiple knowledge-infused responses and use these responses in model distillation (Section 3.4). With multiple teacher labels, the knowledge-injected language model further improves its faithfulness in a generation.

In summary, I) RA2FD is model-agnostic and compatible with fine-tuning various pre-trained language models, and II) RA2FD can considerably enhance a dialogue system's fluency and faithfulness.

Faithfulness V.S. Number of labels: To explore how the quantity of labels generated by the teacher influences the faithfulness of the student model, we adjust the label count from 1 to 7 and plot the corresponding performance trends. Our findings in Figure 3 show a significant enhancement in the faithfulness of the student retrieval-free generation model as the label count increases from 1 to 3 for all three datasets. The improvements converge when the label count exceeds 5. Consequently, we use five teacher-generated responses in our method.

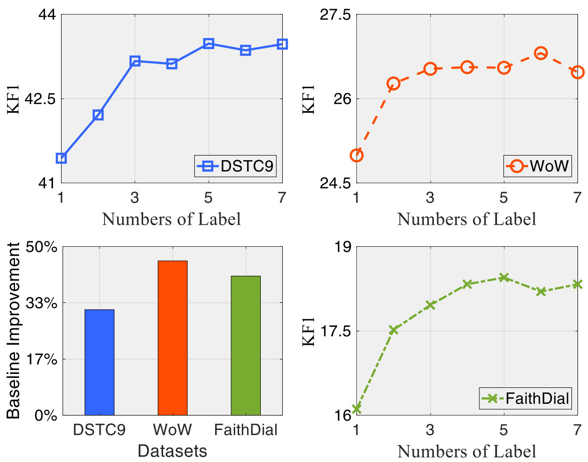

Figure 3: The performance of RA2FD + BART tends to converge when the number of labels is larger than 5. The left bottom shows the average improvement against the fine-tuned language model (i.e., Method: RFG).

5.4 Efficiency Analysis

Since the FaithDial dataset uses the same knowledge pool as the WoW dataset, we only compared the inference efficiency of retrieval-augmented and retrieval-free generation methods on the DSTC9 and WoW datasets.

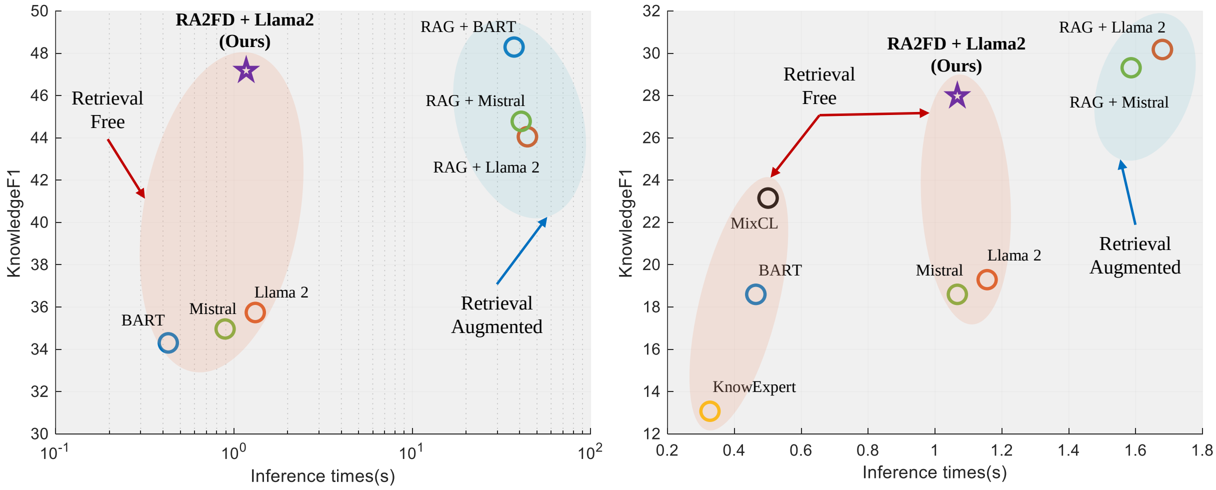

Figure 4: We investigate the correlation between inference latency (in seconds) per sample and the faithfulness of the response on the DSTC9 (left, in **logarithmic scale**) and WoW (right) datasets. The previous retrieval-free generation method boasts rapid inference speeds but experiences a severe decline in performance. In contrast, our proposed RA2FD performs on par with retrieval-augmented approaches while offering swift reasoning speed.

	Faithfulness			Humanness		
	DSTC9	WoW	FaithD	DSTC9	WoW	FaithD
BART	2.68	1.96	1.84	3.32	3.56	3.48
Llama 2	2.76	2.01	2.04	3.43	3.58	**3.59**
Mistral	2.72	1.98	2.04	3.48	3.60	3.52
RA2FD	**3.06**	**2.24**	**2.26**	**3.52**	**3.62**	3.56

Table 6: Our proposed RA2FD + Llama 2 (RA2FD) achieves the best performance in faithfulness without compromising humanness in the human evaluation.

Appendix A.5 thoroughly analyzes the computational resources required for training the RAG and the proposed RA2FD method.

As outlined in Figure 4, while retrieval does enhance the faithfulness of the language model's generation (i.e., retriever + Llama 2 vs. Llama 2), it notably increases inference latency, especially when pulling knowledge from an extensive knowledge base. Notably, the latency of inference times on the DSTC9 dataset is on a logarithmic scale since each model inference requires knowledge retrieval from a database containing 12,039 entries. Thus, the retriever takes roughly 50 times longer than the generation process at each inference time.

The time spent retrieving the WoW dataset is approximately double the generation time, with each retrieval set at around 70 candidate knowledge entries. Compared to the retrieval-augmented method, our proposed RA2FD performs on par with retrieval-augmented methods and offers faster reasoning speed for both datasets.

5.5 Human Evaluation

We randomly select 100 dialogues from each test set of three datasets for evaluation. We provided five master-level annotators with the dialog context, the model response, and the associated knowledge. Annotators assign **Faithfulness** scores ({1: bad}, {3: moderate}, {5: perfect}) to evaluate the alignment of the generated response with the given knowledge. They also assign **Humanness** scores ({1: bad}, {3: moderate}, {5: perfect}) to assess fluency and naturalness.

Table 6 presents the results of human evaluations for four distinct methods of retrieval-free generation. The discrepancy in faithfulness between the DSTC9 and WoW/FaithDial datasets demonstrates the inherently open-ended nature of the conversation process in open-domain chatbots. Notably, our approach performs best regarding faithfulness in both datasets without compromising fluency.

6 Conclusions

In this paper, we proposed a Retrieval Augmented to Retrieval Free Distillation (RA2FD) training scheme to improve the faithfulness of the retrieval-free dialogue generation model. Extensive experiments conducted on the DSTC9, WoW, and FaithDial datasets demonstrate that RA2FD outperforms existing retrieval-free generation methods and achieves a state-of-the-art result on all datasets. Moreover, the faithfulness of our proposed method is comparable to retrieval-augmented generation methods while offering a faster inference speed.

Limitations

Since the proposed method relies on a retrieval-augmented teacher model to help improve the faithfulness of the retrieval-free student model, a dialogue dataset that pairs with external knowledge is required. Thus, an efficient knowledge update method for the proposed method is required when the external knowledge changes, potentially becoming our future research direction.

Ethical Considerations

All the pre-trained language models used in our paper are downloaded from the Huggingface publicly released model card, and we strictly follow the user license. Our study conducts all experiments using publicly available datasets and strictly follows their usage terms to sidestep any ethical issues. Although the method proposed in this paper significantly improves the faithfulness of a retrieval-free generation model, there is still a risk of potential misuse. For example, when given misleading information as input, dialogue systems may spread misinformation. Thus, adding harmful information detection to the retrieval-augmented and retrieval-free dialogue system is necessary for practical use.

References

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

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-domain-adapters: Decoupling and injecting domain knowledge to pre-trained language models' memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*.

Denis Emelin, Daniele Bonadiman, Sawsan Alqahtani, Yi Zhang, and Saab Mansour. 2022. Injecting domain knowledge in language models for task-oriented dialogue systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates.

Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. 2021. Multi-sentence knowledge selection in open-domain dialogue. In *Proceedings of the 14th International Conference on Natural Language Generation*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Huang He, Hua Lu, Siqi Bao, Fan Wang, Hua Wu, Zheng-Yu Niu, and Haifeng Wang. 2024. Learning to select external knowledge with multi-scale negative sampling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:714–720.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

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)*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2018. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 278–289.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022a. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yanyang Li, Jianqiao Zhao, Michael Lyu, and Liwei Wang. 2022b. Eliciting knowledge from large pre-trained models for unsupervised knowledge-grounded conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized pretraining approach.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

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

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. DialoKG: Knowledge-structure aware task-oriented dialogue generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

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.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth*

Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2023. Task-oriented document-grounded dialog systems by hltpr@ rwth for dstc9 and dstc10. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, Dublin, Ireland.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tur. 2023. KILM: Knowledge injection into encoder-decoder language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. “what do others think?”: Task-oriented conversational modeling with subjective knowledge. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Dialogue history
. . .
User: let's do the uhhh regular holiday inn fisherman's wharf uhhh can you give me the zip code.
Bot: ok. sure that zip code is nine four one three three.
User: ok uhhh do they speak french?
Ground Truth Knowledge Question: Do you provide multi-lingual services? Answer: Danish, English, Spanish, French, Italian, Chinese, and Japanese are spoken at Holiday Inn Fisherman's Wharf.
BART: Yes, they do speak French at the Holiday Inn Express. Would you like to make a reservation?
RA2FD + BART: The languages spoken at the Holiday Inn Express are English, Spanish, French, and Chinese. Can I help you with anything else?

Table 7: An example of model outputs on the DSTC9 test set. Our approach accurately lists all the languages the Holiday Inn Express hotel supports. In contrast, the baseline method provides incomplete information or presents incorrect details.

Dialogue history
. . .
Bot: I love toga party's my Halloween party was based off Greek costumes for this theme.
User: I have never heard about this before. Can you explain to me what it means. Toga party!
Bot: Really it is very awesome toga party's is based solely off games costumes etc of Roman or Greek themes.
User: I see. Why do you chose so much interest in it?
Ground Truth Knowledge Toga parties held by college or university students are associated with keg parties and excessive drinking.
BART: I just like toga parties because the theme is based off the fact that the togas or togas allowed women to cover their bodies in a style similar to what was seen in ancient Rome and Greece.
RA2FD + BART: I like toga parties, because toga parties are associated with keg parties and excessive drinking.

Table 8: An example of model outputs on the WoW test set. Our approach delivers precise background knowledge about toga parties, whereas other methods yield dull responses or offer information containing wrong information.

A Appendix

A.1 Case Study

We provide examples of model output in Table 7 and Table 8. The information provided only in our method is highlighted in red, demonstrating that our proposed method improves the faithfulness of the retrieval-free generation model.

A.2 Baseline and Implementation Details

This subsequent section details the baseline methods featured in our primary results. Llama 2-7b, Mistral-7b, BART-Large are general pre-trained language models. In contrast, MixCL and KnowExpert are specialized retrieval-free dialogue generation models designed for the WoW dataset.

Llama 2-7b (Touvron et al., 2023): A most popular open-access large language model trained on 2 trillion tokens with much longer context length.

Mistral-7b (Jiang et al., 2023a): A widely used large language model, which outperforms Llama 2-7b in various benchmarks.

BART-Large (Lewis et al., 2020): An encoder-decoder transformer (Vaswani et al., 2017) pre-trained on a large corpus including Wikipedia using de-noising techniques.

MixCL (Sun et al., 2023): A BART-Large-based model using a novel mixed contrastive objective to optimize the implicit knowledge elicitation process of the language model.

KnowExpert (Xu et al., 2022): It employs a mixed-adapter architecture to adapt the GPT-2 model to open-domain dialogues.

We describe the training and Inference configurations used in our experiments as follows:

I) Training: We trained all methods in our paper using NVIDIA A100 GPUs, evaluating every epoch until the model outputs on the validation set have no further improvement in evaluation metrics. We use a batch size of 64 and the AdamW (Loshchilov and Hutter, 2018) optimizer with a linear learning rate (LR) scheduler.

We fine-tune BART on the WoW and FaithDial datasets for 20 and 10 epochs with a learning rate of 5e-5 and 7e-6. For the DSTC9 dataset, we fine-tune BART with an LR of 3e-6 for 40 epochs.

For large language models, Llama 2-7b and Mistral-7b, we adopt LoRA (Hu et al., 2022) to fine-tune Llama 2 and Mistral on three downstream datasets for 10 epochs with a learning rate of 1e-4. Each method's detailed training time cost and memory usage are shown in Table 10 and Table 11.

The hyper-parameters selection process on the DSTC9 dataset is described as follows: We first set α to 1 and ρ to 0 and change M to find the best value of 5 for M. Then, we fix M to 5 and α to 1 and change ρ to find the best value 6 for ρ. Finally, we fix M to 5 and ρ to 6 and vary α to find the best value of 0.5 for α. Figure 3 in our paper depicts the effect of M on our method's performance when set

Dataset	Method	Fluency			Faithfulness		Latency	Accuracy
		B4	MT	R-L	KF1	BT	Time(s)	R@1
DSTC9	BM25 + Llama 2	14.30	19.91	33.48	30.26	86.50	1.60	11.36
	Bi-encoder + Llama 2	14.86	20.40	33.98	33.87	87.14	2.11	23.98
	Cross-encoder + Llama 2	17.58	22.98	37.60	44.58	89.18	36.89	68.75
	RA2FD + Llama 2	18.24	23.82	39.03	46.99	89.69	1.11	\
WoW	BM25 + Llama 2	2.50	8.36	14.98	14.65	82.41	1.07	4.59
	Bi-encoder + Llama 2	3.23	9.07	16.46	18.59	83.31	1.48	8.98
	Cross-encoder + Llama 2	6.32	11.02	19.50	30.02	85.56	1.60	25.02
	RA2FD + Llama 2	5.18	9.75	17.20	27.85	84.95	1.07	\

Table 9: We use the RAG method using different retrievers to compare our proposed RA2FD against the retrieval augment dialogue generation method. RA2FD consistently outperforms the BM-25 and Bi-encoder-based RAG methods across all fluency and faithfulness metrics and achieves the fastest inference speeds.

ρ to 6 and α to 0.5. A similar selection process is performed for other datasets.

In summary, for the hyper-parameters used in our method, we set parameter α to 0.1 for the WoW and FaithDial datasets and 0.5 for the DSTC9 dataset. The number of teacher-generated labels M is set to 5 for all three datasets, and the margin ρ is set to 6.

II) Inference: For the inference results on the test and validation set, we employ beam search with a max sequence length of 60 tokens and a beam width of 5.

A.3 Cross-Encoder Retriever Implementation

The cross-encoder-based knowledge retriever used a neural network to distinguish knowledge snippets from a knowledge base.

We first randomly sample $C-1$ knowledge snippets from the knowledge base as negative candidates for model training. The negative candidates, along with the ground truth knowledge, can be denoted as $S = \{K_1, K_2, \cdots, K_g, \cdots, K_C\}$, where g is the index of the ground truth knowledge.

Then the dialogue history U_t is contacted with each knowledge candidates to construct C history-knowledge pairs $\{[U_t, k_1], \cdots, [U_t, k_C]\}$.

These history-knowledge pairs were passed through RoBERTa (Liu et al., 2020) to obtain a sequence-level representation averaged on the last hidden state of each token of the history-knowledge pair, which can be written as $H_u = [\mathbf{h}_{u,1}, \mathbf{h}_{u,2}, \cdots, \mathbf{h}_{u,C}] \in \mathbb{R}^{d\times C}$, where d is the dimension of sentence level representation vector. Finally, the sequence-level representation is passed through a linear layer $W \in \mathbb{R}^{1\times d}$ to obtain classification distributions p_u.

$$p_u = [p_{u,1}, \cdots, p_{u,C}] = \mathrm{softmax}(WH_u). \quad (8)$$

We use a cross-entropy loss on the classification logits to guide the network to choose the ground truth knowledge in C knowledge candidates, which can be written as:

$$\mathcal{L}_{CE} = -\log\left(p_{u,g}\right). \quad (9)$$

During inference, the selected knowledge snippet can be written as $K_S = \{l_k \mid \arg\max p(l_k \mid W_t), k \in K\}$, where l_k is a knowledge candidate in the knowledge base.

We utilize five negative candidates to train the retriever model on the DSTC9, WoW, and FaithDial datasets. The batch size for fine-tuning the pre-trained model is set to 64. We adopt an AdamW optimizer with a learning rate of 1e-5 and an ϵ of 1e-8, and the total training epoch is set to 10.

A.4 Ablation Study on Retriever

We use the cross-encoder-based retriever to retrieve relevant knowledge from the external knowledge base for better retrieval accuracy.

In this section, we compare our proposed RA2FD with two additional retrievers, BM25 (Robertson et al., 1995) and Bi-encoder-based (Thulke et al., 2023) retriever, to further demonstrate its effectiveness. The "Latency" column in Table 9 represents the time required for inferring one sample with the model, while the "Accuracy" column reflects the retrieval accuracy.

The Bi-encoder calculates the similarity between the OpenAI embedding (text-embedding-3-small)

DSTC9		Time Cost		Memory	
Method	Model	Train/ epoch(s)	Infer/ sample(s)	Train	Infer
\	Retriever	345	35.87	15G	6G
RAG	BART	945	36.33	32G	10G
	Llama 2	1661	36.92	80G	23G
	Mistral	1679	36.85	80G	23G
RFG	BART	917	0.42	15G	4G
	Llama 2	1517	1.02	63G	17G
	Mistral	1505	0.91	63G	17G
RA2FD+ (Ours)	BART	932	0.45	16G	4G
	Llama 2	1520	1.11	65G	17G
	Mistral	1569	0.91	65G	17G

Table 10: Computational resources analysis on the DSTC9 dataset shows that RA2FD uses the same resources as RFG but less than the RAG part. Furthermore, RA2FD significantly outperforms RFG and nearly matches RAG in faithfulness.

WoW		Time Cost		Memory Usage	
Method	Model	Train/ epoch(s)	Infer/ sample(s)	Train	Infer
\	Retriever	907	0.53	15G	6G
RAG	BART	2486	1.04	34G	10G
	Llama 2	4481	1.65	83G	24G
	Mistral	4579	1.58	83G	24G
RFG	BART	2411	0.52	18G	4G
	Llama 2	4257	1.12	65G	18G
	Mistral	4153	1.05	65G	18G
RA2FD+ (Ours)	BART	2451	0.55	18G	4G
	Llama 2	4280	1.07	67G	18G
	Mistral	4128	1.04	67G	18G

Table 11: Computational resources analysis on the WoW dataset indicates the same conclusion as in the DSTC9 dataset. The proposed RA2FD significantly improves the faithfulness of the original pre-trained language model while not requiring additional resources.

[2] of dialogue history and knowledge candidates to select the candidate with the highest similarity score as the retrieved knowledge.

As shown in Table 9, these two retrievers boost the inference speed, especially in the DSTC9 dataset, compared with the cross-encoder-based retriever used in our paper's main results. However, the inference speed and response faithfulness are inferior to our proposed RA2FD method. The main reason is that a retriever with lower retrieval accuracy will feed incorrect external knowledge into the generation model. Given the dialogue history and incorrect knowledge, this wrong information will mislead the dialogue system in generating a faithful response.

Nevertheless, our proposed RA2FD method consistently surpasses the BM-25 and Bi-encoder-based RAG methods across all fluency and faithfulness metrics, achieving the fastest inference speeds. The comparison between various RAG methods with different retrievers and our proposed RA2FD further validates the effectiveness of our proposed method.

A.5 Computational Resources Analysis

This section provides a detailed analysis of computational resources for the methods presented in our main results.

According to the ablation study shown in Table 3 and Table 4, the ablative variants of $+\mathcal{L}_{\text{IN}}$ (i.e., M=1) contribute the most performance gains to our proposed RA2FD method. Thus, in the following analysis, we set the number of responses M generated by the teacher model to one, considering the trade-off between overall performance and computing complexity.

The time cost and memory usage shown in Table 10 and Table 11 demonstrate that our proposed RA2FD training scheme significantly improves the overall performance of the RFG counterpart while not requiring additional computational resources.

It is important to note that RA2FD + Llama2 and RA2FD + Mistral exhibit superior performance to RA2FD + BART due to their larger parameter scales, but their overall effectiveness remains comparable. However, RA2FD + RART requires only half the training and inference time needed by RA2FD + Llama2 and RA2FD + Mistral and consumes just a quarter of the GPU memory usage. Thus, the RA2FD + RART is more cost-effective considering the computational resources and overall performance trade-offs.

[2] https://openai.com/index/new-embedding-models-and-api-updates/