20240711Meeting

In-Context Retrieval-Augmented Language Models

In-Context RALM

- 問題背景
  - 語言模型存在外部知識訪問限制和事實性錯誤問題
- 現有解決方案的缺點
  - 現有的RALM方法通常需要修改LM架構並進行再訓練，增加了部署的複雜性
- 本文的貢獻
  - 提出In-Context RALM，使用現成的LM和檢索器，不需要修改LM架構或進行再訓練
  - 展示了這種方法在各種語料庫和模型規模上的巨大性能增益
- 實驗結果
  - 使用In-Context RALM的LM性能提升相當於增加2–3倍的參數數量
  - 透過文檔排名方法提升性能

Retrieval-Augmented Language Modeling (RALM) system

- Document selection： selecting the set of documents upon which to condition
- Document reading： determining how to incorporate the selected documents into the LM generation process

Figure 2: An example of *In-Context RALM*: we simply prepend the retrieved document before the input prefix.

Datasets & Models

- 5 datasets
  - WikiText-103 (Merity et al., 2016)
  - RealNews (Zellers et al., 2019)
  - from The Pile (Gao et al., 2021): ArXiv, Stack Exchange and FreeLaw
- Models
  - four models of GPT-2 (110M– 1.5B)
  - three models of GPTNeo and GPT-J (1.3B–6B)
  - eight models of OPT (125M–66B)
  - three models of LLaMA (7B–33B)

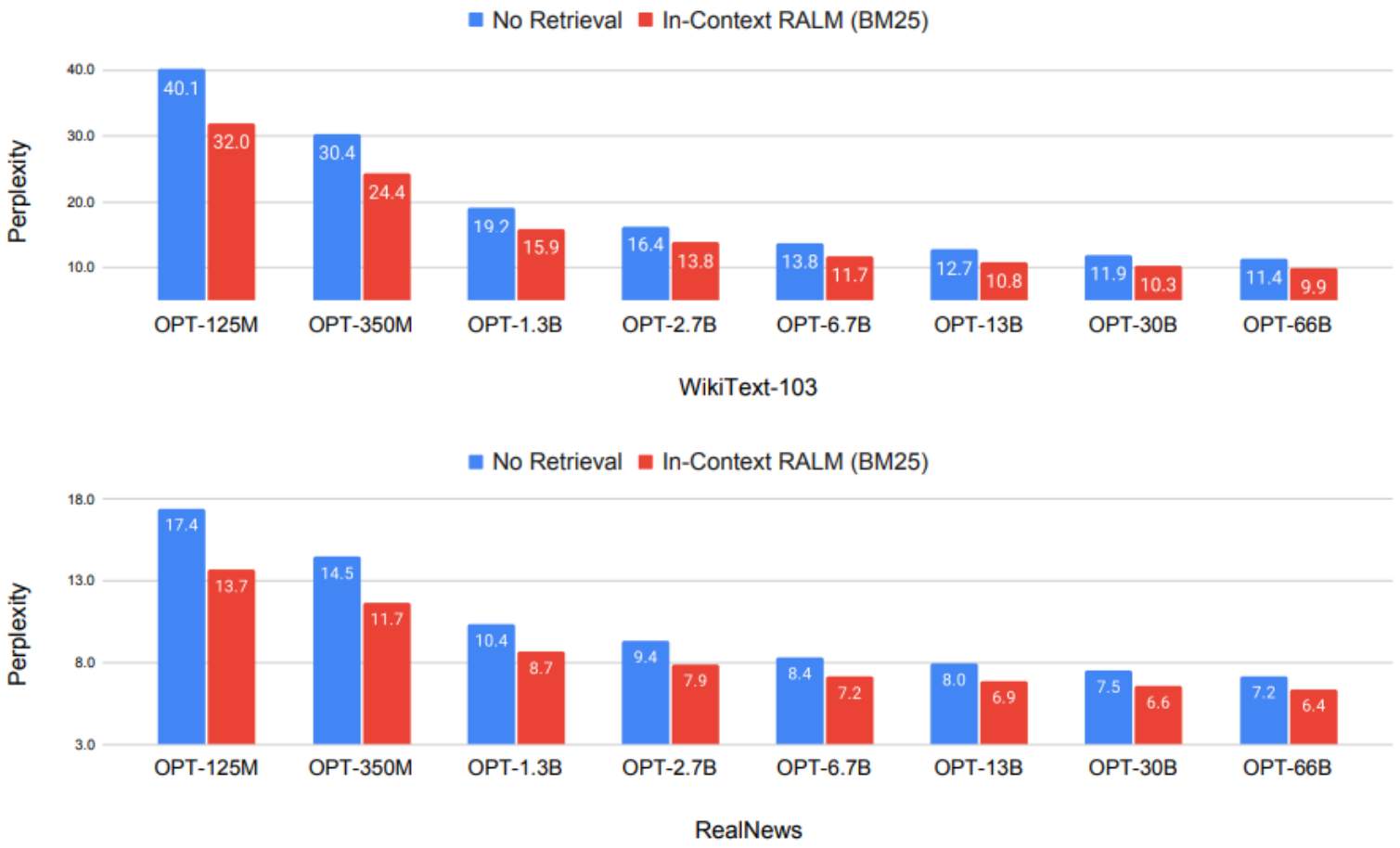
Framework

- 文檔選擇
  - 使用現成的檢索器(如BM25或Dense Passage Retrieval, DPR)從大規模語料庫中選擇與query相關的一組文檔
- 文檔整合
  - 將選擇的文檔預先添加到LM的輸入文本中，並生成輸出文本

RALM Design Choice

- Retrieval Strides (s)
  - 每次檢索之間的跨度，即每隔多少token進行一次檢索
  - 控制檢索的頻率
- Retrieval Query Length (l)
  - 每次檢索所使用的query長度
  - 控制在檢索過程中使用多少上下文信息

Results



Model	Retrieval	WikiText-103 word ppl
LLaMA-7B	-	9.9
	BM25, §5	8.8
LLaMA-13B	-	8.5
	BM25, §5	7.6
LLaMA-33B	-	6.3
	BM25, §5	6.1

Table 2: The performance of models from the LLaMA family, measured by word-level perplexity on the test set of WikiText-103.

Results

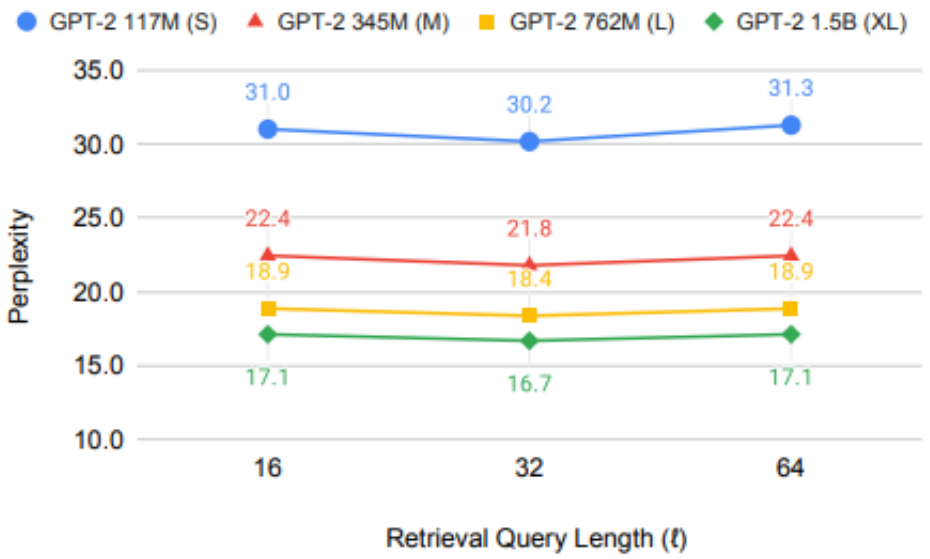


Figure 6: An analysis of perplexity as a function of *the number of tokens in the query* ℓ for BM25 on the development set of WikiText-103. In the appendix, we show similar trade-offs for dense retrievers within WikiText-103. Throughout the paper, we use a query length of $\ell = 32$ tokens.

For sparse BM25 retriever,
retrieval query length(l)=32 was optimal

Results

Figure 9: An analysis of perplexity as a function of *the number of tokens in the query* for an off-the-shelf BERT retriever on the development set of WikiText-103.

Figure 10: An analysis of perplexity as a function of *the number of tokens in the query* for Contriever on the development set of WikiText-103.

For dense retriever, retrieval query length(l)=64 was optimal

Results

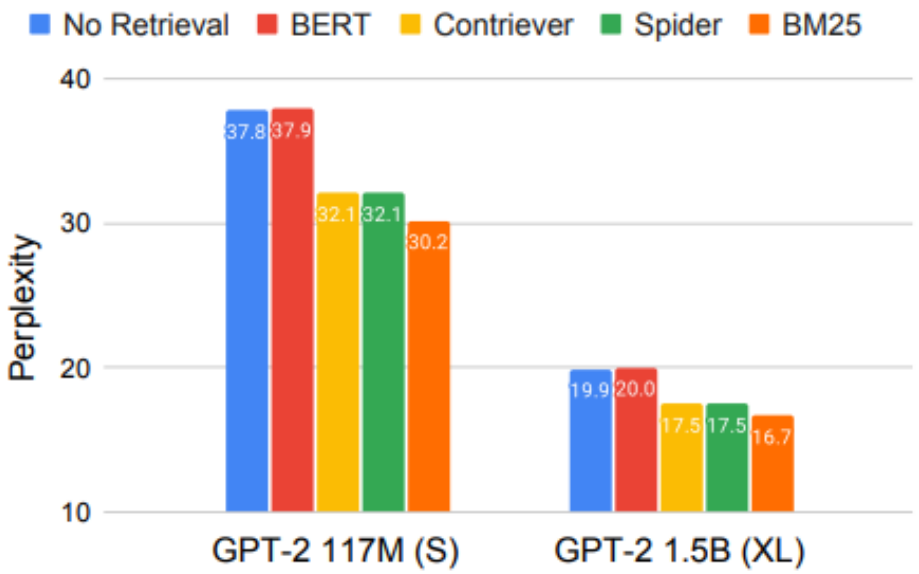


Figure 3: The performance of four off-the-shelf retrievers used for In-Context RALM on the development set of WikiText-103. All RALMs are run with $s = 4$ (*i.e.*, retrieval is applied every four tokens). For each RALM, we report the result of the best query length ℓ (see Figures 6, 9, 10).

BM25 retriever outperforms others

Results

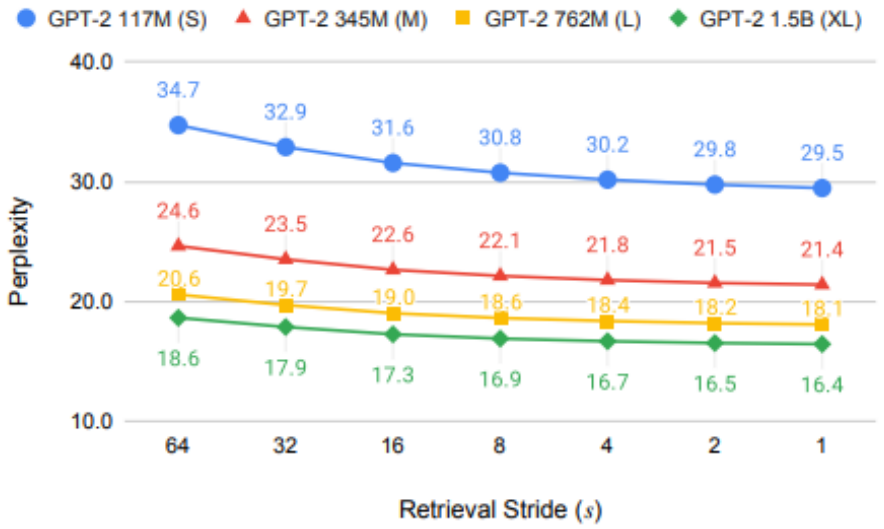


Figure 5: An analysis of perplexity as a function of s, the *retrieval stride*, *i.e.*, the number of tokens between consecutive retrieval operations, on the development set of WikiText-103. Throughout the paper, we use $s = 4$ to balance perplexity and runtime.

LM performance improved as the retrieval operation became more frequent

Improving In-Context RALM

- LM-Oriented Reranking
  - 重新排序從檢索器獲得的文檔，以選擇最有助於生成的文檔
  - 使用語言模型來評估和篩選文檔，而不是僅依賴檢索器的初始排序
- 步驟
  - 檢索：從大規模語料庫中檢索相關文檔
  - 初步篩選：使用LM對這些文檔進行初步排序
  - 重新排序：基於LM的評分重新排序文檔，選擇得分最高的文檔作為最終輸入

Zero-Shot Rerankers

- 無需額外的訓練或調整
- 方法
  - 文檔評估：給定query，使用LM對每個檢索到的文檔進行評分
  - 排序：根據LM的評分結果對文檔進行排序，選擇得分最高的文檔作為最終輸入

Predictive Reranking

- 訓練專門的reranker
- 根據預測模型的評分結果對檢索到的文檔進行重新排序
- 方法：
  - 初步檢索：從大規模語料庫中檢索初步相關的文檔集合。
  - 預測評分：使用預測模型對每個文檔進行評分，評估其與query的相關性。
  - 重新排序：根據預測模型的評分結果對文檔進行排序，選擇得分最高的文檔作為最終輸入。

Results

Table 1: Perplexity on the test set of WikiText-103, RealNews and three datasets from the Pile.

Model	Retrieval	Reranking	WikiText-103 word ppl	RealNews token ppl	ArXiv token ppl	Stack Exch. token ppl	FreeLaw token ppl
GPT-2 S	–	–	37.5	21.3	12.0	12.8	13.0
	BM25 §5	–	29.6	16.1	10.9	11.3	9.6
	BM25	Zero-shot §6.1	28.6	15.5	10.1	10.6	8.8
	BM25	Predictive §6.2	26.8	–	–	–	–
GPT-2 M	–	–	26.3	15.7	9.3	8.8	9.6
	BM25 §5	–	21.5	12.4	8.6	8.1	7.4
	BM25	Zero-shot §6.1	20.8	12.0	8.0	7.7	6.9
	BM25	Predictive §6.2	19.7	–	–	–	–
GPT-2 L	–	–	22.0	13.6	8.4	8.5	8.7
	BM25 §5	–	18.1	10.9	7.8	7.8	6.8
	BM25	Zero-shot §6.1	17.6	10.6	7.3	7.4	6.4
	BM25	Predictive §6.2	16.6	–	–	–	–
GPT-2 XL	–	–	20.0	12.4	7.8	8.0	8.0
	BM25 §5	–	16.6	10.1	7.2	7.4	6.4
	BM25	Zero-shot §6.1	16.1	9.8	6.8	7.1	6.0
	BM25	Predictive §6.2	15.4	–	–	–	–

Sparse BM25 retriever
Retrieval query length(l)=32
Retrieval strides(s)=4