Instruction-tuned Language Models are Better Knowledge Learners

Zhengbao Jiang[2*] **Zhiqing Sun**[2] **Weijia Shi**[1,3] **Pedro Rodriguez**[1] **Chunting Zhou**[1]
Graham Neubig[2] **Xi Victoria Lin**[1] **Wen-tau Yih**[1] **Srinivasan Iyer**[1]
[1]FAIR at Meta [2]Carnegie Mellon University [3]University of Washington
{zhengbaj,gneubig}@cs.cmu.edu {victorialin,scottyih,sviyer}@meta.com

Abstract

In order for large language model (LLM)-based assistants to effectively adapt to evolving information needs, it must be possible to update their factual knowledge through continued training on new data. The standard recipe for doing so involves continued pre-training on new documents followed by instruction-tuning on question-answer (QA) pairs. However, we find that LLMs trained with this recipe struggle to answer questions, even though the perplexity of documents is minimized. We found that QA pairs are generally straightforward, while documents are more complex, weaving many factual statements together in an intricate manner. Therefore, we hypothesize that it is beneficial to expose LLMs to QA pairs *before* continued pre-training on documents so that the process of encoding knowledge from complex documents takes into account how this knowledge is accessed through questions. Based on this, we propose **pre-instruction-tuning (PIT)**, a method that instruction-tunes on questions prior to training on documents. This contrasts with standard instruction-tuning, which learns how to extract knowledge after training on documents. Extensive experiments and ablation studies demonstrate that PIT significantly enhances the ability of LLMs to absorb knowledge from new documents, outperforming standard instruction-tuning by 17.8%.

1 Introduction

Large language models (LLMs) store vast amounts of factual knowledge in their parameters through large-scale pre-training, and this knowledge can be used to answer various questions such as "where is the world's largest ice sheet located" (Brown et al., 2020; OpenAI, 2023; Chowdhery et al., 2022; Zhang et al., 2022; Touvron et al., 2023a,b; Gemini Team, 2023). However, this factual knowledge is static, meaning that it can become outdated as the world evolves, or prove insufficient when LLMs are used in specialized or private domains.

To keep LLMs up-to-date, it is common to continue pre-training on new documents to store knowledge in parameters, which allows LLMs to effectively answer queries that require up-to-date information (Jang et al., 2022). A widely held view is that the factual knowledge stored in parameters can be elicited through prompting (Brown et al., 2020; Petroni et al., 2019; Roberts et al., 2020), and that instruction-tuning (also known as supervised fine-tuning or alignment) makes this elicitation more effective (Sanh et al., 2022; Wei et al., 2022; Ouyang et al., 2022). In the first part of this paper (§ 4), we conduct extensive experiments using Llama-2 (Touvron et al., 2023b) to answer the following question: *to what extent can we augment the knowledge stored in modern LLMs by continued pre-training on new documents, either with or without subsequent instruction-tuning*? We find that, as we train LLMs repeatedly over documents to the extent that perplexity is minimized to one, the percentage of questions regarding those documents that LLMs answer correctly increases consistently to 27.6%. Subsequent instruction-tuning further improves it to 30.3%, confirming that this widely used practice is useful to elicit more knowledge from LLMs.[1] However, the amount of elicited knowledge is still limited, even though the perplexity of documents is minimized, a phenomenon we refer to as the "perplexity curse".[2]

In the second part of the paper (§ 5), we study methods to mitigate the perplexity curse by making LLMs more adept at absorbing knowledge from documents. Zhu and Li (2023a) presented an intriguing finding that training a randomly initialized

[*]Majority of the work done during an internship at Meta.

[1]This capacity might be underestimated by previous works due to using relatively small LMs or randomly initialized transformers, or lack of exhaustive training or instruction-tuning (Wang et al., 2021; Hu et al., 2023; Zhu and Li, 2023a).

[2]Inspired by the "reversal curse" of Berglund et al. (2023).

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5421–5434
August 11-16, 2024

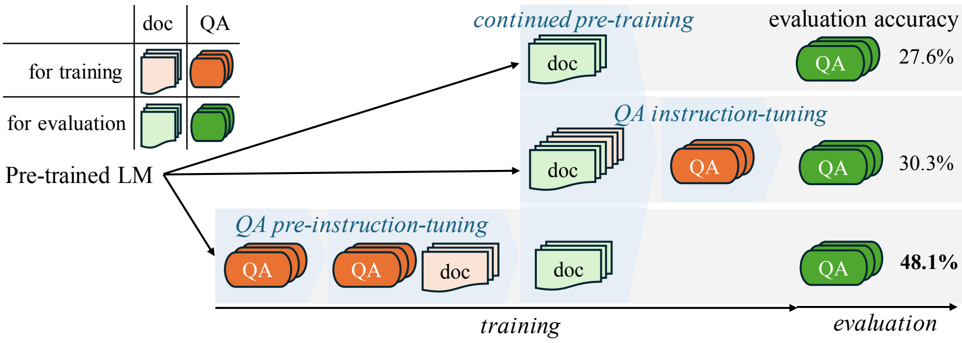


Figure 1: Illustration of continued pre-training (first row), continued pre-training followed by instruction-tuning (second row), and pre-instruction-tuning before continued pre-training (last row), along with their accuracies on evaluation questions. Each right-pointing light-blue triangle indicates a training phase.

transformer from scratch on a mix of biographies and related questions resulted in strong generalization to new questions. However, understanding the reasons behind this finding and exploring ways to practically apply it for absorbing knowledge from new documents requires further investigation. We found that question-answer (QA) pairs are generally straightforward and easily digestible, while documents tend to be more complex and cluttered, often weaving many factual statements together in a more intricate manner. Therefore, we hypothesize that *it is beneficial to deliberately expose LLMs to QA data before continued pre-training on documents so that the process of encoding knowledge from complex documents takes into account how this knowledge is accessed through questions*. We refer to this as **pre-instruction-tuning (PIT)** and conduct comprehensive experiments to benchmark different variations of this method. As shown in Fig. 1, our best-performing variation starts with training exclusively on QA pairs (e.g., "who handled the editing of Oppenheimer") to grasp how knowledge is accessed. This is followed by training on a combination of these QA pairs and associated documents (e.g., "who handled the editing of Oppenheimer" and a document about "Oppenheimer"). In this phase, LLMs enhance their ability to absorb knowledge from information-dense documents, building upon the QA pairs that they have already mastered. To study continual knowledge acquisition, we build a dataset named `Wiki2023`, which includes a collection of documents from Wikipedia that are relevant to the year 2023. Comprehensive experiments on `Wiki2023` demonstrate that after PIT, LLMs exhibit an enhanced ability to absorb knowledge from new documents (e.g., a document about "Barbie"). Detailed ablation studies reveal that this ability primarily stems from prioritizing learning how to access knowledge over learning to encode knowledge from documents. Overall, PIT significantly outperforms the standard instruction-tuning approach (§ 5.1 and § 5.2), improving QA accuracies by 17.8% on Llama-2 7B (30.3% → 48.1%) and 16.3% on Llama-2 70B (46.4% → 62.7%). Moreover, PIT also enhances the ability to absorb knowledge from documents of a *different* domain, shedding light on the potential to scale this method up to a wider variety of documents and instructions for more robust generalization (§ 5.4).

2 Building a Dataset to Study Continual Knowledge Acquisition

To assess the ability of LLMs to learn knowledge from new documents, it is essential to use a document corpus with minimal overlap with the original pre-training corpus. This ensures that when an LLM correctly answers questions, we can confidently attribute this capability to its learning from the new documents, rather than encountering similar questions in its original pre-training corpus. In this section, we describe a methodology for building such a corpus from Wikipedia.

2.1 Wiki2023 Document Corpus

In the following experiments (§ 4 and § 5), we use Llama-2 (7B and 70B) (Touvron et al., 2023b) since it is one of the best-performing LLMs. We use Wikipedia articles classified under the "2023" Category including topics from diverse domains such as films, arts, economics, politics, events, etc.[3] The likelihood that this factual information is not included in the original training corpus is supported by the low QA performance in Tab. 1 (9.5%/17.2%

[3] `https://en.wikipedia.org/wiki/Category:2023`

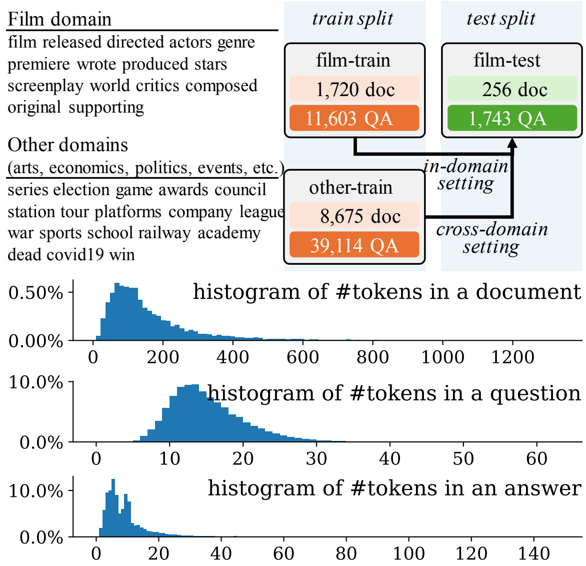


Figure 2: The `Wiki2023` dataset. **Top-right**: the number of documents and QA pairs; **Top-left**: frequent keywords in questions; **Bottom**: the distribution of token counts in documents, questions, and answers.

An example document about "Oppenheimer"
<bos> Oppenheimer (OP-ən-hy-mər) is a 2023 epic biographical thriller film written and directed by Christopher Nolan. It stars Cillian Murphy as J. Robert Oppenheimer, … the film chronicles the career of Oppenheimer, with the story predominantly focusing on his studies, his direction of the Manhattan Project during World War II, and his eventual fall from grace due to his 1954 security hearing. … **Editing was handled by Jennifer Lame**, and the score was composed by Ludwig Göransson. … Oppenheimer premiered at Le Grand Rex in Paris on July 11, 2023, and was theatrically released …

Example QA about "Oppenheimer"
<bos> Question: Who wrote and directed the film Oppenheimer?
Answer: Christopher Nolan. <eos>
<bos> Question: Who stars as J. Robert Oppenheimer in the film?
Answer: Cillian Murphy. <eos>
<bos> Question: What aspects of Oppenheimer's life does the film focus on?
Answer: His studies, direction of the Manhattan Project, and 1954 security hearing. <eos>
<bos> **Question: Who handled the editing of Oppenheimer?**
Answer: Jennifer Lame. <eos>
<bos> Question: When did Oppenheimer premiere in Paris?
Answer: July 11, 2023. <eos>

Figure 3: An example document about "Oppenheimer" and corresponding QA pairs from `Wiki2023`. Tokens used for computing losses are highlighted in green.

for 7B/70B).[4] To accelerate the training process, we only use the first section of each article, which offers a thorough summary and contains many factual statements. The number of collected documents and an example document about "Oppenheimer" can be found in Fig. 2 and Fig. 3. We refer to this as the `Wiki2023` dataset.

2.2 Wiki2023 Question-answer Pairs

To collect QA pairs for either instruction-tuning or performance evaluation, we employ publicly available LLMs to generate diverse questions and their respective answers given the article as context, following the Prompt 1 in Appendix A. On average, 4.93 questions are generated for each article. Fig. 2 and Fig. 3 show the detailed statistics and example QA pairs about "Oppenheimer", respectively.

2.3 Splits

Among all domains, we select the film domain for evaluation and randomly select 256 articles as the test split (`Wiki2023-film-test`). We continually train LLMs on documents from the test split (`Wiki2023-film-test-doc`), and assess their performance based on the accuracy of corresponding questions (`Wiki2023-film-test-QA`). The remaining 1720 articles and corresponding QA pairs (`Wiki2023-film-train`) will be used to study different training strategies, which corresponds to the in-domain setting in Fig. 2. We also train on other domains before evaluation on the film domain to study the effectiveness of different methods across domains, which corresponds to the cross-domain setting in Fig. 2.

3 Experimental Settings

3.1 Objectives

When training on documents, we prepend a <bos> token and compute the standard next-token prediction loss by averaging over all tokens in the document: $L_{\boldsymbol{d}} = -\sum_t \log P(\boldsymbol{d}_t|\boldsymbol{d}_{<t})/|\boldsymbol{d}|$.[5] When training on QA pairs, we compute the average negative log-likelihood loss only on tokens in the answer given the question as the prefix: $L_{\boldsymbol{a}} = -\sum_t \log P(\boldsymbol{a}_t|\boldsymbol{q}, \boldsymbol{a}_{<t})/|\boldsymbol{a}|$. Fig. 3 presents an example document alongside QA pairs, where tokens used for computing losses are highlighted.

3.2 Hyperparameters

When pre-training on documents, we use a batch size of 256 documents and an initial learning rate of 3e-5. During instruction-tuning on QA pairs, we use the same batch size of 256 QA pairs but opt for a reduced initial learning rate of 5e-6 because the number of tokens in a single batch is lower. Details can be found in Appendix B.

3.3 Evaluation Metrics

Since most answers are relatively short, we use exact match (EM) as our primary metric

[4]It is important to note the difficulty in completely avoiding factual overlap between `Wiki2023` and the pre-training corpus of Llama-2. For example, a film released in 2023 might have had information available before 2023. Data duplication detection is an active research direction, which falls beyond the focus of this study.

[5]We do not append a <eos> token at the end of documents because we only use the first section, which does not signify the conclusion of the entire article.

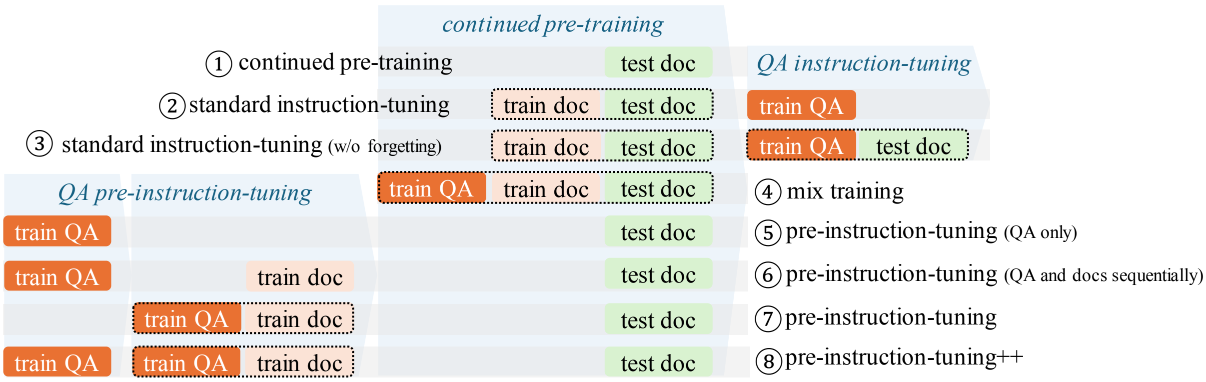


Figure 4: Different experimental settings examined in this paper. Each row represents a different experimental setting with a unique name and number, and each vertical section highlighted by a right-pointing light-blue triangle indicates a training phase. Models are assessed on test QA across all settings. Whenever multiple datasets are enclosed within a dashed square, they are mixed together during the training process.

(Kwiatkowski et al., 2019). To assess longer responses and accommodate minor lexical differences, we also report answer recall and ROUGE-L. Details can be found in Appendix C.

4 How Much Knowledge Can LLMs Absorb via Continued Pre-training Followed by Instruction-tuning?

Factual knowledge stored in the parameters of LLMs can be accessed and applied to answering questions through prompting without additional training (Brown et al., 2020; Petroni et al., 2019; Jiang et al., 2020; Roberts et al., 2020). With additional instruction-tuning (also known as supervised fine-tuning) on high-quality data (Sanh et al., 2022; Wei et al., 2022), knowledge seems to be more effectively elicited from LLMs. However, when LLMs correctly answer a question, the source of the knowledge is unclear due to the diversity of the pre-training data. For instance, when answering the question "where is the world's largest ice sheet located", do LLMs derive their response by recalling and generalizing information from a seen document about the Antarctic ice sheet, or do they merely repeat answers from similar questions encountered in the training data? This distinction is crucial, as the former scenario implies an ability to comprehend documents and effectively store knowledge within parameters in a way that can be elicited later, whereas the latter is mere rote memorization.

Several works have studied this problem and the predominant finding is that LMs struggle to answer questions about documents they have been trained on (Wang et al., 2021; Zhu and Li, 2023a). It is important to note, however, that these experiments were mainly conducted using relatively small LMs such as BART, T5, or GPT-2 (Wang et al., 2021; Jang et al., 2022; Hu et al., 2023), using randomly initialized transformers (Zhu and Li, 2023a), or without instruction-tuning (Ovadia et al., 2023). This makes us wonder *what are the actual limits of modern LLMs to absorb knowledge from new documents and answer questions about them using the standard continued pre-training followed by instruction-tuning recipe*. In this section, we run extensive experiments using Llama-2 7B and 70B on `Wiki2023-film` to test their limits.

4.1 Vanilla Continued Pre-training and Instruction-tuning

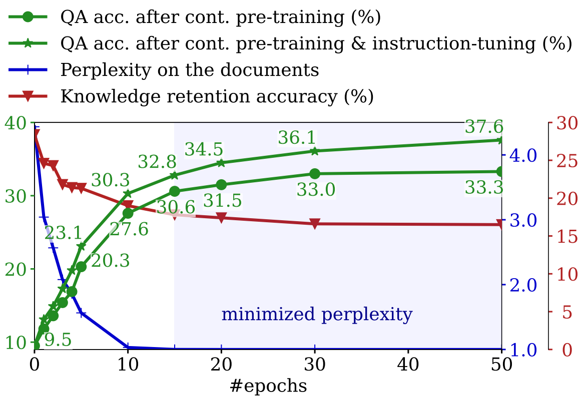
Experimental settings We experiment with two standard settings and assess their performance by answering associated questions.

- Continued pre-training: train on test documents without instruction-tuning (Fig. 4 ①).[6]
- Standard instruction-tuning: train on both train and test documents before instruction-tuning on train QA pairs (Fig. 4 ②).

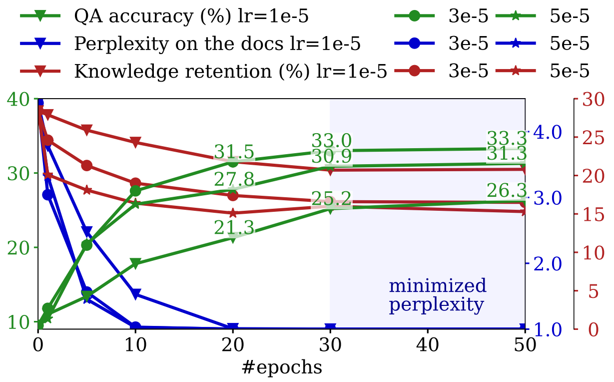
We perform instruction-tuning for a single epoch since more epochs usually result in diminished performance. For training on documents, we opt for multiple epochs (10/5 for a 7B/70B model), which allows for effective knowledge acquisition and remains affordable for corpora of moderate sizes.

Experimental results As shown in Tab. 1, the relatively low performance of the original Llama-2 model (9.5%/17.2% for 7B/70B) indicates that

[6]We found that LLMs struggle to adhere to the QA format after training on raw documents for multiple epochs. Therefore, we include a small set of QA pairs (64) during continued pre-training to prevent LLMs from forgetting the QA format.



(a) Training dynamics w/ (Fig. 4 ②) and w/o instruction-tuning (Fig. 4 ①). Reduction in perplexity consistently leads to improvement in QA accuracy, indicating that factual knowledge acquisition necessitates exhaustive loss minimization.



(b) Training dynamics with different learning rates (Fig. 4 ①). After perplexity is minimized, larger learning rates usually lead to less overfitting to deceptive patterns in documents and better generalization when responding to questions.

Figure 5: We vary the number of epochs (Fig. 5(a)) and learning rate (Fig. 5(b)) during continued pre-training to study the training dynamics of Llama-2 7B. The left axis is QA accuracies for test questions, measured by exact match. On the right axis, we display 2 metrics indicated by distinct colors: the perplexity of all tokens in the documents, and the knowledge retention accuracy, measured by QA accuracy on the Natural Questions dataset. We highlight situations where perplexity of all document tokens is minimized to 1 .

most knowledge in the test documents is not included in the original pre-training corpus. After continued pre-training on documents, performances increase to 27.2%/41.7%, indicating that LLMs can absorb some amount of knowledge. Instruction-tuning further increases the performance to 30.3%/46.4%, confirming the effectiveness of this standard recipe. This observation is different from Zhu and Li (2023a), which demonstrates that instruction-tuning after pre-training is ineffective on a randomly initialized GPT-2-like transformer. The difference probably arises because Llama-2, through its pre-training on diverse corpora comprising raw documents and QA data, has developed a certain degree of proficiency in extracting knowledge from its parameters via questions. We also report the performance where the corresponding document is directly provided to Llama-2 as context ("open-book w/ doc" in Tab. 1). The significant gap between closed-book and open-book settings suggests that retrieving knowledge from the parameters of LLMs is still challenging.

4.2 Analyzing the Training Dynamics: Perplexity and Generalization

How does lower perplexity of documents lead to generalization to answering related questions? We vary the number of epochs (Fig. 5(a)) and learning rate (Fig. 5(b)) for continued pre-training on documents and monitor three metrics to study the training dynamics.[7]

- **Knowledge acquisition** QA accuracies on test questions measured by exact match.
- **Perplexity of documents** We compute perplexity (PPL) on all tokens within the documents.
- **Knowledge retention** We approximate the retention of accumulated knowledge during pre-training by assessing the QA accuracy on the Natural Questions (NQ) dataset. NQ was released in 2019, and primarily includes questions based on Wikipedia articles from that time.

Experiment results

- As shown in Fig. 5(a), QA accuracy consistently improves as perplexity approaches one, indicating that *factual knowledge learning necessitates exhaustive loss minimization over all tokens*. This contrasts with learning general skills, where overly optimizing leads to overfitting.
- As shown in Fig. 5(a) and Fig. 5(b), among all cases where LLMs have minimized perplexity on documents, for reasonably small learning rates (5e-5 is too large and leads to overfitting), cases trained with more epochs or larger learning rates typically exhibit superior QA performance. We

[7]Since we always decay the learning rate to 10% of its initial value, training for more epochs is not the same as continuing training from a checkpoint obtained after fewer epochs.

Settings	Llama-2 7B EM	Rec.	R-L	Llama-2 70B EM	Rec.	R-L
closed- and open-book performance before training						
closed-book	9.5	10.0	21.2	17.2	18.1	31.4
open-book w/ doc	72.2	75.4	91.5	78.2	80.6	94.9
closed-book performance w/ standard methods						
cont. pre-training ①	27.6	31.6	43.8	41.7	45.8	60.2
+instruction-tuning ②	30.3	34.7	47.4	46.4	50.9	64.1
mix all data ④	39.4	44.6	56.7	57.1	63.4	72.4
closed-book performance w/ pre-instruction-tuning (PIT)						
PIT (QA only) ⑤	28.6	32.7	45.2	49.7	53.7	67.9
PIT (QA → docs) ⑥	32.5	37.2	49.0	54.6	60.0	73.8
PIT ⑦	**45.4**	**51.2**	**63.2**	**62.7**	**68.6**	**78.8**

Table 1: Comparison of QA performance (%) between standard instruction-tuning and pre-instruction-tuning. The best results are in bold. Rec. is short for answer recall, and R-L refers to ROUGE-L.

hypothesize that *more aggressive training leads to less overfitting to deceptive patterns in documents and better generalization when responding to questions*.

In summary, lower perplexity does lead to stronger generalization when responding to questions, but it comes at the expense of forgetting previously acquired knowledge.

5 Improving LLMs in Absorbing Knowledge from Documents

The amount of knowledge elicited through the standard instruction-tuning is still limited, even though the perplexity of documents is minimized, a phenomenon we refer to as the "perplexity curse". Our next question is how can we improve the ability of LLMs to absorb knowledge from documents to mitigate the perplexity curse. The main challenge is the gap between the way knowledge is presented in raw documents and how it is accessed through question-answering. We found that QA pairs are generally straightforward, while documents tend to be more complex and cluttered, weaving many factual statements together in a more intricate manner. Using Fig. 3 as an example, the answer to the question "who handled the editing of Oppenheimer" is included in a sentence in the middle of the article "Editing was handled by Jennifer Lame ...", which does not explicitly mention "Oppenheimer". During training, LLMs must understand the context and deduce that "editing" refers to "the editing of Oppenheimer" to effectively encode this knowledge in the parameters.

Zhu and Li (2023a) studied this problem by training a randomly initialized GPT-2-like transformer from scratch on synthetic biographies and evaluated its ability to answer questions about the individuals. They found that training on a mix of biographies and questions related to half of those biographies led to strong generalization when answering questions about the remaining half of biographies, which resembles setting ④ in Fig. 4. In contrast, training on biographies and QA pairs sequentially failed. However, the key contributor to the success remains uncertain because the data were blended together, and it is unclear how to apply this practically to absorb knowledge from new documents. Inspired by our observation of the different difficulty levels between QA pairs and documents, and the finding from Zhu and Li (2023a), we hypothesize that *it is beneficial to deliberately expose LLMs to instruction-tuning data before continued pre-training so that the process of encoding knowledge from complex documents takes into account how this knowledge is accessed.* We refer to this as **pre-instruction-tuning (PIT)** and study various implementations of PIT prior to continued learning (§ 5.1), followed by detailed ablations identifying the keys contributor to performance (§ 5.2 and § 5.3), and finally assess how well PIT performs across domains (§ 5.4). We adhere to the hyperparameters outlined in § 3.2 and perform PIT for 3 epochs unless specified otherwise.

5.1 Variants of Pre-instruction-tuning

Pre-instruction-tuning w/ QA only We start with exposing instruction-tuning data before continued pre-training on documents—training on topically related QA pairs before training on test documents (Fig. 4 ⑤). This can be directly compared with the continued pre-training setting (Fig. 4 ①). The intuition is that questions help LLMs recognize key types of information, enabling LLMs to focus on important information during pre-training on subsequent documents, even though the questions are not directly tied to the documents. For example, training on a question like "who handled the editing of Oppenheimer" could help LLMs pay attention to screenwriters when training on new documents like "Barbie". As shown in Tab. 1, this method outperforms continued pre-training, especially on larger LLMs (27.6%/41.7% → 28.6%/49.7% for 7B/70B). The ablation that trains on QA data after training on documents ("instruction-tuning w/o train doc" in Tab. 2) is ineffective, confirming the importance of training on questions as a warm-up before encoding documents.

Setting names	Setting configurations	EM	Rec.	R-L
	baselines			
continued pre-training ①	test doc	27.6	31.6	43.8
+instruction-tuning ②	train doc + test doc → train QA	30.3	34.7	47.4
+instruction-tuning (w/o forget) ③	train doc + test doc → train QA + test doc	30.2	34.1	46.4
+instruction-tuning (w/o train doc)	test doc → train QA	27.1	30.7	42.3
weighted continued pre-training	test doc (weighted)	27.7	32.7	43.3
adapted continued pre-training	train doc → test doc	26.9	32.7	44.2
mix all data ④	train QA + train doc + test doc	39.4	44.6	56.7
	various pre-instruction-tuning (PIT) methods and ablation studies			
PIT ⑦	train QA + train doc (3 epochs) → test doc	45.4	51.2	63.2
	ablation studies of the number of epochs			
	1 epoch	33.3	39.1	50.3
	5 epochs	45.8	52.1	63.6
	10 epochs	46.5	52.3	61.9
	ablation studies of different learning mechanisms			
	QA before doc (grouped)	38.2	43.2	56.3
	QA after doc (grouped)	27.2	31.1	42.1
	QA before doc (interleaved)	45.9	51.3	64.5
	QA after doc (interleaved)	43.2	49.1	61.6
PIT--	train QA + train doc → train QA → test doc	44.4	51.3	63.4
PIT++ ⑧	train QA → train QA + train doc → test doc	**48.1**	**54.4**	**66.4**

Table 2: Comparison (%) of various pre-instruction-tuning methods and ablation studies to identify the key contributors to improved performance using Llama-2 7B. Different background colors indicate different pre-instruction-tuning methods. The best results are in bold.

Pre-instruction-tuning on QA and documents sequentially Our second implementation trains on QA and associated documents sequentially (Fig. 4 ⑥), with the intuition that the ability to absorb knowledge from documents can be strengthened if an LLM is trained on the complex documents after it has grasped the associated simpler QA pairs. For instance, if an LLM has already learned that "Jennifer Lame" is the answer to "who handled the editing of Oppenheimer", training on the document "Editing was handled by Jennifer Lame" can more efficiently refine its storage of knowledge in its parameters. As shown in Tab. 1, PIT on QA pairs and documents sequentially surpasses the QA-only variant (Fig. 4 ⑤) and standard instruction-tuning (Fig. 4 ②) (30.3%/46.4% → 32.5%/54.6% for 7B/70B), demonstrating its effectiveness.

Pre-instruction-tuning The effectiveness of PIT depends on ensuring that the associated QA pairs are already learned before encoding the respective documents. However, we observed that after training on documents (train doc in Fig. 4 ⑥), the accuracy for corresponding questions (train QA in Fig. 4 ⑥) dropped from almost perfect to 30%, indicating severe forgetting. To fix this, we train on the associated QA pairs and documents together (Fig. 4 ⑦). As shown in Tab. 1, this significantly improves the performance, outperforming all other approaches, including mixing all data together (Fig. 4 ④), by a large margin (39.4%/57.1% → 45.5%/62.7% for 7B/70B). Training on both QA pairs and documents prevents forgetting, but it also obscures how the learning process works. It is unclear whether LLMs grasp QA pairs before encoding knowledge from documents, or if it works the other way around. In the following section, we deliberately arrange the order of QA pairs and documents during training to examine this, which leads us to propose an improved version of PIT.

5.2 Pre-instruction-tuning++

We first study how the performance varies with different numbers of epochs. As shown in Tab. 2, training for 1 epoch is insufficient, and the performance of 3, 5, or 10 epochs is similar. We fix the number of epochs to 3 and arrange the order of QA pairs and corresponding documents as shown in Fig. 6 in Appendix D. The interleaved arrangement cycles through all the data 3 times, ensuring that in each epoch, questions either precede or follow their associated documents. On the other hand, the grouped arrangement clusters each example's 3 appearances together, guaranteeing that the repeated questions are positioned either before or after their respective repeated documents. As shown in Tab. 2, positioning QA pairs before corresponding documents achieves better performance in both grouped and interleaved arrangements, indicating that during PIT, the learning mechanism prioritizes under-

Settings	Llama-2 7B EM	Rec.	R-L	Llama-2 70B EM	Rec.	R-L
standard instruction-tuning ②						
in-domain	30.3	34.7	47.4	46.4	50.9	64.1
cross-domain	23.6	28.2	38.4	42.8	49.7	58.5
pre-instruction-tuning ⑦						
in-domain	45.4	51.2	63.2	62.7	68.6	78.8
cross-domain	36.9	43.2	54.9	55.2	66.7	74.0

Table 3: In-domain and cross-domain PIT.

Settings	EM	Rec.	R-L
generalization to the biography dataset `bioS`			
closed-book	2.9	2.9	11.0
open-book w/ doc	95.2	95.4	95.6
continued pre-training ①	29.6	29.8	38.7
pre-instruction-tuning ⑦	**58.1**	**58.4**	**61.9**
generalization to questions by real users from Google			
standard instruction-tuning ②	21.5	30.1	36.8
pre-instruction-tuning ⑦	**29.0**	**35.5**	**48.2**

Table 4: Generalization of the Llama-2 7B model trained with pre-instruction-tuning.

standing how to access knowledge before learning to absorb information from the more complex and information-dense documents.

Based on this, we propose an improved variant called pre-instruction-tuning++, which trains exclusively on QA pairs to understand patterns of knowledge access, then progresses to training on a combination of QA and document data to align knowledge access through questions and knowledge encoding from documents (Fig. 4 ⑧). As shown in Tab. 2, PIT++ significantly outperforms PIT (Fig. 4 ⑦) from 45.4% to 48.1%, while training on QA data after on the mix (PIT-- in Tab. 2) does not yield additional benefits. This reinforces our hypothesis that understanding how knowledge is accessed aids in absorbing knowledge from documents, and therefore, should be prioritized.

5.3 Ablation Studies

Standard instruction-tuning is inferior not due to forgetting A drawback of standard instruction-tuning is that knowledge in test documents might be forgotten after training on QA pairs (a phenomenon also known as the "alignment tax" (Ouyang et al., 2022)). To show that the lower performance of standard instruction-tuning is not due to forgetting, we add a setting where we mix train QA with test documents during instruction-tuning to prevent forgetting (Fig. 4 ③). As shown in Tab. 2, this does not help, confirming our hypothesis.

Pre-instruction-tuning is not simply upweighting salient tokens from documents We include an ablation inspired by Hu et al. (2023) which upweights tokens when pre-training on documents to focus on salient information. We assign a weight of 1.0 to tokens in documents that are included in the answers (e.g., "Jennifer Lame" in the sentence "Editing was handled by Jennifer Lame"), and assign a lower weight of 0.5 to other tokens. As shown in Tab. 2, this weighted continued pre-training is ineffective, confirming our hypothesis.

5.4 Cross-domain Generalization

We validated the effectiveness of PIT by training and evaluation on data from the same domain (`Wiki2023-film`). *Can PIT make LLMs better at absorbing knowledge from documents of a different domain?* To this end, we follow the cross-domain setting outlined in Fig. 2—training on other domains (`Wiki2023-other-train`) and testing on the film domain (`Wiki2023-film-test`). The results of standard instruction-tuning and PIT, in both in-domain and cross-domain settings, are detailed in Tab. 3. Even though it is not as effective as the in-domain counterparts, cross-domain PIT still significantly outperforms instruction-tuning, demonstrating that it can generalize across different domains. This finding sheds light on the potential to scale this method up to a broader range of documents and instructions for more robust generalization.

We also evaluate the effectiveness of PIT in two other scenarios: (1) when applied to non-Wikipedia documents, and (2) when addressing questions asked by real users. For the first scenario, we take the Llama-2 7B model trained with PIT on `2023Wiki-other` and further train it on biographies synthesized in Zhu and Li (2023a) (`bioS`). Then, we evaluate based on questions about the individuals. For the second scenario, we manually search Google using questions generated by LLMs from `Wiki2023-film-test`, collect a total of 93 similar questions from real users by leveraging Google's "People Also Ask" feature, and then evaluate Llama-2 7B on these questions. As shown in Tab. 4, PIT outperforms baselines in both scenarios, demonstrating its generalization ability.

6 Related Work

6.1 Continual Knowledge Acquisition

Several works have studied whether LMs can answer questions about information in documents they have been trained on. Wang et al. (2021); Jang et al. (2022); Hu et al. (2023) use relatively small LMs such as BART (Lewis et al., 2020a), T5 (Raffel et al., 2020), or GPT-2 (Radford et al., 2019). Ovadia et al. (2023) focus on the comparison between RAG and continued pre-training approaches without using instruction-tuning. Zhu and Li (2023a,b) examine this problem from a similar angle as ours using a GPT-2-like transformer trained from scratch on synthetic biographies and fine-tuned on QA pairs related to the individuals. They examined a mixed training setting on both biographies and QA pairs, which is our major motivation to study different strategies to incorporate QA data before continued pre-training. Other works study adapting LLMs to new domains via various strategies (Zhang et al., 2023; Cheng et al., 2023; Han et al., 2023; Wu et al., 2023; Nguyen et al., 2023; Zhao et al., 2023).

6.2 Instruction-tuning or Alignment

Instruction-tuning (also known as supervised fine-tuning) on high-quality annotated data (Sanh et al., 2022; Wei et al., 2022; Mishra et al., 2022; Iyer et al., 2022; Kopf et al., 2023; Zhou et al., 2023; Sun et al., 2023b,a) and/or data generated by proprietary models (Taori et al., 2023; Chiang et al., 2023; Wang et al., 2023b; Ivison et al., 2023), or alignment with reinforcement learning from human feedback (RLHF) or direct preference optimization (DPO) (Ouyang et al., 2022; Touvron et al., 2023b; Rafailov et al., 2023; Tian et al., 2023) has been a central topic recently because it elicits knowledge from LLMs and enhances various abilities to handle questions from users. We focus on factuality and study the best way to perform instruction-tuning to elicit factual knowledge from LLMs.

6.3 Analyzing the Training Dynamics of LMs

Many works study the training dynamics of LMs from different perspectives. Carlini et al. (2022) quantifies memorization across model sizes and the frequency of data duplication. Tirumala et al. (2022) finds that larger LMs memorize training data faster with less overfitting. Xia et al. (2023) shows that perplexity is more predictive of model behaviors than other factors. Dery et al. (2022) studies end-task aware pre-training using classification tasks and RoBERTa models. Jia et al. (2022) adds a pre-training objective to encourage the vector for each phrase to have high similarity with the vectors for all questions it answers. Our work differs in that we specifically focus on the capacity of recalling and generalizing information from a seen document to answer questions.

6.4 Retrieval-augmented Generation

Retrieval-augmented generation (RAG) is a widely used approach to incorporate new knowledge into LLMs by augmenting fixed LLMs with retrieved information from external sources (Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020b; Borgeaud et al., 2022; Wang et al., 2023a; Alon et al., 2022; He et al., 2021; Sachan et al., 2021; Izacard et al., 2023; Lee et al., 2022; Jiang et al., 2022; Shi et al., 2023; Jiang et al., 2023; Asai et al., 2023; Nakano et al., 2021; Qin et al., 2023; Lin et al., 2023). While RAG is effective in reducing hallucinations commonly experienced when relying solely on knowledge stored in parameters, its retrieval and generation process adds extra latency and complexity. In contrast, continued pre-training to store knowledge in parameters and utilizing the stored knowledge to answer questions in a closed-book manner are simpler and faster at inference time. Enhancing this capability is also scientifically significant, as it represents a fundamental step in employing LLMs as dependable assistants for accessing information. Therefore, this paper focuses on exploring parametric approaches.

7 Conclusion

We study the best way of continued training on new documents with the goal of later eliciting factual knowledge and propose pre-instruction-tuning that learns how knowledge is accessed via QA pairs prior to encoding knowledge from documents. Extensive experiments and ablation studies demonstrate the superiority of pre-instruction-tuning versus standard instruction-tuning. Future directions include scaling this method up to a broader range of documents and instructions for more robust generalization.

8 Limitations

The `Wiki2023` dataset provides a relatively clean testbed for studying continual knowledge acquisition. However, its scope is limited to Wikipedia, which restricts the trained models' adaptability to other sources like web pages from Common Crawl or scientific documents from arXiv. We focus on eliciting factual knowledge with instruction-tuning on QA data in this paper. The effectiveness of pre-instruction-tuning with different types of data for enhancing other skills like reasoning or comprehension is something that needs to be explored in future studies.

Acknowledgements

We would like to thank Zeyuan Allen-Zhu, Zexuan Zhong, Shuyan Zhou, Frank F. Xu, Qian Liu, and Ruohong Zhang for their help with the experiments and constructive feedback.

References

Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International Conference on Machine Learning*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *CoRR*, abs/2309.12288.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *CoRR*, abs/2202.07646.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *CoRR*, abs/2309.09530.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Lucio M. Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. 2022. Should we be pre-training? an argument for end-task aware training as an alternative. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca - an open-source collection of medical conversational AI models and training data. *CoRR*, abs/2304.08247.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Conference on Empirical Methods in Natural Language Processing*.

Nathan Hu, Eric Mitchell, Christopher D. Manning, and Chelsea Finn. 2023. Meta-learning online adaptation of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4418–4432. Association for Computational Linguistics.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing LM adaptation with tulu 2. *CoRR*, abs/2311.10702.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Robin Jia, Mike Lewis, and Luke Zettlemoyer. 2022. Question answering infused pre-training of general-purpose contextualized representations. In *ACL (Findings)*, pages 711–728. Association for Computational Linguistics.

Zhengbao Jiang, Luyu Gao, Zhiruo Wang, Jun Araki, Haibo Ding, Jamie Callan, and Graham Neubig. 2022. Retrieval as attention: End-to-end learning of retrieval and reading within a single transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2336–2349. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7969–7992. Association for Computational Linguistics.

Andreas Kopf, Yannic Kilcher, Dimitri von Rutte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Rich'ard Nagyfi, ES Shahul, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. *ArXiv*, abs/2304.07327.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher D. Manning, and Kyoung-Gu Woo. 2022. You only need one model for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3047–3060. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. 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. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2023. RA-DIT: retrieval-augmented dual instruction tuning. *CoRR*, abs/2310.01352.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciuca, Charlie O'Neill, Ze-Chang Sun, Maja Jablonska, Sandor Kruk, Ernest Perkowski, Jack W. Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Rózanski, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill P. Naiman, Jesse Cranney, Kevin Schawinski, and UniverseTBD. 2023. Astrollama: Towards specialized foundation models in astronomy. *CoRR*, abs/2309.06126.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *CoRR*, abs/2312.05934.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Webcpm: Interactive web search for chinese long-form question answering. *CoRR*, abs/2305.06849.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *CoRR*, abs/2305.18290.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 25968–25981.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023a. SALMON: self-alignment with principle-following reward models. *CoRR*, abs/2310.05910.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023b. Principle-driven self-alignment of language models from scratch with minimal human supervision. *CoRR*, abs/2305.03047.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *CoRR*, abs/2311.08401.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar, and Bryan Catanzaro. 2023a. Shall we pretrain autoregressive language models with retrieval? A comprehensive study. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7763–7786. Association for Computational Linguistics.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can generative pre-trained language models serve as knowledge bases for closed-book qa? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3241–3251. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *CoRR*, abs/2306.04751.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13711–13738. Association for Computational Linguistics.

Ruohong Zhang, Luyu Gao, Chen Zheng, Zhen Fan, Guokun Lai, Zheng Zhang, Fangzhou Ai, Yiming Yang, and Hongxia Yang. 2023. A self-enhancement approach for domain-specific chatbot training via knowledge mining and digest. *CoRR*, abs/2311.10614.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *CoRR*, abs/2305.11206.

Zeyuan Allen Zhu and Yuanzhi Li. 2023a. Physics of language models: Part 3.1, knowledge storage and extraction. *CoRR*, abs/2309.14316.

Zeyuan Allen Zhu and Yuanzhi Li. 2023b. Physics of language models: Part 3.2, knowledge manipulation. *CoRR*, abs/2309.14402.

A `Wiki2023` Dataset

Prompt 1: question-answer generation prompt

Given the following summary about the subject {topic}, generate a comprehensive list of questions and corresponding answers that cover all aspects. To make the question clear, always include {topic} in the question. Answers should be concise, consisting of a few short phrases separated by commas.

Output in the following format:
Q: an open-domain question about the subject {topic} (the subject {topic} should always be included)
A: phrase1, phrase2, ...

Summary:
{summary}

B Hyperparameters

We use AdamW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.1. We decay the learning rate to 10% of its initial value using a cosine scheduler without warm-up. When pre-training on documents, we use a batch size of 256 documents and an initial learning rate of 3e-5. During instruction-tuning on QA pairs, we use the same batch size of 256 QA pairs, but opt for a reduced initial learning rate of 5e-6 because the number of tokens in a single batch used for computing losses is lower. The number of epochs varies depending on the setting and is detailed in the corresponding sections.

C Evaluation Metrics

At inference time, we use greedy decoding to generate answers given questions as context, following the format in Fig. 3. To evaluate the original Llama-2, we add 5 QA pairs as in-context exemplars to make sure it follows the QA format. Since most questions are simple factoid questions and most answers are relatively short, we use exact match (EM) as our primary metric (Kwiatkowski et al., 2019), which measures whether the model's output matches the gold answer exactly after normalization (e.g., remove articles and punctuations). To assess longer responses and accommodate minor lexical differences, we also report answer recall, which

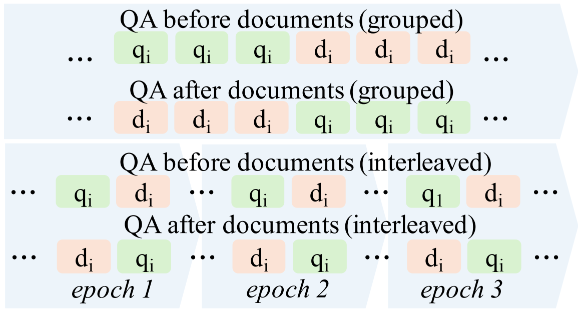


Figure 6: Different arrangements between QA pairs and corresponding documents. The ellipses represent other examples.

assesses if the gold answer appears in the model's output, and ROUGE-L, which measures the longest common subsequence between the model's output and the gold answer.

D Details of Ablation Studies

We arrange the order of QA pairs and corresponding documents as shown in Fig. 6 to study the learning mechanism of pre-instruction-tuning.