Should We Fine-Tune or RAG?
Evaluating Different Techniques to Adapt LLMs for Dialogue

Simone Alghisi [†] **, Massimo Rizzoli** [†] **, Gabriel Roccabruna,**
Seyed Mahed Mousavi, Giuseppe Riccardi
Signals and Interactive Systems Lab, University of Trento, Italy
{s.alghisi, massimo.rizzoli, giuseppe.riccardi}@unitn.it

Abstract

We study the limitations of Large Language Models (LLMs) for the task of response generation in human-machine dialogue. Several techniques have been proposed in the literature for different dialogue types (e.g., *Open-Domain*). However, the evaluations of these techniques have been limited in terms of base LLMs, dialogue types and evaluation metrics. In this work, we extensively analyze different LLM adaptation techniques when applied to different dialogue types. We have selected two base LLMs, Llama2$_C$ and Mistral$_I$, and four dialogue types *Open-Domain*, *Knowledge-Grounded*, *Task-Oriented*, and *Question Answering*. We evaluate the performance of in-context learning and fine-tuning techniques across datasets selected for each dialogue type. We assess the impact of incorporating external knowledge to ground the generation in both scenarios of Retrieval-Augmented Generation (RAG) and gold knowledge. We adopt consistent evaluation and explainability criteria for automatic metrics and human evaluation protocols. Our analysis shows that there is no universal best-technique for adapting large language models as the efficacy of each technique depends on both the base LLM and the specific type of dialogue. Last but not least, the assessment of the best adaptation technique should include human evaluation to avoid false expectations and outcomes derived from automatic metrics.

1 Introduction

In recent years, Large Language Models (LLMs) have been employed for the task of response generation in human-machine dialogues (Hosseini-Asl et al., 2020a; Izacard and Grave, 2021; Komeili et al., 2022). Such models have been applied to several dialogue types, including Open-Domain Dialogues (i.e. informal conversations about trivial matters), Knowledge-Grounded Dialogues (i.e. conversations with a system that provides factual responses), Task-Oriented Dialogues (i.e. conversations where the system helps a user to achieve a specific goal), and Question Answering (i.e. question-answer exchanges given context).

However, recent studies have shown the shortcomings of LLMs as dialogue model surrogates as they are prone to generate toxic, biased, and irrelevant responses (Zhang et al., 2020; Mousavi et al., 2022, 2023; Lin and Chen, 2023). To adapt LLMs to dialogue types, different techniques have been employed such as in-context learning (Brown et al., 2020; Chen et al., 2023; Meade et al., 2023; Hudeček and Dusek, 2023) and fine-tuning (Wang et al., 2022; Komeili et al., 2022; Huang et al., 2023). Furthermore, strategies such as grounding (Gopalakrishnan et al., 2019; Zhao et al., 2023) and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Borgeaud et al., 2022) have been proposed to improve the generation quality.

Currently, the performance of the aforementioned techniques in adapting LLMs across different dialogue types is understudied. Previous studies have evaluated these techniques in a specific dialogue type only (Raposo et al., 2023; Zhang et al., 2023). Such studies are based on different base models and are assessed via incomparable evaluation methodologies.

In this work, we conduct an extensive study on the efficacy of different techniques to adapt LLMs for multiple dialogue types. We select Llama-2 Chat (Llama2$_C$) (Touvron et al., 2023) and Mistral Instruct (Mistral$_I$) (Jiang et al., 2023) as base LLMs, and experiment with in-context learning and fine-tuning in the context of four dialogue types: a) Open-Domain Dialogues (ODDs), b) Knowledge-Grounded Dialogues (KGDs), c) Task-Oriented Dialogues (TODs), d) Question Answering (QA). Besides, we assess the impact of incorporating external knowledge by considering retrieved knowledge

[†] Equal contribution.

Proceedings of the 17th International Natural Language Generation Conference, pages 180–197
September 23–27, 2024.

and gold knowledge. In the retrieved knowledge scenario, we use RAG to add the knowledge to the model's input. We assess the performance of each technique using the same automatic metrics and comparable human evaluation. We further compute the contribution of each segment of the input vector by using integrated gradients as an explainability attribution method. We evaluate the models using an open human evaluation protocol (Mousavi et al., 2022) designed for dialogue contextualization, appropriateness, correctness, and validity. In summary, the main contributions of this paper are:

- Adaptation of Llama2$_C$ and Mistral$_I$ using fine-tuning and in-context learning[1] in four different dialogue types and corresponding corpora;
- Assessment of the impact of grounding the response generation on external knowledge, both in cases of retrieved knowledge and gold knowledge;
- Extensive study on the efficacy of each technique using automatic evaluations and human evaluation, including explainability and categorization analysis of natural language generation errors.

2 Literature Review

Open-Domain Dialogue (ODD) In earlier studies, sequence-to-sequence models have been trained for response generation in open-domain dialogues (Li et al., 2017). However, such models suffered from generating generic or inappropriate responses (Zhang et al., 2020). To improve the generation quality, studies grounded the generation on external knowledge, such as persona statements (Wolf et al., 2019; Kasahara et al., 2022; Xu et al., 2022b), the personal graph of user interactions (Mousavi et al., 2023), and retrieved documents (Huang et al., 2023). While the previous works developed data-driven models using training/fine-tuning, recent studies have explored the potential of in-context learning with LLMs (Qian et al., 2023).

Knowledge-Grounded Dialogue (KGD) Sources such as Wikipedia have been used as unstructured knowledge to ground the generated responses (Dinan et al., 2019; Gopalakrishnan et al., 2019; Komeili et al., 2022) to generate consistent and factual answers. To improve the generation quality, previous works have studied the impact of knowledge selection (Qin et al., 2023; Sun et al., 2023), different knowledge representations (Mousavi et al., 2023; Yang et al., 2023), additional knowledge elements (e.g. dialogue acts, topics) (Hedayatnia et al., 2020), training without knowledge supervision (Han et al., 2023), and in-context learning (Chen et al., 2023).

Task-Oriented Dialogue (TOD) LLMs have been fine-tuned for TOD modeling for joint dialogue state tracking and response generation (Hosseini-Asl et al., 2020b; Kulhánek et al., 2021; Wang et al., 2022; Ding et al., 2024), and robustness to spoken interactions (Thulke et al., 2024; Mousavi et al., 2024). Recent studies focus on augmenting the TOD modeling with unstructured knowledge access (Feng et al., 2020; Kim et al., 2020, 2021). In this regard, He et al. (2024) have proposed a pipeline for retrieval and grounded response generation. Raposo et al. (2023) compared in-context-learning and fine-tuning, but considered retrieved replies from previous dialogues as knowledge.

Question Answering (QA). In the most general setting, relevant documents need to be retrieved to provide an answer (Lee et al., 2019; Qu et al., 2020). Some studies have proposed to select the documents with the highest similarity with the question computed between their BERT encodings (Lee et al., 2019; Karpukhin et al., 2020). With this retrieval strategy, some studies have fine-tuned LLMs to condition the generation on the retrieved documents through grounding (Lewis et al., 2020; Izacard and Grave, 2021) or cross-attention (Borgeaud et al., 2022). Other works generated the answers using in-context learning with zero-shot (Levine et al., 2022; Cho et al., 2023). A survey compared existing generation-only, retrieval-only, and RAG models (Zhang et al., 2023) but with different base models, hindering the comparison of the techniques.

3 Experiments

We study and compare in-context learning and fine-tuning as techniques to adapt LLMs for human-machine dialogues. We select Llama-2 Chat (Llama2$_C$) (Touvron et al., 2023) and Mistral Instruct (Mistral$_I$) (Jiang et al., 2023) as base LLMs, and experiment in the context of four dialogue types: Open-Domain Dialogue

[1]The code is available at `https://github.com/sislab-unitn/Fine-Tune-or-Rag`

(ODD), Knowledge-Grounded Dialogue (KGD), Task-Oriented Dialogue (TOD), and Question Answering (QA). For each technique and dialogue type, we assess the impact of grounding the generation on documents in the scenarios of retrieved knowledge (RAG) and gold knowledge.

3.1 Datasets

In our experiment, we have selected a dataset for each of the four dialogue types (see §A.1 for selection). The statistics of these datasets are summarized in Table 1.

Open-Domain Dialogue (ODD) We select DailyDialog (Li et al., 2017), a widely-used dataset of human-human dialogues crawled from various websites used by English learners to practice. The final dataset contains 13k written dialogues with an average of 8 turns per dialogue.

Knowledge-Grounded Dialogue (KGD) We experiment on Wizard of Wikipedia (Dinan et al., 2019), a dataset of dialogues between two participants with the roles of apprentice and wizard. At each turn, the wizard can access a set of documents (passages from Wikipedia) and use it to incorporate factual knowledge in their reply. The dataset contains 20k dialogues about one of 1359 distinct topics and provides an unseen set of documents for testing.

Task-Oriented Dialogue (TOD) We select the dataset proposed for the first track of the ninth Dialogue System Technology Challenge (Kim et al., 2020), an augmented version of MultiWOZ 2.1 (Eric et al., 2020). The dataset spans over 7 domains and contains 9k multi-domain dialogues. The dialogues include turns where the system needs to access an unstructured knowledge base of 2900 documents (FAQs) to provide a correct response.

Question Answering (QA) We select NarrativeQA (Kočiský et al., 2018), a dataset of 47k questions with free-form answers based on 1.5k books and movie scripts. The question-answer pairs are formulated based on summaries of the books and movies.

3.2 Techniques

We evaluate in-context learning and fine-tuning as techniques to adapt LLMs for response generation in the selected dialogue types. In-context learning is a technique that uses instructions and examples to condition the generation. Instead, fine-tuning further trains the model (completely or partially) on the task of interest using a smaller-scale dataset than the pre-training phase. In a dialogue setting, fine-tuning should *teach* LLMs to behave as dialogue models and account for each state of the conversation between speakers.

Type	**Dataset**	**#Dials**	**Avg. #Turns**	**#Ext. Know.**
ODD	DailyDialog	13k	8	—
KGD	WoW	20k	9	†61
TOD	DSTC9 Track 1	9k	19	2900
QA	NarrativeQA	*47k	2	1572

Table 1: Selected datasets for each dialogue type: Open-Domain Dialogue (ODD), Knowledge-Grounded Dialogue (KGD), Task-Oriented Dialogue (TOD), and Question Answering (QA). #Ext. know. indicates the number of documents in the unstructured knowledge base. † In KGD the content of the knowledge base differs at each turn with an average of 61 ± 22 documents. * Question-answer exchanges.

As a baseline, for both techniques, we consider the context (i.e. the question for QA, the history for ODD, KGD, and TOD) as the input and use the default prompt structure of the models to separate user and system turns. Additionally, for TOD we append the dialogue state (a summary of user requirements), following previous work on this dialogue type (Wang et al., 2022; Ding et al., 2024). For KGD, we prepend the topic to the start of the dialogue.

3.3 Knowledge

Incorporating external knowledge for the task of response generation has been shown to improve the factual accuracy (He et al., 2024) and contextualization (Mousavi et al., 2023) of responses.

For each of the selected types but for ODD, we consider their corresponding unstructured knowledge base. Regarding KGD, we consider passages from Wikipedia, while for TOD we consider FAQs related to services and places (e.g. restaurants, hotels, taxi booking). For QA we consider all the summaries of the books and movies.

For both in-context learning and fine-tuning, we study the impact of knowledge on the generated responses, in two scenarios:

- **Retrieved knowledge**: we retrieve k documents from the unstructured knowledge base;
- **Gold knowledge**: we use the ground truth document.

For the retrieved knowledge scenario, we use the Retrieval Augmented Generation (RAG) strategy.

Model	Technique	External Knowledge	Perplexity			
			ODD	KGD	TOD	QA
Llama2$_C$	*In-Context Learning*	No Know.	64.13	35.17	25.15	1442.26
		Retrieved Know.		33.10	24.72	625.08
		Gold Know.		24.40	23.81	298.16
	Fine-Tuning	No Know.	**5.67** $\pm$ **0.01**	7.63 $\pm$ 0.01	**3.06** $\pm$ **0.01**	12.03 $\pm$ 0.06
		Retrieved Know.		6.95 $\pm$ 0.01	3.97 $\pm$ 0.01	5.47 $\pm$ 0.02
		Gold Know.		**4.38** $\pm$ **0.01**	3.12 $\pm$ 0.01	**4.98** $\pm$ **0.01**
Mistral$_I$	*In-Context Learning*	No Know.	14.19	15.31	9.82	91.42
		Retrieved Know.		14.75	9.76	42.58
		Gold Know.		9.81	9.37	16.74
	Fine-Tuning	No Know.	**6.41** $\pm$ **0.01**	8.67 $\pm$ 0.01	**3.56** $\pm$ **0.01**	14.11 $\pm$ 0.01
		Retrieved Know.		7.78 $\pm$ 0.01	3.61 $\pm$ 0.01	5.97 $\pm$ 0.01
		Gold Know.		**5.17** $\pm$ **0.01**	3.58 $\pm$ 0.01	**4.88** $\pm$ **0.01**

Table 2: **Automatic Evaluation** Perplexity of Fine-Tuning and In-Context Learning with `Retrieved` (top-3) and `Gold` (ground-truth) knowledge, on Llama2$_C$ and Mistral$_I$, in different dialogue types: Open-Domain Dialogues (ODDs), Knowledge Grounded Dialogues (KGDs), Task-Oriented Dialogues (TODs), and Question Answering (QA). Results for fine-tuned models report mean and standard deviation over three runs.

We use an off-the-shelf retriever[2] (model details in §A.2) to retrieve documents from the unstructured knowledge base. First, we encode all the documents considering their content together with their topic (KGD), place or service name (TOD), or title (QA) (Karpukhin et al., 2020). Then, at each turn, we retrieve the k most similar documents based on L2 distance with the encoded context. Finally, we feed the retrieved documents to the base models together with the context to generate a response.

In the gold knowledge scenario, we directly feed the model with the ground truth documents. This serves as an upper bound for RAG. Additionally, this strategy allows us to study the ability of the techniques to incorporate knowledge in the responses.

3.4 Models

We select the widely-used 7B version of Llama2$_C$ and Mistral$_I$ as base models. For in-context learning, we experiment with three instructions for each dialogue type and select the best based on the development set performance. For fine-tuning, we use LoRA, a parameter-efficient technique that has shown comparable performance to fine-tuning all parameters (Hu et al., 2021). Further details about the parameters are reported in §A.2.

4 Evaluation

We conduct a comparative study on the impact of in-context learning and fine-tuning to adapt LLMs for dialogues. We select Llama2$_C$ and Mistral$_I$ as base LLMs and experiment in four dialogue types: ODDs, KGDs, TODs, and QA. For each dialogue type, we study the impact of external knowledge, both retrieved and gold. Further details about the implementation and the resources used are available in the appendix (§A.2).

4.1 Automatic Evaluation

Currently available automatic metrics used for the task of response generation are not interpretable and correlate poorly with human judgments (Liu et al., 2016; Sai et al., 2022; Mousavi et al., 2022). Therefore, we focus on perplexity as it is derived from the objective function used to fine-tune the models, and present other metrics in §A.3.

Table 2 reports the perplexity of Llama2$_C$ and Mistral$_I$ on the test set of each dialogue type. In all dialogue types, fine-tuned models have obtained better performance compared to in-context learning. When considering the impact of external knowledge, models fine-tuned on TODs show that knowledge slightly increases perplexity. The high perplexity obtained by in-context learning models on QA can be explained by two reasons: first, besides the knowledge, only the question is used as context; second, while the ground truths are particularly short (4.26 tokens on average), these models generate long responses, making them unlikely to include the correct answer in the first few tokens. This does not happen for fine-tuned models since they are trained to generate shorter responses. Nevertheless, the best results have been obtained with

[2] `https://github.com/langchain-ai/langchain`

Model	Dialogue Type	Technique	% of Tokens w. Significant Contribution in Each Segment			
			Instruction	Topic/Dialogue State	Dialogue History	Knowledge
Llama2$_C$	*KGD*	In-Context Learning	21.85	28.60	15.97	**33.58**
		Fine-Tuning		39.43	13.80	**46.77**
	TOD	In-Context Learning	25.98	19.54	16.46	**38.02**
		Fine-Tuning		27.19	8.04	**64.77**
Mistral$_I$	*KGD*	In-Context Learning		**69.01**	14.89	16.10
		Fine-Tuning		**65.55**	11.00	23.45
	TOD	In-Context Learning	**69.05**	10.19	11.24	9.52
		Fine-Tuning		14.55	29.06	**56.39**

Table 3: **Explanability Study** Percentage of tokens with significant contribution to the generation in different segments of the input vector for each model in Knowledge-Grounded Dialogues (KGDs), and Task-Oriented Dialogues (TODs). All rows sum to 100. For KGD, the second column reports the contribution of the `Topic`, while for TOD it reports the contribution of the `Dialogue State`. The `Instruction` segment is only present for In-Context Learning.

gold knowledge. We report automatic evaluation results including retriever accuracy, overlap between knowledge and response tokens, and other automatic metrics in §A.3.

4.1.1 Explainability Study

To understand the contribution of each segment of the input vector (i.e. instruction, context, knowledge, topic, and dialogue state), we compute integrated gradients (Sarti et al., 2023)[3] of input elements and select the most contributing input tokens (top-25%). Table 3 reports the percentage of most contributing tokens that fall in each segment (normalized by the length of the segment). In general, in both KGD and TOD, the dialogue history is the least contributing segment, which might indicate that only a part of the history is significant for response generation. On the other hand, in KGD the topic has a higher score than the dialogue history, suggesting its importance for response generation for this dialogue type. Interestingly, Mistral$_I$ gives considerably more importance to the topic than Llama2$_C$, decreasing the importance of the knowledge segment. For the TOD type, the most contributing segment is often the knowledge, reaching over 50% with fine-tuning. This suggests that knowledge is more relevant for TOD and that relevance changes with respect to the dialogue type.

4.2 Human Evaluation

Considering the uninterpretability of automatic evaluations, we conducted a human evaluation of the generated responses to gain more insight into the models' performance. Mousavi et al. (2022) proposed four dimensions to evaluate response generation based on the most common errors and qualities. We evaluate the responses using their protocol and three of their dimensions:

- **Contextualization**: the response includes explicit or implicit references to the dialogue history (ODD, KGD, TOD) or the gold knowledge (QA);
- **Appropriateness**: the response is coherent and makes sense as a continuation of the dialogue;
- **Correctness**: the response is grammatically and syntactically correct.

According to these dimensions, we evaluate the responses for all techniques, models, and knowledge scenarios, in all dialogue types. The only exception is QA, where we do not evaluate "Appropriateness" since the dimension considers coherence with respect to a dialogue history but QA only has question-answer exchanges. Instead, we extend the protocol[4] by proposing a new dimension for QA:

- **Validity**: the response includes adequate information to answer the question.

For TOD we do not include a dimension to evaluate whether the response is in line with user requirements, as this can be measured automatically (via

[3]We use Inseq to compute integrated gradients.

[4]The extended protocol is available at `https://github.com/sislab-unitn/Human-Evaluation-Protocol/tree/v1.1`

Model	Technique	External Knowledge	Contextualization				Appropriateness			Validity
			ODD	KGD	TOD	QA	ODD	KGD	TOD	QA
Llama2$_C$	*In-Context Learning*	No Know.	**85**	70	70	50	**80**	70	60	10
		Retrieved Know.		75	65	70		75	45	35
		Gold Know.		**90**	40	**90**		**85**	45	**80**
	Fine-Tuning	No Know.	45	60	70	15	50	65	60	15
		Retrieved Know.		65	**90**	45		80	80	45
		Gold Know.		80	85	85		65	**85**	75
Mistral$_I$	*In-Context Learning*	No Know.	**90**	80	70	20	**85**	**85**	65	20
		Retrieved Know.		75	65	40		65	60	25
		Gold Know.		90	55	**75**		70	55	**80**
	Fine-Tuning	No Know.	55	90	**85**	25	55	80	80	20
		Retrieved Know.		**95**	**85**	30		**85**	**90**	40
		Gold Know.		80	75	70		65	70	70
Ground-Truth			95	80	95	90	100	85	95	90

Table 4: **Human Evaluation** Percentage of Contextualized, Appropriate (ODD, KGD, TOD), and Valid (QA) responses for In-Context Learning and Fine-Tuning with `Retrieved` (top-3) and `Gold` (ground-truth) knowledge, on Llama2$_C$ and Mistral$_I$, in different dialogue types: Open-Domain Dialogues (ODDs), Knowledge Grounded Dialogues (KGDs), Task-Oriented Dialogues (TODs), and Question Answering (QA).

dialogue state tracking metrics e.g., Joint Goal Accuracy). The dimensions can either have a positive or negative answer value, as well as "I don't know" to avoid forcing erroneous judgments on any of the two sides. For "Contextualization" and "Appropriateness", we also ask the annotators to motivate the negative judgments with the explanations proposed in the original protocol. We present the explanations and related results in §4.3.

We recruited 75 annotators on the Prolific platform[5], and we assigned 5 dialogues to each annotator. After performing quality control, we approved 65 annotators with a compensation of 9.00£/hour (marked as good on the Prolific platform). Due to the large number of responses, each annotator evaluated a different set of model responses for a given dialogue. For the purpose of quality control, for each dialogue type, two dialogues were overlapping among five annotators, while the remaining dialogues were annotated by one crowd-worker with an overlap only on the ground truth. The inter-annotator agreement measured with Fleiss' κ (Fleiss, 1971) was 0.65 (substantial agreement).

As results of the human evaluation (Table 4), we report the percentage of positively judged responses (Contextualized, Appropriate, Valid) for Llama2$_C$ and Mistral$_I$ when considering different adaptation techniques (Fine-Tuning and In-Context Learning) and knowledge (No Knowledge, Retrieved Knowledge, and Gold Knowledge) across different dialogue types. As for ODDs, we report no results for the Retrieved and Gold Knowledge scenarios since no knowledge was used for this dialogue type. Additional results on "Correctness" are reported in §A.4.

Open-Domain Dialogue (ODD) Models fine-tuned for ODD tend to generate considerably less contextualized responses than models adapted using in-context learning. In particular, fine-tuning Llama2$_C$ reduces contextualization by 40%, while for Mistral$_I$ by 35%. Similarly, fine-tuning reduces their appropriateness by 30% compared to their in-context learning version. This contrasts with automatic evaluation (Table 2), where in-context learning obtained a higher perplexity (i.e. worse results) compared to fine-tuning.

Knowledge-Grounded Dialogue (KGD) Concerning KGD, the results are model-dependent. When considering Llama2$_C$, in-context learning provides, regardless of the knowledge, 10% more contextualized responses compared to fine-tuning. On the other hand, fine-tuning Mistral$_I$ on Retrieved Knowledge leads to the highest contextualization (95%). However, using Gold instead of Retrieved Knowledge reduces the contextualization of the fine-tuned model by 15%. Furthermore, when considering the best models, Llama2$_C$ and Mistral$_I$ have a higher contextualization than the ground truth (10 to 15%), suggesting that models copy more from the dialogue history. Similarly to contextualization, adapting Llama2$_C$ with in-context learning and Gold Knowledge provides

[5]https://www.prolific.com/

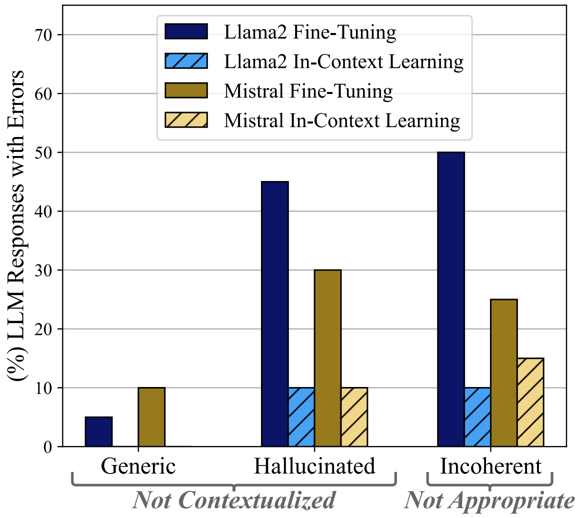

Figure 1: Percentage of LLM responses (y-axis) for each error type (*Not Contextualized* and *Not Appropriate*) and their explanation (Generic, Hallucinated, and Incoherent) (x-axis), for Llama2$_C$ and Mistral$_I$, adapted with In-Context Learning and Fine-Tuning in Open-Domain Dialogues (ODDs).

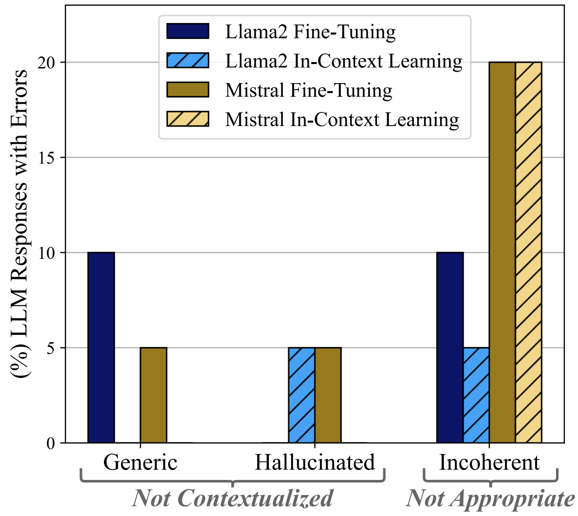

Figure 2: Percentage of LLM responses (y-axis) for each error type (*Not Contextualized* and *Not Appropriate*) and their explanation (Generic, Hallucinated, and Incoherent) (x-axis), for Llama2$_C$ and Mistral$_I$, adapted with In-Context Learning and Fine-Tuning in Knowledge-Grounded Dialogues (KGDs).

the highest percentage of appropriate responses (85%). Instead, fine-tuning (on Retrieved Knowledge) or adapting Mistral$_I$ with in-context learning (using No Knowledge) provides comparable appropriateness (85%). While according to automatic evaluation (Table 2) fine-tuning is always the best technique, human evaluation results show comparable appropriateness and contextualization for in-context learning and fine-tuning.

Task-Oriented Dialogue (TOD) When adapting Llama2$_C$ and Mistral$_I$ to TOD, the results clearly show that fine-tuning is preferable over in-context learning. In particular, if we consider the best model for each technique, when fine-tuned Llama2$_C$ generates 20% more contextualized responses, while Mistral$_I$ generates 15% more. Although fine-tuned models benefit from external knowledge, Retrieved and Gold Knowledge visibly reduce contextualization of in-context learning models (at most by 30% for Llama2$_C$ and 15% for Mistral$_I$). Similar behavior can be observed for in-context learning in terms of appropriateness, where Gold Knowledge reduces Llama2$_C$ results by 15% and Mistral$_I$ by 10%. This is in line with the explainability study (Table 3), where models adapted with in-context learning have a lower contribution from the knowledge segment than their fine-tuned version. In general, if we consider the best models for each technique, fine-tuned models generate 25% more appropriate responses.

Question Answering (QA) In QA, results show improved contextualization and validity when including knowledge, with the best results obtained with gold knowledge. When considering the best model for each technique, in-context learning increases the percentage of contextualized responses by 5%. These results greatly differ from Table 2 and show how unreliable automatic evaluation can be. Although models fine-tuned on No or Retrieved Knowledge obtain comparable or higher validity than in-context learning, adding Gold Knowledge to adapt Llama2$_C$ and Mistral$_I$ with in-context learning increases their validity respectively by 5% and 10%. Finally, even with Gold Knowledge, no model reaches the validity of the ground truth (90%).

These findings indicate that the best technique depends on the dialogue type and the base LLM. Regarding the techniques, in-context learning leads to more contextualized and appropriate responses in ODDs, while fine-tuning improves contextualization and appropriateness in TODs. Regarding the base LLMs, in KGDs adapting Llama2$_C$ with in-context learning leads to the best results, while Mistral$_I$ benefits the most from fine-tuning. Furthermore, in QA the quality of knowledge impacts contextualization and validity the most, while adaptation techniques have a minor effect.

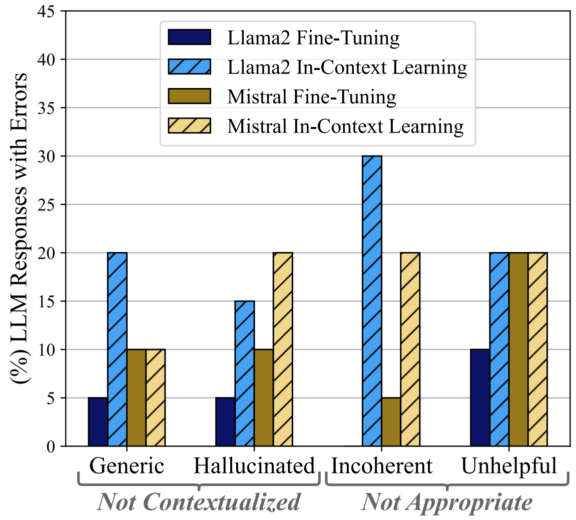

Figure 3: Percentage of LLM responses (y-axis) for each error type (*Not Contextualized* and *Not Appropriate*) and their explanation (Generic, Hallucinated, Incoherent, and Unhelpful) (x-axis), for Llama2$_C$ and Mistral$_I$, adapted with In-Context Learning and Fine-Tuning in Task-Oriented Dialogues (TODs).

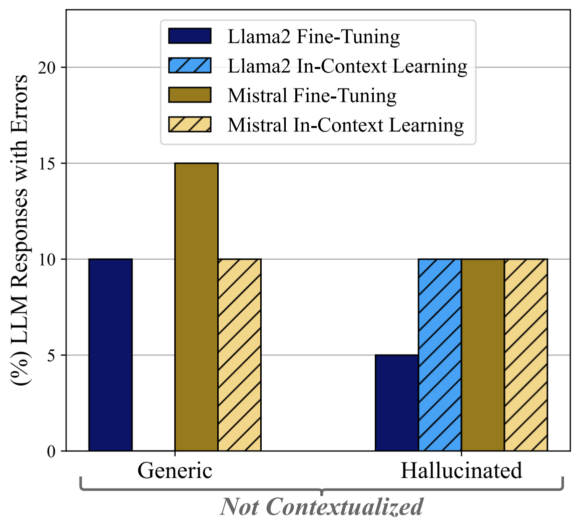

Figure 4: Percentage of LLM responses (y-axis) for each error type (*Not Contextualized*) and their explanation (Generic, and Hallucinated) (x-axis), for Llama2$_C$ and Mistral$_I$, adapted with In-Context Learning and Fine-Tuning in Question Answering (QA).

4.3 Explaining Negative Human Judgments

To better understand the shortcomings of the techniques, we investigate the motivations provided by the annotators to support their negative judgments. For each technique, we considered the scenario with gold external knowledge as the theoretical upper bound (except for ODDs where no external knowledge is required). Following the original protocol, we consider two explanations for *Not Contextualized* responses:

- **Generic**: the response is generic or does not contain any reference (implicit or explicit) to the dialogue history (ODD, KGD, TOD) or the gold knowledge (QA);
- **Hallucinated**: the response is inconsistent with the information contained in the dialogue history (ODD, KGD, TOD) or the gold knowledge (QA).

Regarding *Not Appropriate* responses, the protocol has proposed one explanation (as an alternative to a free-form explanation):

- **Incoherent**: the response is not coherent with the context.

To better characterize errors in TODs, we propose an additional explanation:

- **Unhelpful**: the response candidate is not helpful in fulfilling the user's request.

In this section, we report the percentage of negatively judged responses with a certain explanation out of all the responses.

Open Domain Dialogue (ODD) In ODDs (Figure 1), fine-tuning causes the generation of few generic responses, while for in-context learning none are present. Moreover, fine-tuned models generate around 30% more hallucinated responses, and around 25% more incoherent responses.

Knowledge-Grounded Dialogue (KGD) In KGDs (Figure 2), fine-tuning causes the generation of a few generic responses. Regarding hallucinated responses, fine-tuning slightly reduces them for Llama2$_C$ but increases them for Mistral$_I$. Differently, fine-tuning slightly increases the incoherent responses for Llama2$_C$, but has no impact for Mistral$_I$.

Task-Oriented Dialogue (TOD) For the TOD type (Figure 3), while for Mistral$_I$ fine-tuning has no impact on generic responses, it reduces generic responses by 15% for Llama2$_C$. For both models, fine-tuning reduces the number of hallucinated responses by 10%, and improves coherence by around 20% both models. It further reduces unhelpful responses by 10% for Llama2$_C$.

Question Answering (QA) For the QA type (Figure 4), fine-tuned models generate more generic responses than models adapted with in-context learning. Instead, fine-tuning results in fewer hallucinated responses for Llama2$_C$, although it has no effect for Mistral$_I$.

5 Conclusion

We have conducted an extensive analysis on the efficacy of fine-tuning and in-context learning to adapt LLMs for different dialogue types. We have experimented with Retrieval-Augmented Generation (RAG) and gold knowledge to assess the impact of grounding the response generation on external knowledge. We have studied the models' performance using consistent criteria in both automatic (perplexity, explainability studies) and human evaluations.

Our study highlights the limitation of currently available automatic metrics and the necessity of conducting human evaluations to advance human-machine dialogue research, as the evaluations by human judges correlate poorly with automatic metrics. Furthermore, conducted human evaluations indicate that there is no universal best-technique for adapting LLMs to a dialogue type and the performance of each technique depends on the base LLM as well as the dialogue type. In addition, the correct incorporation of external knowledge depends on various factors such as the retriever accuracy, the representation of the knowledge, and the presence of noise (non-gold) documents, as it can be the least contributing element in the input vector according to explainability studies.

Limitations

Due to the limited computational resources, we could experiment with 7B models, hampering us in validating our findings on larger models. Furthermore, the reproducibility of human evaluation results may be subject to variability, due to possible differences in the set of crowd workers.

Acknowledgments

We acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Qinyu Chen, Wenhao Wu, and Sujian Li. 2023. Exploring in-context learning for knowledge grounded dialog generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10071–10081, Singapore. Association for Computational Linguistics.

Sukmin Cho, Jeongyeon Seo, Soyeong Jeong, and Jong Park. 2023. Improving zero-shot reader by reducing distractions from irrelevant documents in open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3145–3157, Singapore. Association for Computational Linguistics.

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

Zeyuan Ding, Zhihao Yang, Yinbo Qiao, and Hongfei Lin. 2024. Kmc-tod: Structure knowledge enhanced multi-copy network for task-oriented dialogue system. *Knowledge-Based Systems*, 293:111662.

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, Marseille, France. European Language Resources Association.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Gunsoo Han, Daejin Jo, Daniel Nam, Eunseop Yoon, Taehwan Kwon, Seungeun Rho, Kyoung-Woon On, Chang Yoo, and Sungwoong Kim. 2023. Efficient latent variable modeling for knowledge-grounded dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2683–2702, Singapore. Association for Computational Linguistics.

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.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020a. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020b. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2023. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2523–2540, Singapore. Association for Computational Linguistics.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

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. 2023. Mistral 7b.

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, Online. Association for Computational Linguistics.

Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2022. Building a personalized dialogue system with prompt-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 96–105, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. 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*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. "how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online. Association for Computational Linguistics.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yoav Levine, Ori Ram, Daniel Jannai, Barak Lenz, Shai Shalev-Shwartz, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2022. Huge frozen language models as readers for open-domain question answering. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tur. 2023. Using in-context learning to improve dialogue safety. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11882–11910, Singapore. Association for Computational Linguistics.

Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. 2023. Response generation in longitudinal dialogues: Which knowledge representation helps? In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 1–11, Toronto, Canada. Association for Computational Linguistics.

Seyed Mahed Mousavi, Gabriel Roccabruna, Simone Alghisi, Massimo Rizzoli, Mirco Ravanelli, and Giuseppe Riccardi. 2024. Are llms robust for spoken dialogues?

Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, and Giuseppe Riccardi. 2022. Evaluation of response generation models: Shouldn't it be shareable and replicable? In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 136–147, Abu Dhabi, United Arab Emirates (Hybrid). 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

Lang Qin, Yao Zhang, Hongru Liang, Jun Wang, and Zhenglu Yang. 2023. Well begun is half done: Generator-agnostic knowledge pre-selection for knowledge-grounded dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4696–4709, Singapore. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 539–548, New York, NY, USA. Association for Computing Machinery.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Gonçalo Raposo, Luisa Coheur, and Bruno Martins. 2023. Prompting, retrieval, training: An exploration of different approaches for task-oriented dialogue generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 400–412, Prague, Czechia. Association for Computational Linguistics.

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

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weiwei Sun, Pengjie Ren, and Zhaochun Ren. 2023. Generative knowledge selection for knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2077–2088, Dubrovnik, Croatia. Association for Computational Linguistics.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2024. Task-oriented document-grounded dialog systems by hltpr@rwth for dstc9 and dstc10. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:733–741.

Jörg Tiedemann. 2009. *News from OPUS—A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume 5, pages 237–248.

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.

Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2022. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2698–2703, New York, NY, USA. Association for Computing Machinery.

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*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Yizhe Yang, Heyan Huang, Yuhang Liu, and Yang Gao. 2023. Graph vs. sequence: An empirical study on knowledge forms for knowledge-grounded dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15846–15858, Singapore. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

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*, pages 309–323, Prague, Czechia. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Datasets

We briefly present the reasons for selecting the datasets.

Open-Domain Dialogue (ODD) Differently from other datasets, DailyDialog dialogues only involve two participants (Tiedemann, 2009; Baumgartner et al., 2020), are not audio transcriptions (Godfrey et al., 1992), have more than two exchanges between the participants (Rashkin et al., 2019), and are not restricted by a persona (i.e. few sentences describing the user's interests) (Zhang et al., 2018; Xu et al., 2022a).

Knowledge-Grounded Dialogue (KGD) Wizard of Wikipedia provides a test set with an unseen set of documents (Zhou et al., 2018; Komeili et al., 2022) and its knowledge has not changed over time (i.e. comparable with previous/future studies) (Gopalakrishnan et al., 2019; Hedayatnia et al., 2020).

Task-Oriented Dialogue (TOD) A few other TOD datasets include unstructured knowledge access but consist only of a spoken test set (Kim et al., 2021), or provide no dialogue state annotation (Feng et al., 2020). The dataset proposed in the ninth Dialogue System Technology Challenge augmented MultiWOZ 2.1 (Eric et al., 2020) with knowledge access turns but removed the dialogue state annotation. To always include the dialogue state in our analysis, we recovered the dialogue state annotation from the original MultiWOZ 2.1 dialogues, and we only considered the dialogues from this dataset.

Question Answering (QA) We choose NarrativeQA because it has a publicly available test set (to evaluate the retriever) and answers are expressed as free-form text (to evaluate response generation) (Rajpurkar et al., 2016, 2018; Yang et al., 2018; Kwiatkowski et al., 2019). Although the original task always provides the correct document, we also wanted to investigate the performance of the retriever when considering documents with an average length of 600 tokens. Additionally, we avoided splitting documents into smaller chunks (e.g. passages or sentences) because this would have made the computation of the retriever performance more challenging.

A.2 Implementation and resources

Models and parameters We fine-tuned the models using LoRA (rank 32 and alpha 64) for a maximum of 10 epochs with an early stopping patience of 2. We chose AdamW (Loshchilov and Hutter, 2017) as the optimizer and used a learning rate of 10^{-4} for Llama2$_C$ and 10^{-5} for Mistral$_I$ (selected based on the performance on the development sets). To obtain an encoding for both documents and queries, we used all-mpnet-base-v2[6]. We have then stored the encoded documents in a FAISS vector store (used for retrieval).

Input structure We separated the segments of the input vector with their name followed by a colon (i.e. "Dialogue state:", "Topic:", "Knowledge:", "Question:", "Answer:") similarly to previous work (Izacard and Grave, 2021; Wang et al., 2022; Chen et al., 2023; Sun et al., 2023). For TOD, we represented the dialogue state as a comma-separated list of domain slot value triplets (Hosseini-Asl et al., 2020b; Wang et al., 2022).

Instructions Table 5 reports the instructions used for in-context learning experiments. For each dialogue type, we have experimented with three different instructions describing the task and the various input segments (e.g. dialogue history, topic, and knowledge). We have selected the best instruction based on the development set performance.

Generation We sampled 10% of the data (in a stratified fashion, based on the length of the responses) from the development set of each dialogue type. For each model, we used grid search to find, for the sampled data, the combination of parameters (top-p, top-k, and temperature) leading to the highest BLEU-4. The best combination of parameters was used to generate the responses for the test set.

GPU Requirements Most computations were performed on a single NVIDIA A100 GPU with 80GB, requiring less than 50 hours to execute. In a few cases, we had to use two (i.e. fine-tuning the models for QA using more than one document) or three (i.e. integrated gradients) A100 with 80GB each.

A.3 Additional Automatic Evaluation

To automatically evaluate the quality of the generated text, we have considered BLEU-4 (Papineni et al., 2002), F1 (i.e. unigram overlap), and ROUGE-L (Lin, 2004). Furthermore, we have used KF1 (Shuster et al., 2021) to measure the overlap between the prediction and the knowledge selected

[6] `https://www.sbert.net/docs/pretrained_models.html`

Dialogue Type	Instruction
ODD	`""`
	`"This is a conversation between two people. Use the context to write an engaging reply for the other person."`
	`"Write a coherent continuation for the proposed conversation."`
KGD	`""`
	`"This is a conversation between two people about a Topic. Use the Dialogue and the additional Knowledge as context to write an engaging reply for the other person.",`
	`"Write a coherent continuation for the proposed conversation based on the additional Knowledge."`
TOD	`""`
	`"In the following conversation a user wants to achieve some goal and needs help from an assistant. Continue the conversation with the response of the assistant."`
	`"Write a coherent continuation for the proposed conversation."`
QA	`""`
	`"You are presented with a user's Question about a movie or book. Answer to the user's Question using the information provided in the Context."`
	`"Answer to the user's question using the provided information (if available)."`

Table 5: Instructions used to adapt the model to a specific dialogue type with in-context learning. We defined three instructions for each dialogue type, describing the task and the various input segments (e.g. dialogue history, topic, dialogue state, and knowledge). We selected the best instruction based on the development set performance.

by the annotators. For reproducibility purposes, we have computed ROUGE-L using the official implementation[7] and all the remaining metrics using ParlAI[8]. No pre-processing was performed on the model-generated answers.

Table 6 reports the performance for each dialogue type. As mentioned in Section 4.1, the best performance is obtained by fine-tuned models. Following, we analyze the results for each dialogue type.

Open-Domain Dialogue (ODD) Although fine-tuning achieves a higher BLEU-4, the results show that both techniques produce very different responses with respect to the ground truth.

Knowledge-Grounded Dialogue (KGD) We report the performance of the models on the unseen test set (i.e. the knowledge base contains documents that are only present in the test set). The results show that models adapted using fine-tuning obtain a higher F1 than in-context learning. Furthermore, the best models tend to copy more from the gold knowledge compared to the annotators (as shown in the ground truth).

Task-Oriented Dialogue (TOD) Differently from the other types, Llama2$_C$ and Mistral$_I$ have obtained the best performance in terms of BLEU-4 when fine-tuned with no additional knowledge. Further investigation suggests this happens because of the high overlap between the knowledge used for training and testing (82%). We report the performance on the documents only available in the test phase in Table 7 (TOD†). In this scenario, gold knowledge does indeed increase the performance of the models.

Question Answering (QA) Although fine-tuned models achieve the highest ROUGE-L, in-context learning models tend to provide longer and possibly more detailed responses, as reported in terms of KF1. Because ground truths are particularly short (4.26 tokens on average), models that generated longer responses (especially models adapted with in-context learning) were awarded a lower ROUGE-L.

A.3.1 Retriever Accuracy

We study the performance of the retriever for each dialogue type and report Recall@K in Figure 5. Because of the size of the knowledge base (Table 1), the retriever achieves the lowest performance on TOD. However, although the knowledge base for QA is bigger than for KGD, the retriever achieves a higher recall for QA. Further study suggest that, although the retriever selects the gold sentence in

[7] https://github.com/google-research/google-research/tree/master/rouge

[8] https://parl.ai

Model	Technique	External Knowledge	BLEU-4		KF1			F1	ROUGE-L
			ODD	TOD	KGD	TOD	QA	KGD	QA
Llama2$_C$	*In-Context Learning*	`No Know.`	0.2	0.85	11.61	13.66	5.26	12.68	5.59
		`Retrieved Know.`		0.83	13.51	12.10	5.65	12.91	14.86
		`Gold Know.`		1.07	25.87	21.03	**6.72**	16.59	23.22
	Fine-Tuning	`No Know.`	**0.3**	**6.72**	17.43	34.04	0.74	18.46	17.25
		`Retrieved Know.`		4.33	25.10	26.85	1.15	20.70	46.21
		`Gold Know.`		5.39	**76.23**	**42.69**	1.44	**38.41**	**73.38**
Mistral$_I$	*In-Context Learning*	`No Know.`	0.2	1.33	10.96	13.01	4.84	11.04	6.94
		`Retrieved Know.`		1.06	13.83	12.53	6.09	12.22	10.26
		`Gold Know.`		1.33	25.95	28.74	**7.07**	15.88	21.74
	Fine-Tuning	`No Know.`	**0.9**	**4.09**	15.47	29.27	0.67	18.63	12.73
		`Retrieved Know.`		3.85	21.63	30.44	1.18	20.49	45.40
		`Gold Know.`		3.94	68.36	**43.04**	1.46	**38.21**	**70.54**
Ground Truth			100	100	37.79	38.48	1.52	100	100

Table 6: **Automatic Evaluation** BLEU-4, KF1, F1 and ROUGE-L for In-Context Learning and Fine-Tuning with `Retrieved` (top-3) and `Gold` (ground-truth) knowledge, on Llama2$_C$ and Mistral$_I$, in different dialogue types: Open-Domain Dialogues (ODDs), Knowledge Grounded Dialogues (KGDs), Task-Oriented Dialogues (TODs), and Question Answering (QA).

Model	Technique	External Knowledge	BLEU-4		KF1	
			TOD	TOD†	TOD	TOD†
Llama2$_C$	*In-Context Learning*	`No Know.`	0.85	0.60	13.66	12.39
		`Retrieved Know.`	0.83	0.44	12.10	10.44
		`Gold Know.`	1.07	2.67	25.87	23.77
	Fine-Tuning	`No Know.`	**6.72**	4.33	34.04	25.73
		`Retrieved Know.`	4.33	3.15	26.85	22.92
		`Gold Know.`	5.39	**8.50**	**42.69**	**45.49**
Mistral$_I$	*In-Context Learning*	`No Know.`	1.33	1.12	13.01	11.91
		`Retrieved Know.`	1.06	1.02	12.53	10.36
		`Gold Know.`	1.33	3.70	28.74	28.79
	Fine-Tuning	`No Know.`	**4.09**	5.83	29.27	25.47
		`Retrieved Know.`	3.85	4.76	30.44	25.61
		`Gold Know.`	3.94	**10.63**	**43.04**	**49.40**
Ground Truth			100	100	38.48	39.91

Table 7: **Automatic Evaluation** BLEU-4 and KF1 for In-Context Learning and Fine-Tuning with `Retrieved` (top-3) and `Gold` (ground-truth) knowledge, on Llama2$_C$ and Mistral$_I$, in Task-Oriented Dialogues (TODs). † indicates that only test turns with unseen knowledge were included.

only a few cases, the model retrieves a sentence from the same paragraph more than 69% of the time.

A.4 Human Evaluation

Table 8 reports the results for the "Correctness" dimension of Human Evaluations. Except for ODD, fine-tuning tends to improve correctness.

Table 9 presents the question and the answer options for the proposed "Validity" dimension used in QA.

Model	Technique	External Knowledge	Correctness			
			ODD	KGD	TOD	QA
Llama2$_C$	*In-Context Learning*	No Know.	**95**	80	**95**	75
		Retrieved Know.		80	60	60
		Gold Know.		80	70	80
	Fine-Tuning	No Know.	65	**90**	70	75
		Retrieved Know.		**90**	90	55
		Gold Know.		85	85	**85**
Mistral$_I$	*In-Context Learning*	No Know.	**95**	70	75	60
		Retrieved Know.		55	70	50
		Gold Know.		**85**	60	80
	Fine-Tuning	No Know.	65	**85**	80	50
		Retrieved Know.		75	**100**	45
		Gold Know.		70	80	**85**
Ground-Truth			95	70	85	80

Table 8: **Human Evaluation** Percentage of Correct (ODD, KGD, TOD, QA) responses for In-Context Learning and Fine-Tuning with `Retrieved` (top-3) and `Gold` (ground-truth) knowledge, on Llama2$_C$ and Mistral$_I$, for different dialogue types: Open-Domain Dialogues (ODDs), Knowledge Grounded Dialogues (KGDs), Task-Oriented Dialogues (TODs), and Question Answering (QA).

Dimension	Question	Answer Option	Option Definition
Validity	*Is the response candidate valid?*	`Valid`	The response candidate includes the right information from the context to adequately answer the proposed question.
		`Not Valid`	The response candidate does not include the right information from the context to adequately answer the proposed question.
		`I don't know`	The response candidate includes some information that is adequate to answer the proposed question, but some that is not.

Table 9: Question and answer options presented to the annotators for the proposed Validity dimension.

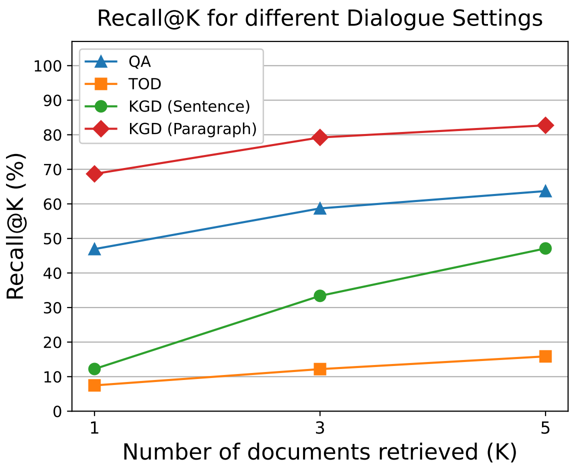

Figure 5: Performance of the off-the-shelf retriever for each dialogue type. The retriever achieves the lowest Recall@K on TOD because of the larger knowledge base size (2900 documents). However, the retriever achieves a higher Recall@K for QA, even though its knowledge base is bigger than the one for KGD (355 vs. 61 $\pm$ 21). Further studies indicate that, despite the model is not capable to retrieve the exact sentence of the annotator (KGD Sentence), the retriever selects a sentence belonging to the same paragraph more than 69% of the time (KGD Paragraph).