A RETRIEVAL CORPUS

We combine the text chunks from the Dec. 20, 2021 Wikipedia dump released by Izacard et al. (2022b) with additional ones from the 2017-2020 CommonCrawl dumps. The Wikipedia dump includes lists and infoboxes in addition to regular articles. The articles are split by section, where long sections are further split into text chunks of equal sizes and contain less than 200 words, leading to a total of 37M text chunks. We randomly sample a subset of articles from the CommonCrawl dumps, and split them into equal-sized text chunks that contain less than 100 white-space-separated words, leading to a total of 362M text chunks.

We use a GPU-based exact k-nearest-neighbor search index implementation[11] released by Izacard et al. (2022b).

B IMPLEMENTATION DETAILS

Retrieval-augmented LM Fine-tuning We use the top-3 retrieved text chunks for a given example (i.e. $\tilde{k} = 3$) to generate the fine-tuning instances. To improve fine-tuning efficiency, we pack multiple examples up to the language model context window limit (2048 tokens). Each example is demacrate by a pair of <bos> and <eos> tokens, and we adopt the document attention masking (Iyer et al., 2022) such that a token only attends to the previous tokens in the same example. We use a dataset mixture that contains 10% unsupervised text and 5% OASST-1 data. For the remaining datasets, we establish a cap on the number of examples per dataset at $\eta = 7500$. We then randomly sample batches in accordance with this adjusted mixture probability.

We fine-tune the 7B, 13B and 65B LLAMA models using 8, 16 and 64 A100 GPUs, respectively. The fine-tuning hyperparameters are detailed in Table 8. Similar to Zhou et al. (2023), we found that the best generalization performance on the dev set can be achieved using a small number of fine-tuning steps. We evaluate the models every 100 steps, and select the best checkpoint based on the average dev set performance over the 6 development KILT tasks shown in Table 11 (early stopping).

Table 8: Hyperparameters for retrieval-augmented LM fine-tuning.

Model	peak lr	end lr	lr scheduler	warm-up	# steps	early stopping	batch size	model parallel	seq len
RA-DIT 7B	1e-5	1e-7	cosine	200	500	500	64	1	2048
RA-DIT 13B	1e-5	1e-7	cosine	200	500	400	128	2	2048
RA-DIT 65B	1e-5	1e-7	cosine	200	500	300	128	8	2048

64-shot Eval Task Fine-tuning Table 9 summarizes our hyperparameters for 64-shot fine-tuning on the 9 KILT eval tasks shown in Table 12 except for MMLU. Given the small amount of examples used ($64 \times 9 = 576$), we fine-tune for a significantly less number of steps at this stage without using warm-up. We evaluate the model every 50 steps, and select the best checkpoint based on the average dev set performance over the 6 development KILT tasks shown in Table 11.

Table 9: Hyperparameters for 64-shot fine-tuning on the eval tasks.

Model	peak lr	end lr	lr scheduler	warm-up	# steps	early stopping	batch size	model parallel	seq len
LLAMA 65B	1e-5	1e-6	linear	0	100	100	8	8	2048
RA-DIT 13B	1e-5	1e-6	linear	0	100	50	32	2	2048
RA-DIT 65B	1e-5	1e-6	linear	0	100	50	32	8	2048

Retriever Fine-tuning We employ both unsupervised text and downstream tasks for retriever fine-tuning. For the *corpus data*, we randomly sample 900k text chunks from our retrieval corpus to

[11] https://github.com/facebookresearch/atlas

form a set of self-supervised data, using the first 50 tokens of each chunk as the input x and the last 50 tokens as the ground-truth output y. In addition, we leverage the multi-task instruction tuning datasets (MTI data) as shown in Table 1, including 10 open-domain question answering and dialog tasks, with a total of 286k training examples. As discussed in §5.1, we observe that, when used alone, the corpus data works slightly better than the downstream tasks. However, combining both types of fine-tuning data yields the best results and outperforms using either source alone. Therefore, we adopt a mixture of 95% corpus data and 5% downstream tasks for retriever fine-tuning in our final model.

We fine-tune the DRAGON+ retriever on 16 A100 GPUs using the dpr-scale codebase[12]. The retriever is fine-tuned using a learning rate of 1e-5 with 1237 warmup steps (DRAGON default), a per-GPU batch size of 32, and a temperature $\tau = 0.01$, for a single epoch over a combination of 5% *MTI data* and 95% *corpus data*. We adopt the KL-divergence loss as discussed in Section 2.4 using the top-10 retrieved chunks for each example. For simplicity and efficiency, we produce the top-10 retrieved chunks and their LSR scores (Eqn. 4) using LLAMA 65B and DRAGON+, and do not update them during R-ft. Furthermore, as only the query encoder is fine-tuned, there is no need to update the chunk embeddings in the retriever index. Model validation is performed once every 500 steps using the same mean reciprocal rank (MRR) metric as in the original DRAGON paper (Lin et al., 2023), on a combined validation set from the 10-task MTI data.

Inference Without further specification, we use the top-10 retrieved text chunks for a given example (i.e. k = 10) and ensemble their predictions during inference. For multi-choice tasks, we compute the weighted average probability of each choice items according to Eq. 2 and select the choice with the highest probability. For generation tasks, we perform decoding using each augmented prompt independently, compute the weighted average probability of each unique generated answer, and output the answer with the highest probability[13]. When computing probabilities of output answers, we use several scoring functions: "nll", "nll_char", "nll_token", and "nll_compl". "nll" is the sum of negative log likelihood across all tokens in the sequence. "nll_char" and "nll_token" are "nll" divided by the numbers of characters and subword units in output answers respectively. "nll_compl" selects answers based on the probability divided by the probability of the answer given "Answer:": $\frac{p(y|x)}{p(y|\text{“}Answer:\text{”})}$.

C FINE-TUNING DATASET TEMPALTES

Table 10: Instruction template used for our fine-tuning datasets. `<inst_s>`, `<inst_e>` and `<answer_s>` are special markers denoting the start and the end of a field.

Category	Instruction Tuning Template	Query Template
Dialogue	Background: {retrieved passage}\n\nQ: {turn$_1$} A: {turn$_2$} Q: {turn$_3$} A: ...	{turn$_1$} {turn$_2$} {turn$_3$} ...
Open-domain QA	Background: {retrieved passage}\n\n<inst_s> {question} <inst_e> <answer_s> {answer}	{question}
Reading Comprehension	Background: {context}\n\n<inst_s> {question} <inst_e> <answer_s> {answer}	{question}
Summarization	Background: {context}\n\nSummarize this article: <inst_e> <answer_s> {summary}	
Chain-of-thought Reasoning	Background: {retrieved passage}\n\n<inst_s> {instructions} {reasoning chain} <answer_s> {answer}	{question}

Table 10 shows the templates we used to serialize our instruction tuning datasets. Following Chung et al. (2022b) and Iyer et al. (2022), we randomize the field markers used during training to avoid overfitting. In pariticular, when serializing a task example, we randomly sample from {"Q:", "Ques-

[12] https://github.com/facebookresearch/dpr-scale

[13] A more sophisticated implementation of ensembling for generation tasks involves computing a weighted ensemble of the output distribution at every step and then sampling from this distribution. However, we opt for the simpler implementation as it performs reasonably well and allows us to execute inference with fewer GPUs.

Table 11: Our evaluation datasets. † indicates the development datasets we used to select fine-tuning hyperparameters.

Task	Dataset name	Acronym	Metric	Score
Open-domain QA	MMLU (Hendrycks et al., 2021a)	MMLU	Acc.	nll
	Natural Questions (Kwiatkowski et al., 2019)	NQ	EM	nll
	TriviaQA (Joshi et al., 2017)	TQA	EM	nll
	†HotpotQA (Yang et al., 2018)	HoPo	EM	nll
	ELI5 (Fan et al., 2019)	ELI5	Rouge-L	nll_token
Fact Checking	†FEVER (Thorne et al., 2018)	FEV	Acc.	nll
Entity Linking	†AIDA CoNLL-YAGO (Hoffart et al., 2011)	AIDA	Acc.	nll
Slot Filling	†Zero-Shot RE (Levy et al., 2017)	zsRE	Acc.	nll
	†T-REx (Elsahar et al., 2018)	T-REx	Acc.	nll
Dialogue	†Wizard of Wikipedia (Dinan et al., 2019)	WoW	F1	nll_token
Commonsense Reasoning	BoolQ (Clark et al., 2019)	BoolQ	Acc.	nll_compl
	PIQA (Bisk et al., 2020)	PIQA	Acc.	nll_char
	SIQA (Sap et al., 2019)	SIQA	Acc.	nll_char
	HellaSwag (Zellers et al., 2019)	HellaSwag	Acc.	nll_char
	WinoGrande (Sakaguchi et al., 2019)	WinoGrande	Acc.	nll_char
	ARC-Easy (Clark et al., 2018)	ARC-E	Acc.	nll_char
	ARC-Challenge (Clark et al., 2018)	ARC-C	Acc.	nll_char
	OpenBookQA (Mihaylov et al., 2018)	OBQA	Acc.	nll_compl

tion: ", and ""} for <inst_s>, set <inst_e> to "\n" and randomly sample from {"A:", "Answer:"} for <answer_s>.

D Evaluation Datasets and Templates

Table 12: Language model prompts and retriever query templates used for our evaluation datasets. We did not perform retrieval for commonsense reasoning tasks evaluation.

Task	LLM Prompt Template	Query Template
Knowledge-Intensive Tasks		
MMLU	Background: {retrieved passage}\n\nQuestion: {question}\nA. {choice}\nB. {choice}\nC. {choice}\nD. {choice}\nA: {answer}	{question}\nA. {choice}\nB. {choice}\nC. {choice}\nD. {choice}
NQ, TQA, ELI5, HoPo, zsRE	Background: {retrieved passage}\n\nQ: {question}\nA: {answer}	{question}
AIDA	Background: {retrieved passage}\n\n{context}\nOutput the Wikipedia page title of the entity mentioned between [START_ENT] and [END_ENT] in the given text\nA: {answer}	{context} tokens between [START_ENT] and [END_ENT]
FEV	Background: {retrieved passage}\n\nIs this statement true? {statement} {answer}	{statement}
T-REx	Background: {retrieved passage}\n\n{entity_1} [SEP] {relation}\nA: {answer}	{entity_1} [SEP] {relation}
WoW	Background: {retrieved passage}\n\nQ: {turn$_1$}\nA: {turn$_2$}\nQ: {turn$_3$} ...\nA: {answer}	{turn$_1$} {turn$_2$} {turn$_3$} ...
Commonsense Reasoning Tasks		
ARC-E, ARC-C	Question: {question}\nAnswer: {answer}	
BoolQ	{context}\nQuestion: {question}\nAnswer: {answer}	
HellaSwag	{context} {ending}	
OpenbookQA	{question} {answer}	
PIQA	Question: {question}\nAnswer: {answer}	
SIQA	{context} Q: {question} A: {answer}	
WinoGrande	{prefix} {answer} {suffix}	

Table 11 shows the evaluation datasets used in our experiments. For dev set evaluation, we use a maximum of 2500 randomly sampled examples from the respective official dev sets to reduce the computational cost. For test set evaluation, we use the full set to ensure fair comparison with

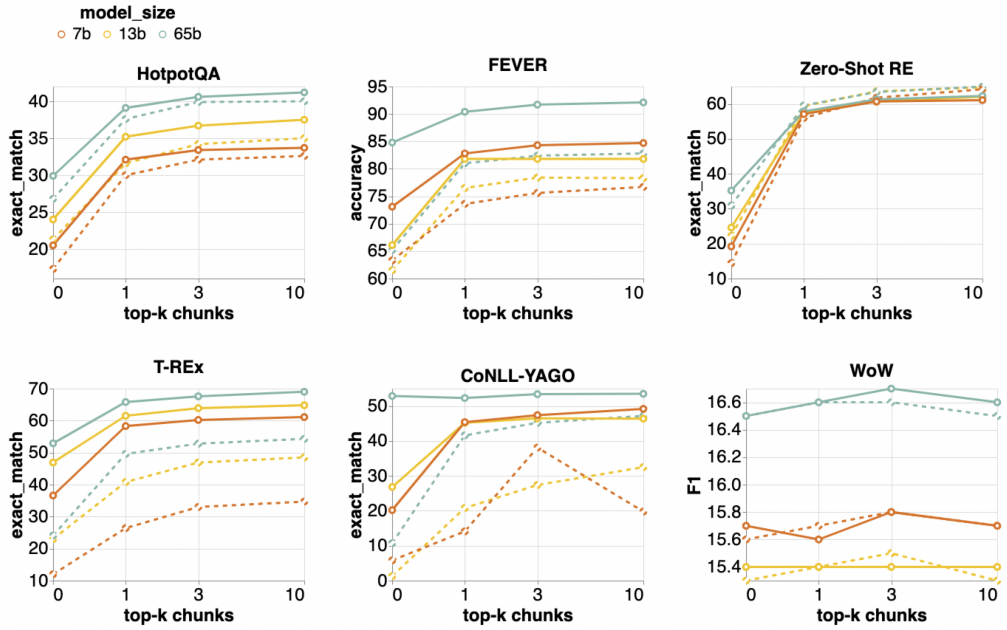

Figure 2: RA-IT model performance (combined with DRAGON+) across sizes 7B, 13B and 65B on our development tasks. 0-shot performance: dashed lines; 5-shot performance: solid lines.

previous work. We also describe the language model instruction templates and retriever queries used in our evaluation in Table 12.

E ADDITIONAL EXPERIMENTS

E.1 SCALING LAWS OF RETRIEVAL AUGMENTED LANGUAGE MODEL FINE-TUNING

We investigate the impact of the base language model size when retrieval-augmented instruction tuning is applied, and summarize the results in Figure 2. We combine the fine-tuned models with the base DRAGON+ retriever in this set of experiments.

Overall, all models substantially benefit from retrieval augmentation, with smaller models witnessing even bigger improvements. We further note that retrieval augmentation can be an effective strategy for enhancing the performance of smaller models (hence reducing pre-training and inference costs), given the 7B model leveraging > 1 retrieved chunks surpassed the performance of the vanilla 65B model on several tasks. This trend also differs across tasks. For tasks that primarily measure one-hop fact look-up abilities (such as Zero-Shot RE and T-REx), retrieval augmentation provides significant improvements across all model sizes and can bring the performance of smaller models closer to that of their larger counterparts. For more complex tasks (such as HotpotQA and WoW), the advantage of using a larger LLM remains prominent.

F EXAMPLES

In this section, we show the task prompts, the corresponding retrieved passages and model predictions generated by LLAMA 65B instruction-tuned with retrieval augmentation (RA-IT 65B) and LLAMA 65B instruction-tuned conventionally (IT 65B) on selected task examples.

F.1 HOTPOTQA

We analyze the performance of the two models on the development set of HotpotQA in the zero-shot setting since under this setting RA-IT 65B outperforms IT 65B by a large margin. Table 13 show two

examples from the HotpotQA development set where RA-IT 65B makes a correct prediction while IT 65B makes a wrong prediction. First, we observed that the dense retriever struggles to return useful text chunks for the multi-hop questoins in the HotpotQA dataset and most of the returned text chunks contains no information that helps the prediction. In this case, the IT 65B model shows a stronger tendency to be misled by distractors within the retrieved text chunk, since it has not been trained with noisy passages during fine-tuning. It also tend to predict "I don't know" more frequently[14], while the RA-IT 65B can ignore the noisy passages retrieved and predict the correct answer based on its parametric knowledge (Mallen et al., 2023). We also observe that in cases where both models generate wrong predictions because of the distractors (e.g. for the third text chunk in the second example), the generation probability of the wrong answer from RA-IT 65B is much lower; and in cases where both models ignore the noisy passages and rely on the parametric knowledge to make a prediction, RA-IT 65B outputs the correct answer with a higher probability (e.g. for the second text chunk in the first example).

[14] As discussed in §2.2, this behavior is induced by fine-tuning on SQuAD v2.0 (Rajpurkar et al., 2018), which trains the model to predict "I don't know" for passages that does not match with the given question.

Table 13: Example predictions in HotpotQA (dev set) in the 0-shot setting ensembling 10 retrieved text chunks. The top-3 retrieved chunks and the corresponding model predictions are shown. RA-IT 65B and IT 65B are used to generate these outputs.

Prompt	p_R	Output		nll_{LM}	
		RA-IT	IT	RA-IT	IT
Input: Charlotte Hatherley initially came to prominence in a band formed in what year? **Label:** 1992. **RA-IT 65B final prediction:** 1992 ✓ **IT 65B final prediction:** 1997 ✗					
Background: Charlotte Hatherley Born in London, Hatherley was brought up in West London and attended Chiswick Community School. Her music career began at the age of 15, when she joined British punk band Nightnurse. Two years later, with Ash looking for a guitarist to add to their live sound, Hatherley was hired after frontman Tim Wheeler saw her play at a Nightnurse gig. Hatherley's Ash debut was at Belfast's Limelight on 10 August 1997, and the following week the new lineup played the 1997 V Festival in front of 50,000 people. Her recording career with the band began later that year on the single Ä Life Less Ordinaryänd continued on the album Nu-Clear Sounds in 1998. Hatherley was a full-time member of Ash for eight years, playing on three studio albums, and wrote a handful of the band's songs, most notably Ğrey Will Fade, on the B-side of the single T̈here's a Star. The song was a cult favourite among fans, and eventually became the title track of Hatherley's debut solo album. On 20 January 2006 it was announced that Hatherley would be leaving Ash in an amicable breakup.\n\nQ: Charlotte Hatherley initially came to prominence in a band formed in what year?\nA:	0.27	1992	1997	1.16	1.01
Background: WM: Charlotte Hatherley only... so CD fans might still have to shell out big bucks for an import. Oh, in case you were wondering who Hatherley is, I first heard of her as the g̈irl guitaristïn the band Ash - a band that I have been a fan of since the early 90s when I was getting into all these Britpop-type bands. She naturally started doing her own solo material and left the band a few years ago. The last I heard of her was she was in the band new waver Client with Kate Holmes (not to be confused with the\n\nQ: Charlotte Hatherley initially came to prominence in a band formed in what year?\nA:	0.21	1992	1992	0.46	0.98
Background: Charlotte Hatherley Charlotte Franklin Hatherley (born 20 June 1979) is an English singer, songwriter, guitarist and soundtrack composer. She initially came to prominence as guitarist and backing vocalist for alternative rock band Ash. Since leaving Ash in 2006, she has pursued a solo career and acted as a touring instrumentalist for Bryan Ferry, KT Tunstall, Bat for Lashes, Cold Specks, Rosie Lowe and Birdy. Hatherley has also been a touring member of NZCA Lines and is currently musical director for South African artist Nakhane.\n\nQ: Charlotte Hatherley initially came to prominence in a band formed in what year?\nA:	0.13	1992	I don't know.	0.54	0.72
Input: Oxley Highway ends at a coastal town that had how many inhabitants in June 2016 ? **Label:** 45,698. **RA-IT 65B final prediction:** 45,698 ✓ **IT 65B final prediction:** I don't know. ✗					
Background: Oxley Electorate: Ipswich Motorway: 1 Dec 2016: House debates (OpenAustralia.org) Oxley Electorate: Ipswich Motorway The Ipswich Motorway is a vital link supporting the Queensland economy. It forms part of the national land freight network providing connectivity for industry to the Acacia Ridge intermodal facility, the major industrial area of Wakool and the Brisbane markets at Rocklea2̌014in the member for Morton's electorate2̌014which are the state's largest fruit and vegetable markets and a major centre for produce on the east coast. The section of the motorway is over capacity with 93,000 vehicles on average each day, including up to 12,000 freight vehicles. Numbers are increasing each year at an average of four\n\nQ: Oxley Highway ends at a coastal town that had how many inhabitants in June 2016 ?\nA:	0.25	10,000	I don't know.	7.27	0.61
Background: Post Offices For Sale NSW — Lotto — Newsagencies — Marlow & Co South Wales about 390 km north of Sydney, and 570 km south of Brisbane. The town is located on the Tasman Sea coast, at the mouth of the Hastings River, and at the eastern end of the Oxley Highway. The town with its suburbs had a population of 45,698 in June 2016. Port Macquarie is a retirement destination, known for its extensive beaches and waterways. Port Macquarie has a humid sub-tropical climate with warm, humid summers and mild winters, with frequent rainfall spread throughout the year. Port Macquarie2̌019s central business district contains two shopping centres, a marina, the beginnings of\n\nQ: Oxley Highway ends at a coastal town that had how many inhabitants in June 2016 ?\nA:	0.15	45,698	45,698	0.18	0.38
Background: The Long Paddock - THE LONG PADDOCK The Long Paddock 4x4, 4WD, caravan, camper trailer, camping products reviews, tests, comparisons by Mark Allen The Long Paddock west, the Oxley Highway is the track you2̌019ll be aiming for and Tamworth is the major western town of reference on the map. Once you2̌019re in the main streets of Port, you2̌019ll wonder no more why in excess of 76,000 people now call the area home. As a rough breakdown, the majority of locals are 25 to 44, followed closely by the 45 to 64 year old bracket 2̌013 just perfect for all you thrill seeking middle aged folk and laid back grey nomads and let2̌019s not forget about the younger set that now have oodles of schooling and after-schooling\n\nQ: Oxley Highway ends at a coastal town that had how many inhabitants in June 2016 ?\nA:	0.12	76,000	76,000	4.85	0.93