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.

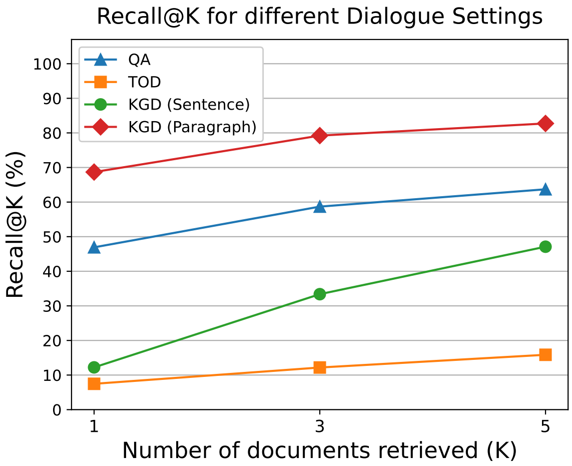

Figure 5: Performance of the off-the-shelf retriever for each dialogue type. The retriever achieves the lowest Recall@K on TOD because of the larger knowledge base size (2900 documents). However, the retriever achieves a higher Recall@K for QA, even though its knowledge base is bigger than the one for KGD (355 vs. 61 ± 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).