RA-DIT

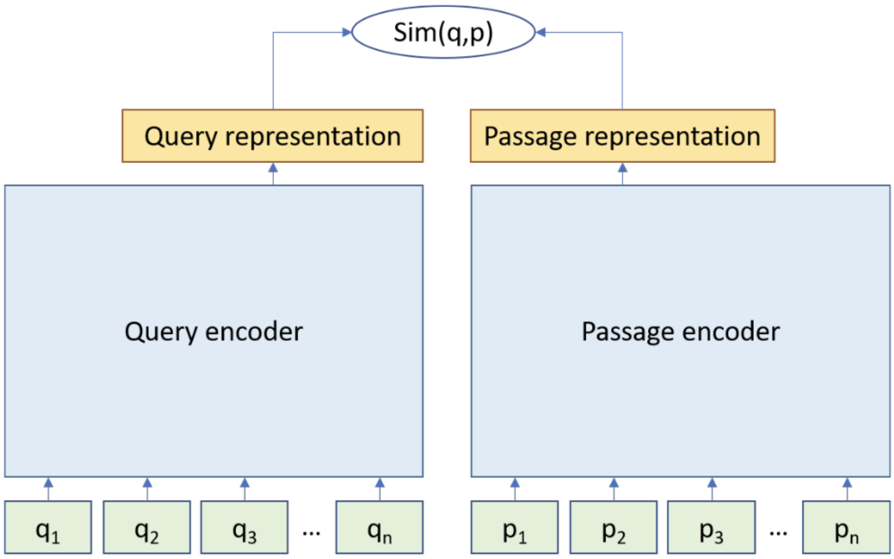
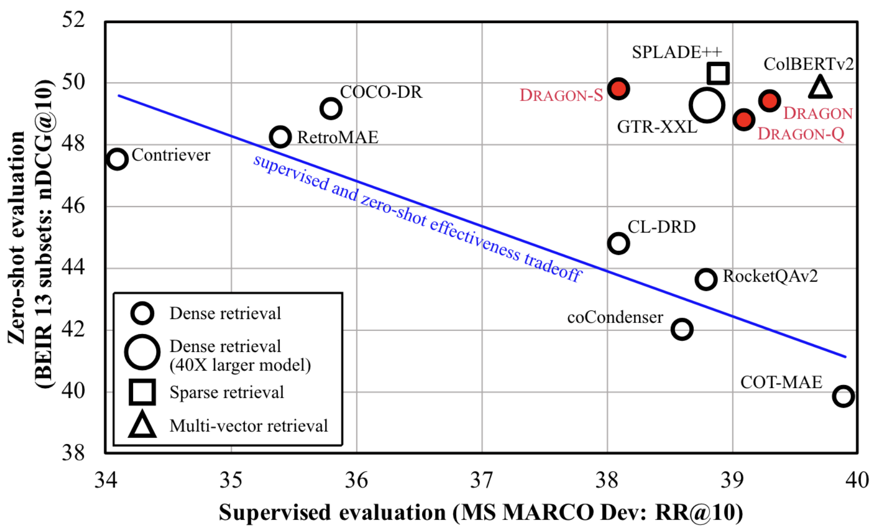
1. 解决什么问题（动机）
    - 需要LLM针对检索得到的知识进行预训练
        - 开销大，成本高
            - atlas：https://arxiv.org/abs/2208.03299
            - REPLUG：https://arxiv.org/pdf/2301.12652
    - 对检索得到的知识进行整合的方法可能会导致模型出现此优的结果
2. 如何解决
    - 解决方案：使用检索的知识微调LLM，使用微调后的LLM生成的回答来微调检索器
    - 实验配置：LLM（LLaMA 65B）、检索器（DRAGON+）
        - DRAGON+
            - 基础模型：BERT-base
            - 

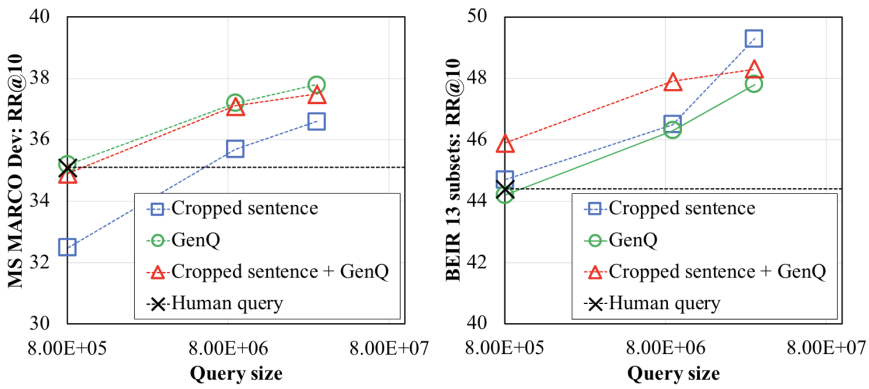
              Figure 1: Bi-encoder architecture for retrieval.
            - 

            - 如何构造对比学习样本?

Table 1: Categorization of existing DR models by their approaches to data augmentation.

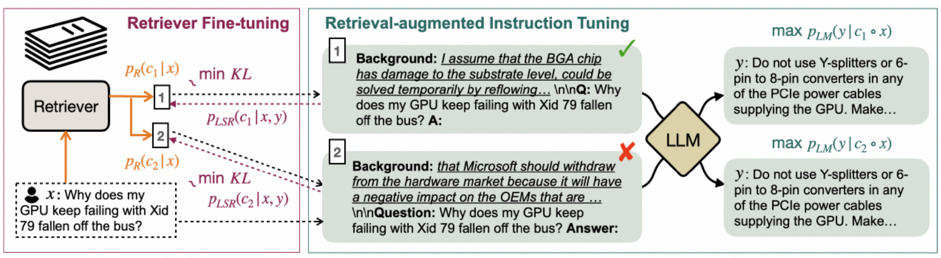
Type	Model	Qry Aug	Label Aug	Corpus
1	RocketQAv2 CL-DRD	✗	CE	MS MARCO
2	coCondenser Contriever COCO-DR	cropping	✗	MS MARCO Wiki+CCnet BEIR
3	GPL PTR	GenQ	✗	BEIR
	DRAGON-S DRAGON-Q DRAGON	cropping GenQ cropping+GenQ	retrievers	MS MARCO

-
    - Query部分：使用裁剪、LLM来做数据增强
    - Label部分：使用教师模型，其中主要用了三个模型（UnCOIL、Contriever、ColBERTv2）
        - top10随机抽取一个作为正样本，46-50随机抽取一个作为负样本作为负样本



(a) MS MARCO Dev (b) BEIR-13

-
- 模型架构



- 训练步骤
    - 第一阶段：使用检索器得到的知识来微调LLM
        - 两个好处：
            - 1.给了模型提示来训练LLM，可以使得模型更好地利用这些提示信息；
            - 2.这些检索器可能会检索到不好的例子，模型在这些bad case的例子下进行优化可以提高模型的鲁棒性
        - prompt准备
            - 依据原始promot来检索相关的文本段
            - 检索到的文本段落拼接到prompt前

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}

- 模型输入长度限制：2048 tokens
- 每个pair使用<bos>和<eos>作为起始和结束分隔符

- 训练损失（加入了阅读理解和摘要抽取）
  - label-loss（预测任务）
  - 预测损失

$$p_R(c|x) = \frac{\exp s(x,c)}{\sum_{c' \in C'} \exp s(x,c')}$$

$$p_{LM}(y|x,\mathcal{C}') = \sum_{c \in \mathcal{C}'} p_{LM}(y|c \circ x) \cdot p_R(c|x),$$

$$\mathcal{L}(\mathcal{D}_L) = -\sum_i \sum_j \log p_{LM}(y_i|c_{ij} \circ x_i).$$

- 使用数据集
  - 20个数据集包含5个不同类别（对话、开源问答、阅读理解、摘要、思维链）
    - 阅读理解使用了数据集：SQuAD 2.0
      - 包含“我不知道”的回复字段，用来训练模型在面对提示无用情况下的回复
    - 注意！
      - 在优化过程中，对于摘要数据集和部分只依赖于上下文的阅读理解任务，作者没有使用检索来增强prompt
      - 对于部分阅读理解任务，会生成k个检索实例+原先就有的实例=k+1个
- 实验参数

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

  - 机器：8, 16 and 64 A100 GPUs

- 关于早停

  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 (?)	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

- 第二阶段：使用更新后LLM生成的答案来微调检索器（只微调query编码器，微调两个编码器导致效果下降！）
  - prompt模版
    - 与第一阶段一样
  - 训练损失–LSR损失
    - $$p_R(c|x) = \frac{\exp s(x,c)}{\sum_{c' \in \mathcal{C}'} \exp s(x,c')}$$
    - $$p_{LSR}(c|x,y) = \frac{\exp\left(p_{LM}(y|c \circ x)/\tau\right)}{\sum_{c' \in \mathcal{C}} \exp\left(p_{LM}(y|c' \circ x)/\tau\right)} \approx \frac{\exp\left(p_{LM}(y|c \circ x)/\tau\right)}{\sum_{c' \in \mathcal{C}'} \exp\left(p_{LM}(y|c' \circ x)/\tau\right)},$$
    - $$\mathcal{L}(\mathcal{D}_R) = \mathbb{E}_{(x,y) \in \mathcal{D}_R} KL\left(p_R(c|x) \parallel p_{LSR}(c|x,y)\right)$$
      - 针对每一个检索的context都计算一次损失
  - 使用数据集
    - 筛选了一部分QA的数据集合、FreebaseQA、MS-MARCO
- 使用的数据集统计

  Table 1: Our intruction tuning datasets. All datasets are downloaded from Hugging Face (Lhoest et al., 2021), with the exception of those marked with ‡, which are taken from Iyer et al. (2022).

Task	HF identifier	Dataset name	$\mathcal{D}_L$	$\mathcal{D}_R$	#Train
Dialogue	oasst1	OpenAssistant Conversations Dataset (Köpf et al., 2023)	✓	✓	31,598
Open-Domain QA	commonsense_qa	CommonsenseQA (Talmor et al., 2019)	✓	✓	9,741
	math_qa	MathQA (Amini et al., 2019)	✓	✓	29,837
	web_questions	Web Questions (Berant et al., 2013)	✓	✓	3,778
	wiki_qa	Wiki Question Answering (Yang et al., 2015)	✓	✓	20,360
	yahoo_answers_qa	Yahoo! Answers QA	✓	✓	87,362
	freebase_qa	FreebaseQA (Jiang et al., 2019)		✓	20,358
	ms_marco*	MS MARCO (Nguyen et al., 2016)		✓	80,143
Reading Comprehension	coqa	Conversational Question Answering (Reddy et al., 2019)	✓		108,647
	drop	Discrete Reasoning Over Paragraphs (Dua et al., 2019)	✓		77,400
	narrativeqa	NarrativeQA (Kočiský et al., 2018)	✓		32,747
	newsqa	NewsQA (Trischler et al., 2017)	✓		74,160
	pubmed_qa	PubMedQA (Jin et al., 2019)	✓	✓	1,000
	quail	QA for Artificial Intelligence (Rogers et al., 2020)	✓		10,246
	quarel	QuaRel (Tafjord et al., 2019)	✓	✓	1,941
	squad_v2	SQuAD v2 (Rajpurkar et al., 2018)	✓		130,319
Summarization	cnn_dailymail	CNN / DailyMail (Hermann et al., 2015)	✓		287,113
Chain-of-thought Reasoning	aqua_rat‡	Algebra QA with Rationales (Ling et al., 2017)	✓		97,467
	ecqa‡	Explanations for CommonsenseQ (Aggarwal et al., 2021)	✓		7,598
	gsm8k‡	Grade School Math 8K (Cobbe et al., 2021)	✓		7,473
	compeition_math‡	MATH (Hendrycks et al., 2021b)	✓		7,500
	strategyqa‡	StrategyQA (Geva et al., 2021)	✓		2,290

  * We only used the question-and-answer pairs in the MS MARCO dataset.

3. 实验结果

- 评估数据集

- 知识密集型任务：不包含于微调任务中的数据集（MMLU、NQ、TriviaQA），还有KILT的6个子集合（HotpotQA、FEVER、AIDA、CoNLL-YAGO、Zero-shotRE、T-REx、Wizard、Wikipedia、ELI5）

- 评估的prompt模版

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}	

- 整体效果

Table 2: Main results: Performance on knowledge intensive tasks (test sets).

	MMLU	NQ	TQA	ELI5	HoPo	FEV	AIDA	zsRE	T-REx	WoW	Avg$^{\diamond}$	Avg
0-shot												
LLAMA 65B	51.2	5.2	55.8	19.5	12.5	59.3	0.6	6.7	1.3	15.6	32.9	22.8
LLAMA 65B REPLUG	59.7	28.8	72.6	19.1	32.0	73.3	41.8	50.8	36.3	16.1	45.1	43.1
RA-DIT 65B	**64.6**	**35.2**	**75.4**	**21.2**	**39.7**	**80.7**	**45.1**	**73.7**	**53.1**	**16.4**	**49.1**	**50.5**
5-shot in-context												
LLAMA 65B	63.4	31.6	71.8	22.1	22.6	81.5	48.2	39.4	52.1	**17.4**	47.2	45.0
LLAMA 65B REPLUG	64.4	42.3	74.9	22.8	**41.1**	89.4	46.4	60.4	**68.9**	16.8	51.1	52.7
RA-DIT 65B	**64.9**	**43.9**	**75.1**	**23.2**	40.7	**90.7**	**55.8**	**72.4**	68.4	17.3	**51.8**	**55.2**

64-shot fine-tuned	NQ	TQA	HoPo	FEV	AIDA	zsRE	T-REx	WoW	Avg
ATLAS[†]	42.4	**74.5**	34.7	**87.1**	66.5	74.9	58.9	15.5	56.8
RA-DIT 65B	**43.5**	72.8	**36.6**	86.9	**80.5**	**78.1**	**72.8**	**15.7**	**60.9**

$^{\diamond}$ Average of MMLU, NQ, TQA, and ELI5.
[†] ATLAS conducts 64-shot fine-tuning for each individual task and evaluates task-specific models individually. For RA-DIT, we perform multi-task fine-tuning using 64-shot examples from each task combined, and report the performance of a unified model across tasks.

- 常识推理任务

Table 3: Performance on commonsense reasoning tasks (dev sets) without retrieval augmentation.

0-shot	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Avg
LLAMA 65B	85.3	82.8	52.3	84.2	77.0	78.9	56.0	**60.2**	72.1
RA-DIT 65B	**86.7**	**83.7**	**57.9**	**85.1**	**79.8**	**83.7**	**60.5**	58.8	**74.5**

- 消融实验
  - 第一阶段模型微调的实验

Table 4: Ablation of language model fine-tuning strategies. All rows report dev set performance.

0 / 5-shot	HoPo	FEV	AIDA	zsRE	T-REx	WoW	Avg
LLAMA 65B	12.5 / 23.8	59.6 / 83.7	0.9 / 64.1	9.7 / 36.0	1.2 / 52.3	15.7 / 17.4	16.6 / **46.2**
IT 65B	20.0 / 30.0	67.8 / 83.2	8.9 / 58.5	19.0 / 35.4	17.3 / 53.5	16.4 / 16.5	24.9 / **46.2**
RA-IT 65B	26.8 / 29.9	65.2 / 84.8	10.7 / 52.9	30.9 / 35.2	24.1 / 52.9	16.5 / 16.5	**29.0** / 45.4
top-1 chunk							
LLAMA 65B + DRAGON+	25.8 / 39.4	72.8 / 89.8	39.1 / 50.7	48.8 / 59.6	31.4 / 69.1	15.8 / 17.1	39.0 / **54.3**
IT 65B + DRAGON+	33.3 / 38.8	84.0 / 90.1	43.9 / 50.3	56.8 / 58.2	44.7 / 66.4	15.7 / 15.6	46.4 / 53.2
RA-IT 65B + DRAGON+	37.6 / 39.1	81.0 / 90.4	41.6 / 52.3	59.6 / 57.9	49.6 / 65.8	16.6 / 16.6	**47.7** / 53.7
top-3 chunks							
LLAMA 65B + DRAGON+	29.6 / 40.8	74.9 / 90.3	43.1 / 52.8	55.9 / 62.9	37.2 / 70.8	16.0 / 17.2	42.8 / **55.8**
IT 65B + DRAGON+	35.2 / 40.0	85.7 / 91.2	49.7 / 52.3	56.2 / 61.9	45.9 / 68.6	15.6 / 15.6	48.1 / 54.9
RA-IT 65B + DRAGON+	39.9 / 40.6	82.4 / 91.7	45.2 / 53.4	63.4 / 61.3	52.8 / 67.6	16.6 / 16.7	**50.1** / 55.2
top-10 chunks							
LLAMA 65B + DRAGON+	31.0 / 41.6	75.4 / 90.8	44.8 / 54.0	58.6 / 63.7	40.2 / 71.9	16.0 / 17.8	44.3 / **56.6**
IT 65B + DRAGON+	33.9 / 40.6	87.0 / 91.8	50.5 / 53.8	53.9 / 62.5	45.7 / 69.4	15.6 / 15.7	47.8 / 55.6
RA-IT 65B + DRAGON+	40.0 / 41.2	82.8 / 92.1	47.2 / 53.5	65.0 / 62.3	54.3 / 69.0	16.5 / 16.6	**51.0** / 55.8

- 第二阶段检索器微调的实验

Table 5: Ablation of retriever fine-tuning strategies. All rows use the LLAMA 65B model and report 5-shot performance on the dev sets.

5-shot	MMLU	NQ	TQA	HoPo	FEV	AIDA	zsRE	T-REx	WoW	Avg°	Avg
DRAGON+	62.6	41.8	72.9	41.5	90.6	54.1	63.7	72.1	17.5	56.6	57.4
MTL instruction tuning data	61.1	43.6	74.0	36.5	91.4	54.6	56.7	72.1	17.1	56.4	57.5
corpus data (FT both encoders)	61.7	43.2	73.8	37.5	88.2	69.8	53.5	57.2	17.5	54.0	55.8
corpus data	62.9	43.0	74.3	41.1	91.6	54.4	63.4	71.8	17.4	56.6	57.8
95% corpus + 5% MTL data	63.0	42.1	74.9	41.2	91.6	54.9	65.2	71.6	17.5	**57.0**	**58.0**

- 检索数量对模型最后效果的影响

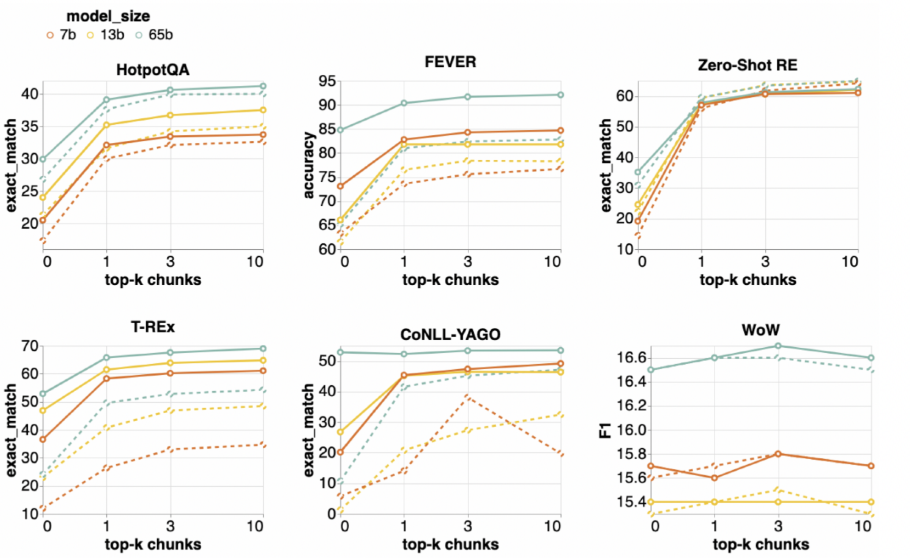


Figure 2: RA-IT model performance (combined with DRAGON+) across sizes 7B, 13B and 65B on our development tasks. 0-shot performance: dashed lines; 5-shot performance: solid lines.

4. 启发（可以借鉴的东西）

- 可以借鉴该文章的思想，微调LLM&检索器
- 更换配置：原先DRAGON+使用的BERT-base，可以考虑使用RoBERT

5. 参考资料：

- Re-Dit论文：https://arxiv.org/pdf/2310.01352
- DRAGON+论文：https://arxiv.org/pdf/2302.07452
  - code：https://github.com/facebookresearch/dpr-scale
- https://yach-doc-shimo.zhiyinlou.com/docs/1Iq7MZORpxFmjxAe/ <RAG技术汇报>