SELF-TUNING: Instructing LLMs to Effectively Acquire New Knowledge through Self-Teaching

Anonymous ACL submission

Abstract

Large language models (LLMs) often struggle to provide up-to-date information due to their one-time training and the constantly evolving nature of the world. To keep LLMs current, existing approaches typically involve continued pre-training on new documents. However, they frequently face difficulties in extracting stored knowledge. Motivated by the remarkable success of the Feynman Technique in efficient human learning, we introduce SELF-TUNING, a learning framework aimed at improving an LLM's ability to effectively acquire new knowledge from raw documents through self-teaching. Specifically, we develop a SELF-TEACHING strategy that augments the documents with a set of knowledge-intensive tasks created in a self-supervised manner, focusing on three crucial aspects: *memorization*, *comprehension*, and *self-reflection*. Additionally, we introduce three Wiki-Newpages-2023-QA datasets to facilitate an in-depth analysis of an LLM's knowledge acquisition ability concerning *memorization*, *extraction*, and *reasoning*. Extensive experimental results on LLAMA2 family models reveal that SELF-TUNING consistently exhibits superior performance across all knowledge acquisition tasks and excels in preserving previous knowledge.

1 Introduction

Armed with a wealth of factual knowledge acquired during the pre-training phase (Zhou et al., 2023a), LLMs (Touvron et al., 2023; OpenAI, 2023) exhibit remarkable proficiency in numerous knowledge-intensive tasks (Cohen et al., 2023; Gekhman et al., 2024). Despite this, the knowledge stored in LLMs can quickly become outdated due to the one-time training of LLMs and the ever-changing nature of the world (Huang et al., 2023; Jiang et al., 2024b). These unavoidable knowledge limitations present notable obstacles to the trustworthiness of LLMs in real-world scenarios (Liu et al., 2023; Mecklenburg et al., 2024). Thus, it is essential to equip LLMs with new knowledge to keep them up-to-date.

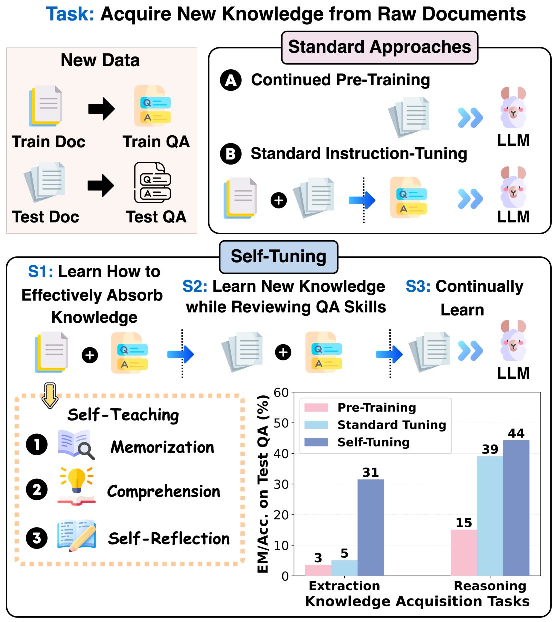


Figure 1: Illustration of the knowledge acquisition task with two standard knowledge injection approaches (in the upper part). Depiction of SELF-TUNING for effective knowledge acquisition from raw documents, which significantly enhances factual accuracy compared to the standard approaches (in the lower part).

In this paper, we focus on injecting new knowledge into the parameters of LLMs. As depicted in the upper part of Figure 1, a standard approach involves continued pre-training (A) on a raw corpus (here, test doc) containing new information (Jang et al., 2022). However, it struggles to extract the embedded knowledge, potentially due to the impaired question-answering (QA) capability (Allen-Zhu and Li, 2023; Cheng et al., 2024). Despite the assistance of subsequent instruction-tuning (B) (Wei et al., 2022; Ouyang et al., 2022a) on QA data, the knowledge retrieved from the LLMs remains notably constrained (Jiang et al., 2024b). Recently, Jiang et al. (2024b) suggests fine-tuning on a mix of QA data and related documents before continuing pre-training, with the aim of teaching the model how to access knowledge from documents and answer questions. Although this method greatly outperforms standard approaches, our initial results suggest that its effectiveness in knowledge extraction remains limited.

Numerous studies (Ambion et al., 2020; Reyes et al., 2021) evidence the effectiveness of the Feynman Technique (Xiaofei et al., 2017) in promoting human learning and knowledge understanding. The remarkable success of this potent learning method is often attributed to its emphasis on "comprehension," "self-reflection" ("identifying gaps and review"), rather than mere "memorization". This encourages our exploration into its potential application in improving LLMs' knowledge acquisition capabilities. As a result, we present SELF-TUNING, a framework that empowers an LLM to effectively internalize and recall new knowledge. As depicted in the lower part of Figure 1, SELF-TUNING consists of three stages: (*i*) Firstly, we train the model using a mix of training documents and associated QA data, equipping it with the ability to efficiently absorb knowledge from raw documents via self-teaching, as well as question-answering skills. Specifically, we design a SELF-TEACHING strategy to present the training documents as plain texts for *memorization* and a series of knowledge-intensive tasks derived from the documents in a self-supervised manner, without any mining patterns (van de Kar et al., 2022), for *comprehension* and *self-reflection*. (*ii*) Next, we deploy the model to apply the learning strategy for spontaneously acquiring knowledge from new documents while reviewing its QA skills. (*iii*) Finally, we continue training the model using only the new documents to ensure thorough acquisition of new knowledge.

In addition, we introduce three Wiki-Newpages-2023-QA datasets to conduct an in-depth study of how an LLM acquires new knowledge *w.r.t., memorization*, *extraction*, and *comprehension* (in this study, *reasoning*) across single-domain, multi-domain, and cross-domain settings. These datasets are carefully curated to ensure minimal overlap with the LLM's pre-training corpora, emphasizing two key knowledge-intensive tasks, *i.e.,* open-ended generation and natural language inference (NLI) tasks. Extensive experimental results on LLAMA2 family models demonstrate that SELF-TUNING significantly outperforms all other compared methods on knowledge memorization and extraction tasks. In addition, SELF-TUNING consistently yields high accuracy on reasoning tasks, while the performance of the compared methods largely fluctuates in different scenarios. Inspiringly, SELF-TUNING exhibits exceptional performance in retaining previously acquired knowledge (*i.e.,* knowledge retention) concerning extraction and reasoning on two well-established benchmarks.

In summary, our contributions are three-fold:

- We present SELF-TUNING, a framework designed to improve an LLM's knowledge acquisition capability via self-teaching.
- We introduce three Wiki-Newpages-2023-QA datasets to enable a comprehensive analysis of an LLM's knowledge acquisition ability *w.r.t.,* memorization, extraction, and reasoning.
- We validate the efficacy of SELF-TUNING on three crucial knowledge acquisition tasks using the Wiki-Newpages-2023-QA datasets.

2 Related Work

Continual Knowledge Injection. The primary research approach for injecting new knowledge into LLMs (Xu et al., 2023; Ovadia et al., 2024; Mecklenburg et al., 2024) is through continued pre-training. This method entails the ongoing pre-training of LLMs on raw corpora containing new knowledge, carried out in a causal auto-regressive manner (Allen-Zhu and Li, 2023; Ibrahim et al., 2024; Ovadia et al., 2024). However, this straightforward approach often encounters hurdles in effectively enabling LLMs to extract the acquired knowledge during the inference phase (Allen-Zhu and Li, 2023; Jiang et al., 2024b; Cheng et al., 2024). To enhance knowledge extraction, instruction tuning on QA data after pre-training has been extensively employed (Wei et al., 2022; Ouyang et al., 2022b). Jiang et al. (2024b) suggests that the effectiveness of this method remains limited, and proposes fine-tuning the model on QA data before continued pre-training. This instructs the model on how to retrieve knowledge from raw corpora, thereby enhancing knowledge extraction. However, such an approach tends to underestimate the importance of comprehending the new knowledge.

Acknowledging the value of knowledge comprehension, Cheng et al. (2024) proposes converting raw corpora into reading comprehension texts. This approach, however, focuses on domain adaptation and preserving general prompting abilities by mining a set of instruction-following tasks from

Wiki-Newpages	Factual Knowledge	Open-Ended Generation (Train & Test Sets)		NLI (Test Set)	
		Statistics	Avg. # Tokens	Statistics	Answer Type
Wiki-Bio (Single-domain)	Birth Date, Profession, Education, *etc.*	Train: 6,136 (# QA); 1,136 (# Docs) Test: 663 (# QA); 127 (# Docs)	8.34 (Q) 4.24 (A) 59.64 (Doc)	729 (# QA) 127 (# Docs)	Yes (65.84%) No (33.47%) Impossible (0.69%)
Wiki-Multi (Multi-domain)	News, TV series, Sports, *etc.*	Train: 10,004 (# QA); 1,823 (# Docs) Test: 1,502 (# QA); 281 (# Docs)	10.13 (Q) 5.70 (A) 69.25 (Doc)	1,627 (# QA) 281 (# Docs)	Yes (60.97%) No (36.63%) Impossible (2.40%)
Wiki-Film (Single-domain)	Genre, Language, Director, Released Time, *etc.*	Test: 955 (# QA); 169 (# Docs)	8.83 (Q) 4.61 (A) 58.10 (Doc)	1,387 (# QA) 169 (# Docs)	Yes (62.73%) No (26.53%) Impossible (2.52%)

Table 1: Statistical information of three Wiki-Newpages-2023-QA datasets, *i.e.,* Wiki-Bio, Wiki-Multi, and Wiki-Film. "Impossible": "It's impossible to say". Details about token count distribution can be found in Appendix K.

the document content. In contrast, our work aims to equip the model with the ability to effectively absorb new knowledge from raw documents and employ the learned ability to unseen documents. Specifically, we develop a SELF-TEACHING strategy to present the raw document as plain texts for memorization, accompanied by a set of tasks for comprehension and self-reflection, which are created based on raw corpora in a self-supervised manner, without relying on any mining patterns.

Additionally, knowledge editing (Zhang et al., 2024a) and retrieval-augmented generation (Ovadia et al., 2024; Jeong et al., 2024) are recognized as two related research fields. Further details are provided in Appendix A.

3 Wiki-Newpages-2023-QA: Datasets for Studying LLM Knowledge Acquisition

To explore the knowledge acquisition capabilities of LLMs from new documents, *w.r.t.,* memorization, extraction and reasoning, we introduce the Wiki-Newpages-2023-QA datasets (Table 1), which are carefully designed to minimize overlap with the initial pre-training corpus. These datasets comprise new document corpora for studying knowledge memorization and associated QA datasets for two vital knowledge-intensive tasks: open-ended generation and NLI for examining extraction and reasoning, respectively. Due to space constraints, we provide a brief overview of the dataset construction process here, with the complete version available in Appendix B.

3.1 Document Collection and QA Pair Generation

Document Collection. To construct the document corpus, we collect articles from September to October 2023 (4,257 articles in total) from Wikipedia NewPages[1], which include new articles from various domains published after the pre-training cut-off time of the LLMs being evaluated.[2] Following Jiang et al. (2024b), we only use the first paragraph of each article, as it offers a comprehensive summary and contains a wealth of factual information.

QA Pair Generation. We gather QA pairs for generation and NLI tasks using our handcrafted prompts in Tables 17 and 18, aiming to cover all factual information within the given document.

3.2 Splitting

To facilitate an in-depth analysis across single-domain, multi-domain, and cross-domain scenarios, we create three datasets and partition them into training and testing subsets.

Dataset Splitting. We generate three datasets: Wiki-Newpages-2023-10-Bio (Wiki-Bio), Wiki-Newpages-2023-10-Multi (Wiki-Multi), and Wiki-Newpages-2023-(9)10-Film (Wiki-Film) by randomly selecting 1,263 biographical documents, 2,104 multi-domain documents, and 955 film documents from the collected document corpus and their associated QA pairs.

Train-test Splitting. We divide Wiki-Bio and Wiki-Multi datasets into training and testing subsets for single-domain and multi-domain evaluations. We use Wiki-Film as the test set for cross-domain scenarios. Note that the training QA datasets only include open-ended generation task pairs, ensuring fair comparisons.

4 SELF-TUNING

In this section, we introduce the SELF-TUNING framework to improve the LLM's capability to ac-

[1] `https://en.wikipedia.org/wiki/Special:NewPages`

[2] The pre-training cut-off time for the LLAMA2 family models used in this study is 2022.

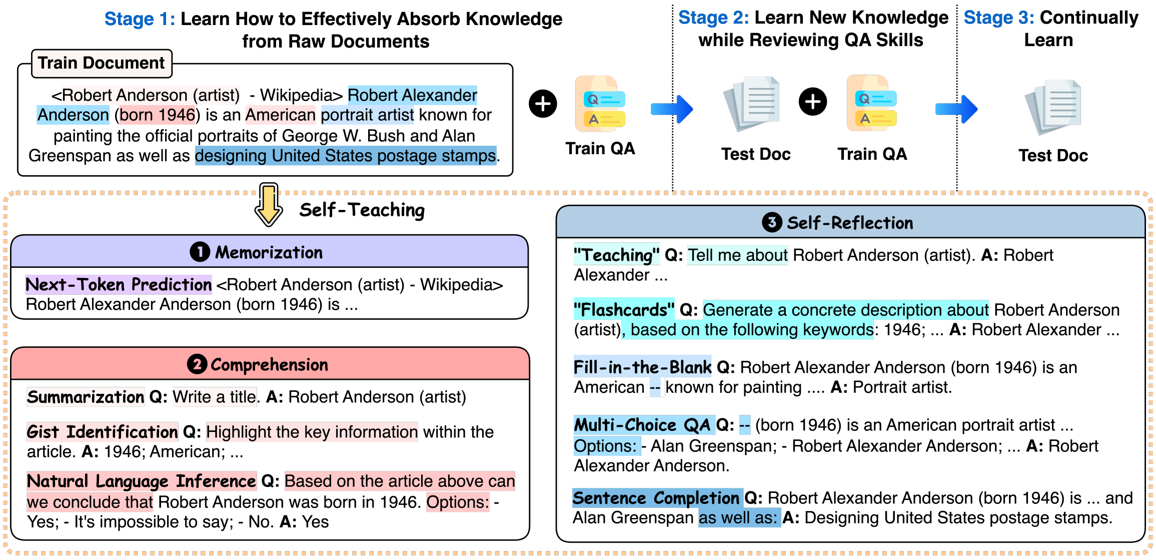


Figure 2: Illustration of the proposed SELF-TUNING. The framework consists of three stages (in the upper part): (*i*) Equipping the model with the ability to deeply absorb knowledge from raw documents using the proposed SELF-TEACHING strategy (in the lower part), along with question-answering capabilities; (*ii*) Applying the learning strategy acquired in Stage 1 to obtain new knowledge from unseen documents and refining QA skills; (*ii*) Continuously learning from unseen documents. See Appendix M for the full training document example in Stage 1.

quire knowledge from new documents, with the devised SELF-TEACHING strategy. We first give an overview of the training process for knowledge acquisition using the proposed SELF-TUNING in Section 4.1. Then, we delve into the SELF-TEACHING strategy in Section 4.2.

4.1 Overview

As depicted in Figure 2, the proposed SELF-TUNING comprises the following three stages.

Stage 1: Learn How to Effectively Absorb Knowledge from Raw Documents. Our objective is to equip an LLM M, parameterized by θ, with the ability to learn how to derive knowledge from raw documents. This is achieved by training the model using a combination of training document dataset D^{Doc}_{train} and associated training QA dataset D^{QA}_{train}, as depicted in the upper left part of Figure 2. To enhance effective knowledge absorption, we present D^{Doc}_{train} along with a series of knowledge-intensive tasks (*a.k.a.* self-teaching tasks) D^{Self}_{train} that are related to their content for SELF-TEACHING (in the lower part of Figure 2). These tasks are generated in a self-supervised manner based on the contents of D^{Doc}_{train}, using the proposed SELF-TEACHING learning approach (Section 4.2). The multi-task training objective is:

$$L^{Stage1}_{\theta} = L_{\theta}(D^{Doc}_{train}) + L_{\theta}(D^{Self}_{train}) + L_{\theta}(D^{QA}_{train}) \quad (1)$$

Stage 2: Learn New Knowledge while Reviewing QA Skills. Our aim is to train the model M to apply the learned strategy for spontaneously extracting new knowledge from unseen documents (*i.e.*, the test document dataset D^{Doc}_{test}). In addition to training on D^{Doc}_{test}, we include D^{QA}_{train}, allowing the model M to review and refine its question-answering ability. This approach enhances knowledge extraction from D^{Doc}_{test}. The objective is:

$$L^{Stage2}_{\theta} = L_{\theta}(D^{Doc}_{test}) + L_{\theta}(D^{QA}_{train}) \quad (2)$$

Stage 3: Continually Learn. Our goal is to ensure that the model M thoroughly absorbs the new knowledge by conducting follow-up training on D^{Doc}_{test}. The objective for this stage is as follows:

$$L^{Stage3}_{\theta} = L_{\theta}(D^{Doc}_{test}) \quad (3)$$

4.2 SELF-TEACHING Learning Strategy

Motivated by the Feynman Technique, we aim to equip the model with systematic knowledge learning abilities from three perspectives: memorization, comprehension, and self-reflection, as shown in the lower part of Figure 2. Specifically, we devise a self-supervised SELF-TEACHING learning strategy that presents the raw documents D^{Doc}_{train} as plain texts for memorization and as a series of knowledge-intensive tasks in a question-answering format related to their content for comprehension and self-reflection (Table 15). This method *does not require any specific mining patterns, making it applicable to any raw texts.*

Memorization. To allow the model M to learn to memorize and capitalize on the factual information embedded in the raw texts, we execute the ***next-token prediction*** task on plain document texts.

Comprehension. Our goal is to facilitate the model's ability to comprehend the factual knowledge within the document in a top-down manner. To achieve this, we conduct the following tasks:

(*i*) ***Summarization*** allows the model to learn to grasp the topic by using the prompt `Write a title:` to encourage the model to summarize the raw text, with the document title serving as the ground truth.

(*ii*) ***Gist identification*** improves the model's ability to pinpoint the key elements (*i.e.,* entities) within the atomic facts. Specifically, we prompt the model with `Highlight the key information within the article:`, and use the entities within the document as gold answers, identified using Spacy[3].

(*iii*) ***Natural language inference*** provides the model with the capability to determine whether a statement can be inferred from specific document contents (*i.e.,* "Yes," "No," or "It's impossible to say"), thus avoiding misconceptions that may arise during knowledge acquisition. Specifically, we use a randomly sampled sentence (identified using NLTK[4]) within the document content as the true statement, and a corrupted version where one entity is replaced by an irrelevant entity from another sentence as the false statement. Then, we prompt the model with `Based on the article above can we conclude that` and the sampled sentence (either initial or corrupted), with the three relations as options and corresponding answers.

Self-Reflection. Our objective is to improve the model's ability to memorize and recall acquired knowledge by "identifying and filling in the knowledge gaps." To this end, we devise the following closed-book generation tasks:

(*i*) ***"Teaching"*** fosters the model's ability to recall its acquired knowledge on a particular topic by "pretending to teach" others, using the prompt `Tell me about {topic}:` with the document content serving as the answer.

(*ii*) ***"Flashcards"*** imparts the model with the ability to recall its learned information based on the topic and associated keywords, using the prompt `Generate a concrete description about {topic} based on the following keywords:`, with the document text as the answer.

(*iii*) ***Fill-in-the-Blank*** equips the model with the ability to conduct a detailed check on the acquired factual information. Specifically, we randomly replace one entity with a "–" symbol to form a cloze question, with the replaced entity serving as the corresponding answer.

(*iv*) ***Multi-choice QA*** helps the model learn to differentiate the correct answer from the available options and prevents confusion with irrelevant content. Specifically, we randomly replace one entity with a "–" symbol to form a cloze question, with the replaced entity and three other entities randomly sampled from the document forming the options, and the replaced entity serving as the correct choice.

(*v*) ***Sentence completion*** allows the model to develop its ability to focus on factual data found towards the end of a sentence. This is crucial since our initial observations indicate that the model frequently encounters difficulties when attempting to extract knowledge from later positions. Additionally, the model is anticipated to learn to emphasize not only entities but also phrase-level factual information. To achieve this, we first employ Spacy to pinpoint prepositions in a randomly chosen sentence from the document. Then, we store the phrase that follows the final preposition as the correct answer and the portion of the sentence preceding the phrase as the question. Comprehensive templates for each task can be found in Table 15.

5 Experiments

5.1 Setup

Datasets and Evaluation Metrics. We validate SELF-TUNING in both knowledge acquisition and retention for a well-rounded analysis.

We carry out assessments on three **knowledge acquisition** tasks. (*i*) *For memorization*, we utilize test document datasets and report perplexity (PPL) (Jelinek et al., 1977). (*ii*) *For extraction*, we employ test QA datasets for open-ended generation tasks and evaluate factual accuracy using exact match (EM), Recall, F1 (Kwiatkowski et al., 2019), and Rouge-L (Lin, 2004; Jiang et al., 2024b). We also assess accuracy using the bidirectional entailment approach with the Deberta-Large-MNLI model (He et al., 2021). (*iii*) *For reasoning*, we use test QA datasets for NLI tasks and report accuracy.

We conduct evaluations on two aspects of **knowledge retention**. (*i*) *For extraction*, we evaluate the

[3] https://spacy.io/usage

[4] A natural language toolkit. https://www.nltk.org/

model's performance in retaining factual knowledge using Natural Questions (NQ) (Kwiatkowski et al., 2019) (*i.e.,* NQ-open (Min et al., 2021)) and report EM and F1 scores. (*ii*) *For reasoning*, we assess the capability in retaining commonsense knowledge using CommonsenseQA (CSQA) (Talmor et al., 2019) and report accuracy.

All evaluations are conducted in a closed-book setting. Details can be found in Appendix N.

Compared Methods. We compare our method with the following representative approaches, as presented in the upper part of Table 5 and report the mean results of three different runs.

- **Continued Pre-training** trains the model on the D^{Doc}_{test} dataset.
- **Standard Instruction-tuning** first trains on both D^{Doc}_{train} and D^{Doc}_{test} datasets, then fine-tunes on D^{QA}_{train} dataset.
- **PIT** (Jiang et al., 2024b) first trains on D^{QA}_{train} and D^{Doc}_{train} datasets to equip the model with the ability to absorb knowledge from raw documents, with the QA pairs placed right before the corresponding document texts, then trains on the D^{Doc}_{test} data.

Due to space limits, we present the comprehensive **implementation details** in Appendix O.

5.2 Main Results

Table 2 (top) presents the evaluation results on LLAMA2-7B in relation to knowledge acquisition and retention in the single-domain scenario using the Wiki-Bio dataset. Due to space limitations, the results on LLAMA2-13B can be found in Appendix D. The following observations are noteworthy:

The curated dataset exhibits minimal overlap with the pre-training data of the LLMs. The extremely low performance in the closed-book setting (*e.g.,* with EM around 2% for knowledge extraction) indicates that the dataset has little in common with the pre-training data, thus ensuring the reliability of the evaluation results. The non-zero EM values might be due to a small number of collected Wikipedia articles that were initially published but underwent revisions after the cut-off time.

SELF-TUNING substantially improves the LLM's knowledge acquisition ability. SELF-TUNING greatly enhances the performance of LLAMA2-7B across three dimensions: (*i*) reducing PPL to nearly 1, signifying effective memorization of the new documents; (*ii*) increasing EM by roughly 20% on the knowledge extraction task, attaining performance comparable to the open-book setting; (*iii*) achieving high accuracy among the compared methods for the reasoning task, demonstrating excellent understanding of the newly acquired knowledge. These results underscore the value of comprehension and self-reflection, beyond simply memorizing document contents. This confirms the effectiveness of the SELF-TEACHING learning approach. We provide in-depth analyses in Appendix F, Appendix G, and Appendix H.

SELF-TUNING excels in knowledge retention. Unlike other methods that display fluctuating performance, SELF-TUNING shows a strong ability to maintain previously acquired knowledge in terms of both knowledge extraction and reasoning. The slight improvements in evaluation metrics, such as F1 (roughly 1% on extracting learned world knowledge) and accuracy (around 13% on commonsense reasoning), compared to the closed-book performance without knowledge injection, suggest that systematically learning new knowledge doesn't necessarily lead to catastrophic forgetting. Instead, it enhances the elicitation and understanding of previously learned knowledge.

In addition, we further validate the efficacy of SELF-TUNING by comparing it with three other representative methods (Appendix C) and present evaluation results on LLAMA2-7B-CHAT (Appendix E) to promote a comprehensive understanding of the performance across various models.

5.3 Results in the Multi-Domain and Cross-Domain Scenarios

To explore the potential of SELF-TUNING for enhancing LLM's knowledge acquisition and retention in real-world scenarios, we evaluate its performance in two challenging settings (Table 2): (*i*) the multi-domain scenario (in the middle part), where both the training documents and test documents come from various complex domains; (*ii*) the cross-domain scenario (in the bottom part), where training data and test documents belong to entirely different domains, *i.e.,* the training data is from Wiki-Bio, while the test data is from Wiki-Film.

SELF-TUNING consistently shows strong potential in enhancing knowledge acquisition and retention in both settings. In Table 2, SELF-TUNING consistently achieves the best performance in both settings, suggesting the potential to expand this method to a wider range of documents containing diverse new knowledge.

Method	Wiki-Newpages-2023-QA (Acquisition)							NQ (Reten.)		CSQA (Reten.)
	Memorization	Extraction					Reason.	Extraction		Reasoning
	PPL (↓)	% Acc.	% EM	% F1	% Rec.	% Rouge	% Acc.	% EM	% F1	% Acc.
Knowledge Acquisition on Wiki-Newpages-2023-10-Bio (Single-Domain Scenario)										
w/o Knowledge Injection										
Open-book w/ test doc	8.41	55.20	31.83	64.48	75.55	62.10	7.96	-	-	-
Closed-book	8.41	4.68	2.87	14.63	16.98	15.07	7.96	16.05	24.67	53.40
w/ Knowledge Injection										
Con. Pre-training	7.28	6.33	3.62	15.96	18.72	16.11	15.09	16.00	24.11	53.40
Standard Ins.-tuning	6.83	6.94	5.13	19.15	19.05	19.48	39.09	15.72	23.67	51.84
PIT	2.08	14.03	11.61	27.15	28.86	27.11	11.93	15.72	**26.31**	57.58
SELF-TUNING	**1.11**	**37.25**	**31.52**	**50.83**	**52.62**	**50.61**	**44.31**	**16.45**	25.67	**66.01**
Knowledge Acquisition on Wiki-Newpages-2023-10-Multi (Multi-Domain Scenario)										
w/o Knowledge Injection										
Open-book w/ test doc	7.84	48.93	26.63	60.37	71.71	58.54	6.33	-	-	-
Closed-book	7.84	4.53	2.73	16.19	18.63	16.38	6.33	16.05	24.67	53.40
w/ Knowledge Injection										
Cont. Pre-training	3.32	5.86	3.40	18.04	20.59	18.42	14.51	**17.02**	25.05	53.56
Standard Ins.-tuning	2.73	8.66	5.73	24.94	25.64	25.31	34.91	15.60	26.26	52.74
PIT	1.96	14.31	8.72	30.26	33.97	30.22	10.69	15.55	**27.02**	55.12
SELF-TUNING	**1.13**	**22.30**	**16.51**	**39.94**	**41.02**	**39.89**	**50.65**	16.34	25.85	**69.29**
Knowledge Acquisition on Wiki-Newpages-2023-(9)10-Film (Cross-Domain Scenario)										
w/o Knowledge Injection										
Open-book w/ film doc	8.30	57.38	34.45	68.64	78.92	66.31	7.35	-	-	-
Closed-book	8.30	3.35	1.88	11.27	12.97	11.49	7.35	16.05	24.67	53.40
w/ Knowledge Injection										
Cont. Pre-training	5.52	3.46	2.30	11.83	14.30	11.98	12.04	**16.79**	25.35	56.02
Standard Ins.-tuning	2.83	5.23	3.77	16.15	17.45	16.45	**51.69**	14.41	25.54	49.80
PIT	1.52	6.39	2.67	16.97	18.92	17.10	3.03	13.06	23.42	54.38
SELF-TUNING	**1.10**	**22.51**	**16.44**	**35.58**	**36.60**	**35.43**	44.92	16.77	**26.44**	**66.34**

Table 2: Five-shot evaluation results on LLAMA2-7B for knowledge acquisition and retention in three scenarios: single-domain (top), multi-domain (middle), and cross-domain (bottom). Results that fall below the baseline performance are highlighted in red.

The capacity to systematically absorb knowledge improves generalization ability. The substantial improvements over all compared methods in the cross-domain setting, *e.g.,* exceeding EM by 13% on the knowledge extraction task, highlight the value of equipping the model with the ability to effectively absorb knowledge from raw documents using the SELF-TEACHING strategy, rather than solely teaching it how to answer questions.

5.4 Training Dynamics

We analyze the training dynamics of SELF-TUNING during continued pre-training (beginning from Stage 2 in Figure 2) on the test documents by varying the number of training epochs for two main reasons: (*i*) to eliminate the possibility that the exceptional performance of SELF-TUNING in enhancing knowledge acquisition is merely a result of early fitting on the test documents, and (*ii*) to conduct an in-depth assessment of its long-term knowledge retention capability. Furthermore,

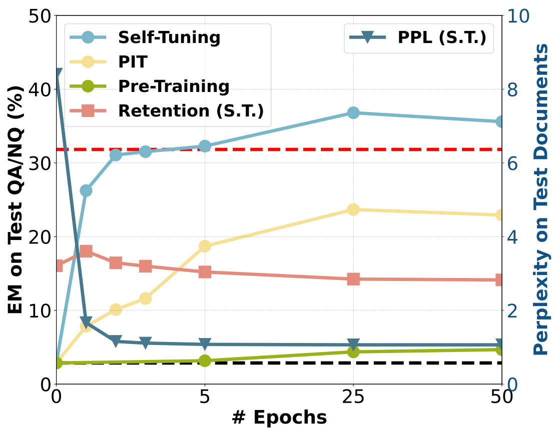


Figure 3: Training dynamics on LLAMA2-7B *w.r.t.,* knowledge memorization, extraction, and retention across different numbers of training epochs. We present the EM scores on NQ datasets to evaluate knowledge retention. The black and red dashed lines represent the baseline closed-book and open-book performances for the knowledge extraction task, respectively.

Method	Wiki-Newpages-2023-10-Bio (Acquisition)							NQ (Reten.)		CSQA (Reten.)
	Mem.	Extraction					Reason.	Extraction		Reasoning
	PPL (↓)	% Acc.	% EM	% F1	% Rec.	% Rouge	% Acc.	% EM	% F1	% Acc.
Continued Pre-training	7.28	4.68	2.87	14.63	16.98	15.07	7.96	16.05	24.67	53.40
SELF-TUNING w/o Review	1.26	28.36	23.68	41.29	41.93	41.11	**50.40**	15.55	24.20	65.11
SELF-TUNING via Read.	1.46	20.97	17.65	34.54	39.19	34.55	39.37	**18.43**	**27.99**	62.74
SELF-TUNING w/ Pre-Review	1.28	29.86	25.94	43.46	44.96	43.31	46.91	16.28	24.80	65.11
SELF-TUNING	**1.11**	**37.25**	**31.52**	**50.83**	**52.62**	**50.61**	44.31	16.45	25.67	**66.01**

Table 3: Five-shot evaluation results of the SELF-TUNING variants on LLAMA2-7B in the single-domain scenario. Results that fall below the baseline closed-book performance (previously shown in Table 2) are highlighted in red.

we integrate the results of PIT and continued pre-training to offer a well-rounded evaluation.

The remarkable performance of SELF-TUNING in enhancing knowledge acquisition does not stem from early-fitting. In Figure 3, we observe that SELF-TUNING not only memorizes new knowledge more rapidly than the compared methods, lowering PPL to almost 1 within 3 epochs, but also consistently achieves the best performance during long-term training. Remarkably, SELF-TUNING begins to outperform the open-book performance from the 5th epoch and reaches its peak at the 25th epoch with a 5% higher EM score on the knowledge extraction task. This observation highlights the importance and potential of incorporating knowledge into the parameters of LLMs.

SELF-TUNING performs well in preserving previously acquired knowledge, with only a small decline in EM of roughly 2-3% over the course of 50 training epochs. This suggests that SELF-TUNING has great potential for real-world applications.

5.5 Variants of SELF-TUNING

Setup. To further investigate the effectiveness of SELF-TUNING, we present three variations, as depicted in Table 5: (1) SELF-TUNING w/o Review, where we continue training on test documents without the reviewing capability; (2) SELF-TUNING via Read., which displays the training documents in a reading-comprehension format (Cheng et al., 2024) (an example is shown in Table 21); (3) SELF-TUNING w/ Pre-Review, which trains on a combination of training documents and training QA in the second stage, before training on test documents.

Results. In Table 3, despite having lower performance than SELF-TUNING, all variations significantly enhance the model's ability for knowledge acquisition compared to continued pre-training, which further validates the effectiveness of SELF-TUNING in improving knowledge acquisition.

Reviewing the QA ability aids in knowledge acquisition and retention. Compared to SELF-TUNING, SELF-TUNING w/o Review also displays inferior performance on the knowledge retention task. Moreover, we suspect that the slightly lower performance of SELF-TUNING w/ Pre-Review is due to reviewing the QA ability during the continuous learning of new knowledge helps in reducing the distribution shift, thereby stabilizing the training process. These findings underscore the importance of reviewing the QA ability during the continuous knowledge acquisition.

Decoupling the knowledge acquisition process into three perspectives is more effective than solely focusing on comprehension. The comparison between SELF-TUNING w/o Review and SELF-TUNING w/ Read. demonstrates that presenting the test document text from three distinct perspectives contributes more to knowledge memorization (1.26% vs. 1.46% on PPL), extraction (23.68% vs. 17.65% on EM), and reasoning (50.40% vs. 39.37% on accuracy) than presenting the test document text with all constructed tasks as a whole.

6 Conclusion

In this study, we introduce SELF-TUNING to enhance an LLM's ability to effectively learn from raw documents through self-teaching. Specifically, we develop SELF-TEACHING, a self-supervised learning strategy that presents documents as plain texts along with various knowledge-intensive tasks derived directly from the documents. Additionally, we present three Wikipedia-Newpages-2023-QA datasets to enable a comprehensive evaluation of an LLM's knowledge acquisition capabilities across three distinct scenarios. Our findings show that SELF-TUNING consistently yields superior performance on the knowledge acquisition tasks while showing impressive knowledge retention performance. These results suggest the potential for broader applications of SELF-TUNING, such as acquiring domain-specific knowledge.

Limitations

While our experimental results show promise, we consider these findings to be preliminary, as there are still many unexplored aspects in this field.

Conducting Experiments on Various LLMs. Due to constraints in time and computational resources, our extensive experiments are conducted on LLAMA2 family models. For future research, we plan to explore the effectiveness of our approach in enhancing the knowledge acquisition capabilities of other models such as Mistral-7B (Jiang et al., 2023), Orca2-7B (Mitra et al., 2023), LLAMA3 models (MetaAI, 2024), and larger-scale models like LLAMA2-70B.

Applying to Broader Scenarios. Our study primarily centers on infusing new factual knowledge into the parameters of LLMs to keep them up-to-date. However, our proposed SELF-TEACHING strategy doesn't necessitate any mining patterns to build the knowledge-intensive tasks based on the new corpus. As a result, we foresee that our SELF-TUNING framework can be utilized in various areas, such as arming the LLMs with domain-specific knowledge (Cheng et al., 2024; Que et al., 2024), and mathematical principles (Xu et al., 2024).

Performing More Comprehensive Evaluations of LLMs' Knowledge Acquisition Capabilities. In this study, we evaluate the knowledge acquisition capabilities of LLMs from three important perspectives: knowledge memorization, extraction, and reasoning. Future work could consider additional evaluation aspects, such as integrating factual knowledge with mathematical reasoning, to explore the model's ability to utilize the learned factual knowledge in solving more complex real-world problems (Zheng et al., 2024).

Ethics Statement

Throughout the research, we have consistently adhered to ethical guidelines. During the Wiki-Newpages-2023-QA data collection process using GPT-4, we carefully constructed prompts to eliminate any language that might discriminate against specific individuals or groups. These measures aimed to minimize potential negative effects on users' well-being. Examples of these thoughtfully designed prompts can be found in Table 17, Table 18, and Table 20. To further ensure dataset quality, the authors manually reviewed the newly collected datasets, following the instructions in Bai et al. (2022). These datasets were confirmed to be of high quality, devoid of offensive content, false information, or any personally identifiable information (Zhou et al., 2023b; Radharapu et al., 2023). Additionally, future research efforts could explore the OpenAI moderation API[5] to systematically filter out inappropriate system responses. The knowledge retention tasks utilized well-established benchmark datasets. Our study is dedicated to advancing knowledge while maintaining a strong commitment to privacy, fairness, and the well-being of all individuals and groups involved.

[5] `https://platform.openai.com/docs/guides/moderation/overview`

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *Preprint*, arXiv:2309.14316.

Ronnel Ian A Ambion, Rainier Santi C De Leon, Alfonso Pio Angelo R Mendoza, and Reinier M Navarro. 2020. The utilization of the feynman technique in paired team teaching towards enhancing grade 10 anhs students' academic achievement in science. In *2020 IEEE Integrated STEM Education Conference (ISEC)*, pages 1–3. IEEE.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1856–1869, Dubrovnik, Croatia. Association for Computational Linguistics.

ContextualAI. 2024. Introducing rag 2.0.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *Preprint*, arXiv:2405.05904.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Preprint*, arXiv:2403.08763.

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 *International Conference on Learning Representations*.

Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and Janet M. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *Preprint*, arXiv:2403.14403.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, Qun Liu, and Wei Wang. 2024a. Learning to edit: Aligning llms with knowledge editing. *Preprint*, arXiv:2402.11905.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen tau Yih, and Srinivasan Iyer. 2024b. Instruction-tuned language models are better knowledge learners. *Preprint*, arXiv:2402.12847.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick 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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *Preprint*, arXiv:2308.05374.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *Preprint*, arXiv:2404.00213.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date. *MetaAI blog*.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen tau Yih. 2021. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. *Preprint*, arXiv:2101.00133.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *Preprint*, arXiv:2311.11045.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Long Ouyang, Jeffrey 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. 2022b. Training language models to follow instructions with human feedback. 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*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in llms. *Preprint*, arXiv:2312.05934.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, Xu Tan, Jie Fu, Wenbo Su, Jiamang Wang, Lin Qu, and Bo Zheng. 2024. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Preprint*, arXiv:2406.01375.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 380–395, Singapore. Association for Computational Linguistics.

Englevert Reyes, Ron Blanco, Defanee Doroon, Jay Limana, and Ana Torcende. 2021. Feynman technique as a heutagogical learning strategy for independent and remote learning. *Recoletos Multidisciplinary Research Journal*, 9:1–13.

Kuniaki Saito, Kihyuk Sohn, Chen-Yu Lee, and Yoshitaka Ushiku. 2024. Where is the answer? investigating positional bias in language model knowledge extraction. *Preprint*, arXiv:2402.12170.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Mozes van de Kar, Mengzhou Xia, Danqi Chen, and Mikel Artetxe. 2022. Don't prompt, search! mining-based zero-shot learning with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7508–7520, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation. *Preprint*, arXiv:2310.03214.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Jason Wei, Maarten Bosma, Vincent 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 *International Conference on Learning Representations*.

Kevin Wu, Eric Wu, and James Zou. 2024a. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *Preprint*, arXiv:2404.10198.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024b. How easily do irrelevant inputs skew the responses of large language models? *Preprint*, arXiv:2404.03302.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *Preprint*, arXiv:2405.15556.

Wang Xiaofei, Chen Qing, Sun Yanyan, Tong Weifeng, and Niu Wenzhi. 2017. The application of the feynman technique for practical teaching of prosthodontics. *Chinese Journal of Medical Education*, 41(9):822.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tür. 2023. KILM: knowledge injection into encoder-decoder language models. *CoRR*, abs/2302.09170.

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, Jie Tang, and Yuxiao Dong. 2024. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *Preprint*, arXiv:2404.02893.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024a. A comprehensive study of knowledge editing for large language models. *Preprint*, arXiv:2401.01286.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024b. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4862–4876. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*.

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. 2023a. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023b. Rethinking machine ethics – can llms perform moral reasoning through the lens of moral theories? *Preprint*, arXiv:2308.15399.

A Additional Efforts for Knowledge Injection

Knowledge editing (Zheng et al., 2023; Yao et al., 2023; Jiang et al., 2024a; Zhang et al., 2024a) and retrieval-augmented generation (RAG) (Lewis et al., 2021; Ovadia et al., 2024; Jeong et al., 2024) are recognized as two related research initiatives in the field of knowledge injection.

(*i*) *Knowledge editing* (Mitchell et al., 2022; Zheng et al., 2023; Yao et al., 2023; Jiang et al., 2024a; Zhang et al., 2024a) concentrates on rectifying outdated or inaccurate factual knowledge stored in the model, without affecting other facts. In contrast, our focus lies in enabling LLMs to efficiently acquire knowledge from raw corpora.

(*ii*) *Retrieval-augmented generation (RAG)* (Lewis et al., 2021; Vu et al., 2023; Ovadia et al., 2024; Jeong et al., 2024) equips LLMs with new knowledge by augmenting off-the-shelf LLMs with retrieved knowledge from external sources. However, its performance is vulnerable to irrelevant or malicious information in the retrieval results (ContextualAI, 2024), potentially leading to inaccurate responses (Zhang et al., 2024b; Wu et al., 2024b; Xiang et al., 2024). Moreover, recent findings (Wu et al., 2024a) emphasize an underlying tension between a model's prior knowledge and the information presented in retrieved documents. Consequently, this paper primarily focuses on exploring the injection of knowledge into the parameters of LLMs.

B Wiki-Newpages-2023-QA: Datasets for Studying LLM Knowledge Acquisition

To explore the knowledge acquisition capabilities of LLMs from new documents, *w.r.t.,* memorization, extraction and reasoning, we introduce the Wiki-Newpages-2023-QA datasets, which are carefully designed to minimize overlap with the initial pre-training corpus. These datasets comprise new document corpora for studying knowledge memorization and associated QA datasets for two vital knowledge-intensive tasks: open-ended generation and NLI for examining extraction and reasoning, respectively. We provide the details on dataset construction in the following subsections.

B.1 Document Collection

Given the well-structured nature of Wikipedia articles, which encompass extensive factual information and cover a wide range of topics across various domains, we gather documents from Wikipedia NewPages[6]. This collection includes new articles from diverse domains published after the pre-training cut-off time of the LLMs being evaluated,

Document: <Sawyer Gipson-Long - Wikipedia> Alec Sawyer Gipson-Long (born December 12, 1997) is an American professional baseball pitcher for ...

QA Pair Example for Generation Task

Question: When was Sawyer Gipson-Long born?
Answer: December 12, 1997.

QA Pair Example for NLI Task

Question: Based on the paragraph above can we conclude that <Alec Sawyer Gipson-Long> Sawyer Gipson-Long was born in December 1997.
Options: -Yes; -It's impossible to say; -No
Answer: Yes

Table 4: A simplified example of a document and its associated QA pair for the open-ended generation task. Factual information related to the QA pairs is denoted in blue.

allowing us to largely ensure that the models have not been exposed to these facts. To construct the document corpus, we specifically gather articles from September to October 2023, resulting in a total of 4,257 articles.[7] Following Jiang et al. (2024b), we only utilize the first paragraph of each article, which provides a comprehensive summary and contains a wealth of factual information.

B.2 QA Pair Generation

To gather QA pairs, we utilize GPT-4 (OpenAI, 2023) along with our manually curated prompts to generate a variety of questions and their corresponding answers, aiming to cover all factual information within the given document. Specifically, we construct QA datasets for the open-ended generation and NLI tasks by employing the prompts shown in Table 17 and Table 18, respectively. A simplified example document with associated QA pairs is provided in Table 4. More detailed examples can be found in Appendix J.

B.3 Splitting

To enable a comprehensive analysis in single-domain, multi-domain, and cross-domain situations, we develop three datasets and divide them into training and testing subsets.

Dataset Splitting. We create three datasets: Wiki-Newpages-2023-10-Bio (Wiki-Bio), Wiki-Newpages-2023-10-Multi (Wiki-Multi), and Wiki-Newpages-2023-(9)10-Film (Wiki-Film). Specifically, we randomly select 1,263 biographical doc-

[6] https://en.wikipedia.org/wiki/Special:NewPages

[7] The pre-training cut-off for the LLAMA2 family models used in this study is 2022.

uments to curate Wiki-Bio, choose 2,104 documents covering various topics for constructing Wiki-Multi, and compile 955 film documents for producing Wiki-Film, using the assembled document corpus along with their associated QA pairs.

Train-test Splitting. We partition the Wiki-Bio and Wiki-Multi datasets, comprising the document corpus and the derived QA datasets, into training and testing subsets for conducting evaluations in single-domain and multi-domain contexts. We directly utilize the Wiki-Film dataset as the test set for the cross-domain scenario. It is crucial to note that the training QA datasets only contain the QA pairs from open-ended generation tasks, ensuring a fair comparison with existing knowledge injection approaches. We provide extensive statistical information for the three datasets in Table 1 and a thorough analysis of the QA types in Appendix L.

C Evaluation Results of Additional Compared Methods

For a thorough assessment, we also examine the efficiency of our suggested SELF-TUNING by contrasting it with three other notable methods, namely standard instruction-tuning without forgetting, PIT^{++}, and mixed training, as displayed in Table 5. We present the evaluation results in Table 6. Our SELF-TUNING consistently shows superior performance, for example, it increases the EM by 11% on the knowledge extraction task.

D Evaluation Results on LLAMA2-13B in the Single-domain Scenario

Table 7 presents the evaluation results on LLAMA2-13B concerning knowledge acquisition and retention in the single-domain scenario using the Wiki-Bio dataset. We make the following observations:
SELF-TUNING consistently demonstrates superior performance in enhancing the model's knowledge acquisition and retention abilities as the model size scales. As the model size scales, SELF-TUNING continues to achieve the highest performance across all evaluation metrics on memorization and acquisition tasks, consistently outperforming the compared methods by a significant margin (*e.g.,* improving EM score by 20% on the extraction task). On the reasoning task, SELF-TUNING consistently attains high accuracy. Additionally, SELF-TUNING consistently exhibits strong performance on knowledge retention tasks. These findings confirm the effectiveness of SELF-TUNING, suggesting the potential and robustness of SELF-TUNING for applications on larger-scale models.

Continued pre-training for knowledge acquisition proves challenging across all three dimensions. We find that continuing pre-training on new documents may result in a decline in knowledge extraction performance on LLAMA2-13B, compared to the baseline performance. This could be due to the fact that merely continuing pre-training might adversely affect its question-answering capability, even when equipped with new knowledge, as demonstrated by the lowered PPL. This observation is consistent with the findings in Cheng et al. (2024). Moreover, the marginal improvements in memorization (reducing PPL by 2%) and reasoning (increasing accuracy by 2%) suggest that such a naive approach fails to help the model memorize and capitalize on new knowledge. This highlights the importance of evaluating the model's knowledge acquisition ability comprehensively across multiple dimensions.

E Evaluation Results on LLAMA2-7B-CHAT in the Single-domain Scenario

In this section, we showcase the evaluation outcomes for LLAMA2-7B-CHAT in Table 8. We find that even after extensive instruction-following training (Ouyang et al., 2022a), LLAMA2-7B-CHAT faces difficulty in extracting newly acquired knowledge after simply continuing pre-training on test documents. Almost all high-performing approaches struggle with knowledge retention, indicating that to incorporate new knowledge, it is preferable to train a base model rather than the version fine-tuned via RLHF (reinforcement learning from human feedback) (Ouyang et al., 2022a), despite its remarkable instruction-following capability. More significantly, SELF-TUNING consistently surpasses all other compared methods by a considerable margin on knowledge acquisition tasks. These promising outcomes further validate the effectiveness of SELF-TUNING. The results imply a potential foundation for exploring the domain of enhancing knowledge acquisition for various

[8] To ensure a fair comparison, all compared approaches train on the test documents for 3 epochs in total, regardless of the number of training stages. For continued pre-training, which is observed to struggle in grasping new knowledge, we train the models for 5 epochs.

Method	Training Data in Each Stage: Stage 1	Stage 2	Stage 3
Continued Pre-training	❶ test doc		
Standard Ins.-tuning	❶ train doc & test doc	❷ train QA	
PIT	❶ train QA train doc	❷ test doc	
SELF-TUNING	❶ train QA & train doc w/ self-teaching tasks	❷ train QA & test doc	❸ test doc
***Variants of* SELF-TUNING**			
SELF-TUNING w/o Review	❶ train QA & train doc w/ self-teaching tasks	❷ test doc	
SELF-TUNING via Read.	❶ train QA & train doc (reading-comp. format)	❷ test doc	
SELF-TUNING w/ Pre-Review	❶ train QA & train doc w/ self-teaching tasks	❷ train QA & train doc	❸ test doc
Additional Compared Methods			
Standard Ins.-Tuning w/o Forget.	❶ train doc & test doc	❷ train QA & test doc	
PIT^{++}	❶ train QA	❷ train QA train doc	❸ test doc
Mixed Training	❶ train doc & train QA & test doc		

Table 5: Depiction of the training stages and associated datasets employed in the compared methods. "Train doc w/ self-teaching tasks": the training documents presented together with the self-teaching tasks. "Reading-comp. format": reading-comprehension format. "Forget.": "Forgetting".[8]

Method	Wiki-Newpages-2023-10-Bio (Acquisition): Mem.	Extraction					Reason.	NQ (Reten.): Extraction		CSQA (Reten.): Reasoning
	PPL (↓)	% Acc.	% EM	% F1	% Rec.	% Rouge	% EM	% EM	% F1	% EM
LLAMA2-7B										
Closed-book	8.41	4.68	2.87	14.63	16.98	15.07	7.96	16.05	24.67	53.40
w/ Knowledge Injection										
Continued Pre-training	7.28	4.68	2.87	14.63	16.98	15.07	7.96	16.05	24.67	53.40
Standard Ins.-Tuning w/o Forget.	2.82	9.35	7.09	21.25	21.72	21.51	36.08	16.05	24.88	54.30
PIT^{++}	1.78	22.78	20.06	37.11	37.62	37.06	42.25	16.39	25.67	57.00
Mixed Training	1.42	24.13	20.67	38.82	39.95	38.66	**55.69**	**19.33**	**28.40**	58.97
SELF-TUNING	**1.11**	**37.25**	**31.52**	**50.83**	**52.62**	**50.61**	44.31	16.45	25.67	**66.01**

Table 6: Five-shot evaluation results of the additional compared methods in the single-domain scenario. Results that are inferior to closed-book performance without knowledge injection are indicated in red.

Method	Wiki-Newpages-2023-10-Bio (Acquisition): Memorization	Extraction					Reason.	NQ (Reten.): Extraction		CSQA (Reten.): Reasoning
	PPL (↓)	% Acc.	% EM	% F1	% Rec.	% Rouge	% Acc.	% EM	% F1	% Acc.
LLAMA2-13B										
w/o Knowledge Injection										
Open-book w/ test doc	8.27	58.97	37.41	70.38	78.64	68.09	3.57	-	-	-
Closed-book	8.27	6.33	4.68	17.45	19.37	17.58	3.57	19.84	28.71	66.34
w/ Knowledge Injection										
Con. Pre-training	6.35	4.98	3.77	17.12	18.95	17.04	5.49	**21.25**	30.35	66.34
Standard Ins.-tuning	3.00	12.67	10.11	26.79	27.42	27.00	52.43	19.95	30.95	65.77
PIT	1.70	22.93	19.61	36.50	36.99	36.25	**59.40**	19.05	31.02	70.93
SELF-TUNING	**1.09**	**44.19**	**39.37**	**58.31**	**60.47**	**57.90**	54.18	20.69	**31.62**	**71.50**

Table 7: Five-shot evaluation results on LLAMA2-13B for knowledge acquisition and retention in the single-domain scenario. Results that are inferior to closed-book performance without knowledge injection are indicated in red.

models.

Method	Wiki-Newpages-2023-10-Bio (Acquisition)							NQ (Reten.)		CSQA (Reten.)
	Memorization	Extraction					Reason.	Extraction		Reasoning
	PPL (↓)	% Acc.	% EM	% F1	% Rec.	% Rouge	% Acc.	% EM	% F1	% Acc.
LLAMA2-7B-CHAT										
w/o Knowledge Injection										
Open-book w/ test doc	12.36	71.34	43.74	75.11	88.38	73.74	31.14	-	-	-
Closed-book	12.36	5.58	4.07	16.05	17.63	16.19	31.14	18.20	26.84	67.16
w/ Knowledge Injection										
Con. Pre-training	8.12	5.73	3.32	15.89	18.60	15.81	24.83	**18.32**	**27.01**	65.19
Standard Ins.-tuning	2.99	12.67	10.56	25.13	25.41	25.38	67.76	14.81	23.72	58.07
PIT	1.85	15.54	13.12	29.03	29.47	29.45	39.51	14.92	23.38	62.33
SELF-TUNING	**1.10**	**33.03**	**29.41**	**46.94**	**47.90**	**47.00**	**72.29**	13.57	22.28	64.21

Table 8: Five-shot evaluation results on LLAMA2-7B-CHAT for knowledge acquisition and retention in the single-domain scenario. Results that are inferior to closed-book performance without knowledge injection are indicated in red.

Method	Q&A Types (% EM)			
	Total	Top-5 (37%)	Time-Related (27%)	Multiple (10%)
PIT	7.00	10.81	3.70	0
SELF-TUNING	**32.00**	**37.84**	**40.74**	**20.00**

Table 9: Fine-grained evaluation results on the open-ended generation task, using the Wiki-Bio test dataset concerning the fact types of QA pairs.

F Fine-grained Comparison

Setup. To fully understand how the ability to systematically acquire knowledge aids in the knowledge extraction task, we conduct fine-grained comparisons of PIT and SELF-TUNING on generated answers for 100 randomly sampled questions from the Wiki-bio dataset. This subset includes 56 QA types in total. Furthermore, we categorize the questions based on the fact types they contain: (*i*) the top-5 most common (accounting for 37%), which includes birthdate, affiliation, nationality, profession, and position/sport; (*ii*) time-related (accounting for 27%), such as birthdate, event date, and time period; (*iii*) multiple-facts (accounting for 10%), which ask about more than one fact, for example, inquiring both birth date and place; and we report the evaluation results separately. We assess the factual accuracy using exact match.

Results. As shown in Table 9, we observe that SELF-TUNING consistently outperforms PIT in the overall evaluation and the fine-grained evaluations related to different QA types. These findings underscore the importance of equipping the model with the ability to systematically acquire new knowledge. Furthermore, we present a qualitative comparison between the answers generated by PIT and SELF-TUNING in Appendix G. To gain insights into potential enhancements for SELF-TUNING, we also conduct a detailed error analysis on the types of factual errors that remain challenging after implementing SELF-TUNING in Appendix I.

G Qualitative Analysis

In Table 10, we provide a qualitative comparison between the answers generated by PIT and SELF-TUNING on the Wiki-Bio test set. We observe that SELF-TUNING performs better in answering questions that inquire about multiple facts and time-related facts, as indicated in the top part of Table 10. Furthermore, as shown in the lower part, PIT tend to fail to recall and extract facts at the end of the documents, *i.e.,* suffering from "positional bias". This observation is consistent with the findings in Saito et al. (2024). Encouragingly, our proposed SELF-TUNING aids in recalling and extracting factual knowledge embedded at the end of the documents. These findings align with the automatic evaluation results, underscoring the effectiveness of SELF-TUNING in enhancing the LLM's knowledge acquisition capability, particularly in knowledge extraction.

H Ablation Study

Setup. We conduct a comprehensive analysis of how comprehension and self-reflection tasks within the self-teaching tasks contribute to enhancing the LLM's knowledge acquisition ability. We focus on two vital aspects: knowledge memorization and extraction. Specifically, we calculate the percentage of the constructed examples for each task type and systematically remove certain tasks to study their

Case study 1: Questions requesting information on multiple facts.

Document: <Helmut Moritz - Wikipedia> Helmut Moritz (1 November 1933 - 21 October 2022) was an Austrian physical geodesist. He was a member of the Austrian Academy of Sciences and of many other international academies and societies. He became internationally known with a fundamental work on Error propagation in Geodesy. From 1991 to 1995, he was president of the International Union of Geodesy and Geophysics (IUGG).
Question: When was Helmut Moritz born and when did he pass away?
Gold Answer: Born on November 1, 1933, passed away on October 21, 2022.
Model Answers

PIT's Answer: Information not provided.

SELF-TUNING's Answer: Born on november 1, 1933, passed away on october 21, 2022.

Case study 2: Questions inquiring about time-related details.

Document: <Brad Smiley - Wikipedia> Brad Smiley (born June 19, 1973) is an American college football coach. He is the head football coach for Southern Arkansas University; a position he has held since 2022. He also was the head coach for Trinity Valley Community College from 2007 to 2017. He also coached for Baylor, Northwestern State, and Tulane.
Question: Since when has Brad Smiley been the head football coach for Southern Arkansas University?
Gold Answer: Since 2022.
Model Answers

PIT's Answer: Since 2016.

SELF-TUNING's Answer: Since 2022.

Case study 3: Questions inquiring about facts encoded in the end of the document, *i.e.*, "positional bias".

Document: <Nathan Saliba - Wikipedia> Nathan-Dylan Saliba (born February 7, 2004) is a Canadian professional soccer player who plays for Major League Soccer club CF Montréal.
Question: Which Major League Soccer club does Nathan Saliba play for?
Gold Answer: CF Montréal.
Model Answers

PIT's Answer: San jose earthquakes.

SELF-TUNING's Answer: CF Montréal.

Table 10: Qualitative analyses comparing the answers produced by PIT and SELF-TUNING on the open-ended generation task using the Wiki-Newpages-2023-10-Bio test dataset. The false answers and correct answers are highlighted in red and blue, respectively.

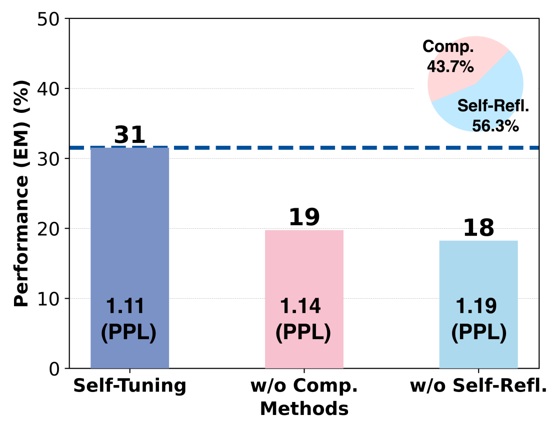


Figure 4: Ablation analysis exploring the impact of removing comprehension and self-reflection tasks from the self-teaching tasks for knowledge memorization and acquisition. The proportion of each task type among the self-teaching tasks in the training documents is shown in the upper right corner.

impacts.

Results. In Figure 4, we observe the following: (*i*) The examples of self-reflection tasks account for a slightly higher ratio than comprehension tasks among the self-teaching tasks. (*ii*) Both comprehension and self-reflection tasks benefit overall performance on the knowledge acquisition tasks. Notably, removing the examples of self-reflection tasks results in a more significant drop in performance, aligning with its higher percentage over comprehension tasks. These findings confirm the efficacy of the developed SELF-TEACHING strategy, underscoring the crucial role of comprehension and self-reflection in learning new knowledge for LLMs.

I Error Analysis

In order to gain insights into potential enhancements for SELF-TUNING, we outline four common errors that persist as challenges after implementing SELF-TUNING. We offer an in-depth analysis of these errors in Table 11, using EM as the evaluation metric.

J In-depth Sample Documents and Corresponding QA Pairs for Open-Ended Generation and Natural Language Inference Tasks

We present detailed sample documents along with their corresponding QA pairs for open-ended generation and natural language inference tasks in Table 12 and Table 13, respectively.

K Token Count Distribution for the Open-ended Generation Task Across the Three Datasets

The distribution of token counts for the open-ended generation task across the three datasets is depicted in Figure 5, Figure 6, and Figure 7, respectively.

L Examination of QA Types in Open-ended Generation QA Datasets

We perform a detailed analysis of the QA types associated with the factual information in the open-ended generation QA datasets, as displayed in Table 14, by using the prompt in Table 20 with GPT-4.

M Detailed Templates used in the SELF-TEACHING Strategy

We provide the detailed templates employed in the SELF-TEACHING strategy in Table 15 and a complete example of a training document accompanied by its associated SELF-TEACHING tasks in Table 16.

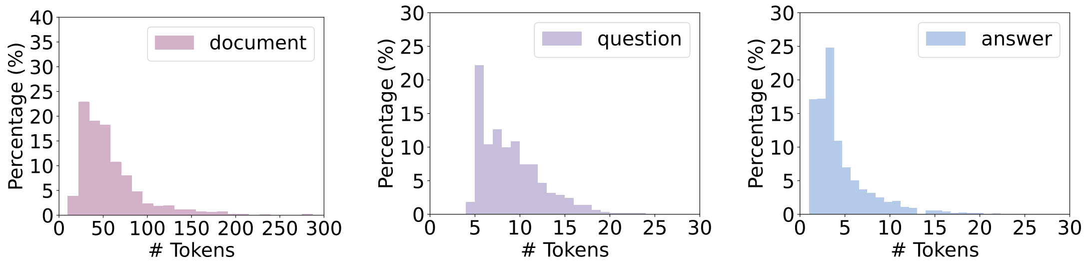
N Datasets and Evaluation Metrics

Evaluation on Knowledge Acquisition. We assess the effectiveness of SELF-TUNING in enhancing the model's knowledge acquisition capabilities on the curated Wiki-Newpages-QA datasets, concentrating on memorization, extraction, and reasoning. (*i*) For memorization, we utilize test document datasets and report perplexity (Jelinek et al., 1977), which measures how well a language model predicts a text sample. (*ii*) For extraction, we employ test QA datasets for open-ended generation tasks. To evaluate the factual accuracy of the generated responses, we use exact match (EM), Recall, and F1 over words in the answer(s), following Kwiatkowski et al. (2019). Additionally, we report Rouge-L (Lin, 2004) to measure the overlap of n-grams between the generated and gold answers, accounting for minor lexical variations, following Jiang et al. (2024b). We also assess accuracy by comparing each response's factual correctness to the gold answer, using the bidirectional entailment approach with the Deberta-Large-MNLI model (He et al., 2021). We report the five-shot evaluation results on the open-ended generation tasks using the prompt in Table 19. (*iii*) Concerning reasoning, we utilize the test QA datasets for NLI tasks and report the accuracy by comparing the generated option with the gold option using EM. We present the zero-shot evaluation results on NLI tasks.

Evaluation on Knowledge Retention. It is well-known that knowledge acquisition is often accompanied by catastrophic forgetting (Allen-Zhu and Li, 2023; Wang et al., 2023). Therefore, we also provide the knowledge retention performance for a comprehensive investigation. Specifically, (*i*) we verify the knowledge extraction performance on world knowledge using natural questions (NQ) (Kwiatkowski et al., 2019) (*i.e.,* NQ-open (Min et al., 2021) in the closed-book setting) and report EM and F1 scores. We report the five-shot evaluation results using the first five QA pairs in the dev sets as prompts. (*ii*) we assess the reasoning capability on Commonsense knowledge using CommonsenseQA (CSQA) (Talmor et al., 2019), employing accuracy to assess the correctness of the selected option, calculated by comparing the generated option against the gold option using EM. We present the five-shot performance on the dev sets, as the test set does not contain golden annotations, and use the first five multi-choice QA pairs in the training set as prompts. We use these two datasets because they were curated before the cut-off time of LLAMA2 family models (*i.e.,* year 2022), making it likely that the models have obtained relevant knowledge in these datasets during the pre-training stage, as evidenced by Touvron et al. (2023).

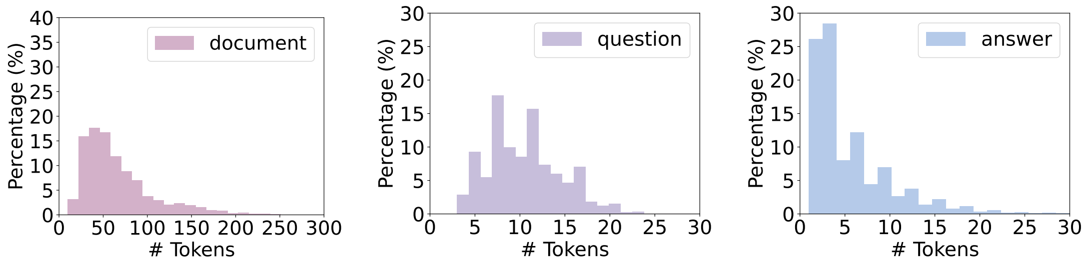
O Implementation Details

Training Details. We utilize LLAMA2-7B and LLAMA2-13B for our investigation and provide an analysis on LLAMA2-7B-CHAT in Appendix E for a comprehensive understanding. We use the following training objectives: (*i*) for training on document data D^{Doc}, we compute the standard next-token prediction loss by averaging over all tokens in the document d (Equation 4); (*ii*) for training on QA data D^{QA}, we compute the average negative log-likelihood loss only on tokens in the answer a given the question q (Equation 5), where $|d|$ and $|a|$ refer to the length of the tokenized document



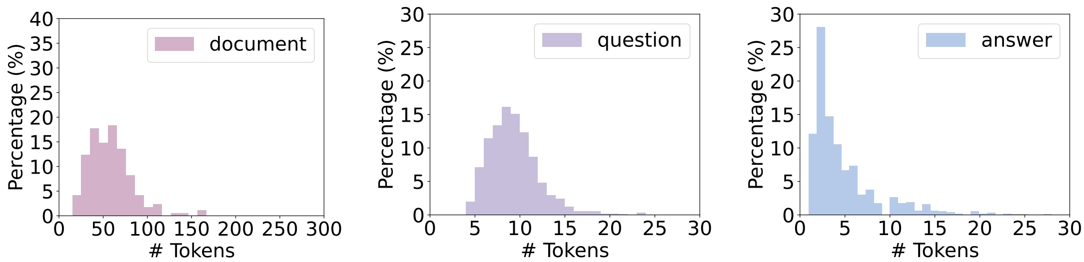
(a) Document length. (b) Question length. (c) Answer length.

Figure 5: Distribution histogram of the token count in a document, a question, and an answer for the open-ended generation task from the Wiki-Newpages-2023-10-Bio dataset, respectively.



(a) Document length. (b) Question length. (c) Answer length.

Figure 6: Distribution histogram of the token count in a document, a question, and an answer for the open-ended generation task from the Wiki-Newpages-2023-10-Multi dataset, respectively.



(a) Document length. (b) Question length. (c) Answer length.

Figure 7: Distribution histogram of the token count in a document, a question, and an answer for the open-ended generation task from the Wiki-Newpages-2023-(9)10-Film dataset, respectively.

sequence and answer sequence, respectively.

$$L_\theta(D^{Doc}) = -\frac{1}{|d|}\sum_t \log p_\theta\left(d_t \mid d_{<t}\right) \tag{4}$$

$$L_\theta(D^{QA}) = -\frac{1}{|a|}\sum_t \log p_\theta\left(a_t \mid q, a_{<t}\right) \tag{5}$$

We train LLAMA2-7B and LLAMA2-7B-CHAT on 8 32GB Tesla V100 GPUs using a batch size of 8 and a learning rate of 5e-6. Additionally, we train LLAMA2-13B on 8 A100-SXM4-40GB GPUs with a batch size of 8 and a learning rate of 5e-6. To ensure a fair comparison, all compared approaches train on the test documents for 3 epochs in total, regardless of the number of training stages. For continued pre-training, which is observed to struggle in grasping new knowledge, we train the models for 5 epochs. The specific number of training epochs used for each approach in Table 5 are as follows:

- **Continued Pre-training** trains the model on the D^{Doc}_{test} dataset for 5 epochs.
- **Standard Instruction-tuning** first trains on both D^{Doc}_{train} and D^{Doc}_{test} datasets, then fine-tunes on D^{QA}_{train} dataset for 3 epochs.

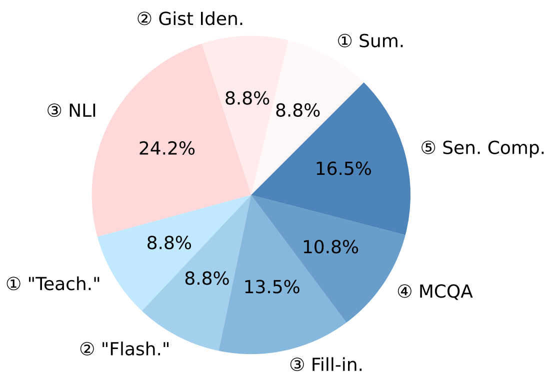


Figure 8: The percentage of constructed examples of each task type in the self-teaching tasks on training documents in Wiki-Newpages-2023-10-Bio dataset.

- **PIT** (Jiang et al., 2024b) first trains on D^{QA}_{train} and D^{Doc}_{train} datasets for 3 epochs, positioning the QA pairs right before the corresponding document texts, then trains on the D^{Doc}_{test} data for 3 epochs.
- **SELF-TUNING** (ours) first trains on D^{QA}_{train} and D^{Doc}_{train} with the created instruction-following dataset D^{Self}_{train} (in the QA format) using the SELF-TEACHING strategy for 2 epochs, then continues training on D^{Doc}_{test} data while reviewing the D^{QA}_{train} data for 1 epoch, and finally continues training on D^{Doc}_{test} data for 2 epochs. In addition, we provide the percentage of SELF-TEACHING task examples on training documents in Wiki-Newpages-2023-10-Bio dataset in Figure 8.

Specifically, in the cross-domain setting, where there is a substantial difference between the domains of the training data and test documents, we continue training on D^{Doc}_{test} data while reviewing the D^{QA}_{train} data for 2 epochs after the initial training stage, followed by further training on D^{Doc}_{test} data for 1 epoch. Furthermore, we adopt the same training strategy when dealing with LLAMA2-7B-CHAT, where the process of knowledge injection poses a significant challenge, as demonstrated by our experimental results. In accordance with Jiang et al. (2024b), for PIT and SELF-TUNING, we include 64 examples and 128 examples randomly sampled from D^{QA}_{train} datasets, respectively, during the final training stages when solely training on the D^{Doc}_{test} data, to prevent the model from losing its question-answering capabilities. It is important to note that all evaluation results are reported at the temperature $T = 1$.

Training Details for SELF-TUNING Variants.

- **SELF-TUNING w/o Review** first trains on D^{QA}_{train} and D^{Doc}_{train} with the created instruction-following dataset D^{Self}_{train} (in the QA format) using the SELF-TEACHING strategy for 2 epochs, then continues training on D^{Doc}_{test} data for 3 epochs.
- **SELF-TUNING via Read.** initially trains on D^{QA}_{train} and D^{Doc}_{train} (in the read-comprehension format, as shown in Table 21 for 3 epochs, then trains on the D^{Doc}_{test} data for 3 epochs.
- **SELF-TUNING w/ Pre-Review** first trains on D^{QA}_{train} and D^{Doc}_{train} with the created instruction-following dataset D^{Self}_{train} (in the QA format) using the SELF-TEACHING strategy for 2 epochs, then continues training on D^{Doc}_{train} and D^{QA}_{train} data for 1 epoch, and finally continues training on D^{Doc}_{test} data for 3 epochs.

Training Details for Additional Compared Methods.

- **Standard Instruction-Tuning w/o Forgetting** initially trains on the mixture of D^{Doc}_{train} and D^{Doc}_{test} for 3 epochs, then on D^{QA}_{train} and D^{Doc}_{test} datasets for 1 epoch.
- **PIT^{++}** (Jiang et al., 2024b) initially trains on D^{QA}_{train} for 1 epoch, then on D^{QA}_{train} and D^{Doc}_{train} datasets for 3 epochs, with the QA pairs placed right before the corresponding document texts, and finally, it trains on the D^{Doc}_{test} data for 3 epochs.
- **Mixed Training** trains on mixture of the D^{Doc}_{train}, D^{Doc}_{test} and D^{QA}_{train} datasets simultaneously for 3 epochs.

Prompts Employed in this Study. The prompts used for constructing the QA datasets for open-ended generation and NLI tasks are presented in Table 17 and Table 18, respectively. The prompt used during the evaluation process is displayed in Table 19. The prompt used by GPT-4 for annotating QA types in the open-ended generation tasks of the Wiki-Newpages-2023-QA datasets is presented in Table 20.

P A Sample Training Document in the Reading-Comprehension Format

Drawing inspiration from Cheng et al. (2024), we restructure the training document in the reading-comprehension text format. Each raw text is enriched with a series of tasks related to its content, constructed using our proposed SELF-TEACHING strategy. An example of a training document is provided in Table 21.

Type	Fraction	Example			
		Document	Question	Gold Answer	Model Answer
Wrong answer	76.47%	<Jalen Mack - Wikipedia> Jalen Mack (born August 5, 2005) is an American professional stock car racing driver who competes part-time in the ARCA Menards Series and ARCA Menards Series East, driving the No. 43 Chevrolet for Tamayo Cosentino Racing. He also competes part time in the ARCA Menards Series West, driving the No. 83 Chevrolet for Mack Motorsports in conjunction with Bill McAnally Racing.	Which team does Jalen Mack drive for in the ARCA Menards Series and ARCA Menards Series East?	Tamayo Cosentino Racing.	Venturini motorsports.
Higher granularity	7.35%	<Andriyko Olha Fedorivna - Wikipedia> Andriyko Olha Fedorivna (born January 28, 1945, Voronkiv, Kyiv region) is a Doctor of Law, Professor, Head of the Department of Constitutional, Administrative and Financial Law of the Kyiv University of Law of the National Academy of Sciences of Ukraine, and Deputy Head of the Department of State and Legal Problems of Management of the V. M. Koretsky Institute of State and Law of the National Academy of Sciences of Ukraine.	What are Andriyko Olha Fedorivna's academic and professional titles?	Doctor of Law, Professor.	Doctor of law, professor, head of the department of constitutional, administrative, and financial law of the kyiv university of law of the national academy of sciences of ukraine.
Lower granularity	5.88%	<Mike Babcock (American football) - Wikipedia> Michael Babcock (born February 13, 1979) is an American college football coach. He is the head football coach for McKendree University; a position he has held since 2013. He also coached for UCLA, Colorado, San Diego, and CSU Pueblo. He played college football for UCLA as a linebacker.	Since when has Mike Babcock (American football) held the head coach position at McKendree University?	Since 2013.	2013.
Paraphrase	10.29%	<Lil Tay - Wikipedia> Tay Tian (born July 29, 2009), known professionally as Lil Tay, is an American-born Canadian internet personality and singer. In 2018, she gained prominence online for a period of three months, proclaiming herself to be the "youngest flexer of the century". During her brief career, she posted rap videos on YouTube and Instagram which garnered tens of millions of views. Her career ended in mid-2018, after her father applied to the superior court of Canada for full custody and control of her career. According to court documents, he was abusive and largely an absentee.	What is Lil Tay's nationality?	American-born Canadian.	Canadian-American.

Table 11: Analysis on the types of factual errors that remain challenging after applying SELF-TUNING.

Sample document and associated QA pairs for open-ended generation tasks

Dataset: Wiki-Newpages-2023-10-Bio

Document: <Helmut Moritz - Wikipedia> Helmut Moritz (1 November 1933 - 21 October 2022) was an Austrian physical geodesist. He was a member of the Austrian Academy of Sciences and of many other international academies and societies. He became internationally known with a fundamental work on Error propagation in Geodesy. From 1991 to 1995, he was president of the International Union of Geodesy and Geophysics (IUGG).

Question: When was Helmut Moritz born and when did he pass away?
Answer: Born on November 1, 1933, passed away on October 21, 2022.
Question: What was Helmut Moritz's profession?
Answer: Austrian physical geodesist.
Question: Which academies and societies was Helmut Moritz a member of?
Answer: Austrian Academy of Sciences, many other international academies, and societies.
Question: What work made Helmut Moritz internationally known?
Answer: A fundamental work on Error propagation in Geodesy.
Question: What position did Helmut Moritz hold from 1991 to 1995?
Answer: President of the International Union of Geodesy and Geophysics (IUGG).

Dataset: Wiki-Newpages-2023-10-Multi

Document: <2018 California Proposition 71 - Wikipedia> Proposition 71, also known as Prop 71, was a California ballot proposition and proposed state constitution amendment to change the effective date of passed ballot measures from the day after the election to the fifth day after the Secretary of State certified the results.\n\n Stated goals of the measure was to ensure results were official before new measures were implemented. Opposers fearing a delay in urgent measures. Kevin Mullin supported the amendment. The California Democratic Party endorsed the amendment. Rural County Representatives of California also endorsed the amendment.

Question: What was the 2018 California Proposition 71, also known as Prop 71?
Answer: A California ballot proposition, proposed state constitution amendment, change effective date of passed ballot measures.
Question: What was the proposed change in the effective date of passed ballot measures in the 2018 California Proposition 71?
Answer: From the day after the election, to the fifth day after the Secretary of State certified the results.
Question: What were the stated goals of the 2018 California Proposition 71?
Answer: To ensure results were official before new measures were implemented.
Question: What concern did opposers of the 2018 California Proposition 71 have?
Answer: A delay in urgent measures.
Question: Who supported the 2018 California Proposition 71 amendment?
Answer: Kevin Mullin.
Question: Which organizations endorsed the 2018 California Proposition 71 amendment?
Answer: The California Democratic Party, Rural County Representatives of California.

Dataset: Wiki-Newpages-2023-(9)10-Film

Document: <Krazy House (film) - Wikipedia> Krazy House is an upcoming Dutch comedy film. It is written, directed, and co-produced by Steffen Haars and Flip van der Kuil in their English-language feature debut. Shot on location in Amsterdam, the film stars Nick Frost, Kevin Connolly and Alicia Silverstone. Maarten Swart is producer for Kaap Holland Films.

Question: What is Krazy House (film)?
Answer: An upcoming Dutch comedy film.
Question: Who are the writers, directors, and co-producers of Krazy House (film)?
Answer: Steffen Haars, Flip van der Kuil.
Question: What is significant about Steffen Haars and Flip van der Kuil's involvement in Krazy House (film)?
Answer: It is their English-language feature debut.
Question: Where was Krazy House (film) shot?
Answer: On location in Amsterdam.
Question: Who is the producer of Krazy House (film) and which production company is involved?
Answer: Maarten Swart, Kaap Holland Films.

Table 12: Sample document and associated QA pairs for open-ended generation tasks in Wiki-Newpages-2023-10-Bio, Wiki-Newpages-2023-10-Multi, and Wiki-Newpages-2023-(9)10-Film datasets.

Sample document and associated QA pairs for natural language inference tasks

Dataset: Wiki-Newpages-2023-10-Bio

Document: <Sawyer Gipson-Long - Wikipedia> Alec Sawyer Gipson-Long (born December 12, 1997) is an American professional baseball pitcher for the Detroit Tigers of Major League Baseball (MLB). He made his MLB debut in 2023.

Question: Based on the paragraph above can we conclude that <Alec Sawyer Gipson-Long> Sawyer Gipson-Long was born in December 1997. Options: -Yes; -It’s impossible to say; -No
Answer: Yes
Question: Based on the paragraph above can we conclude that <Alec Sawyer Gipson-Long> Sawyer Gipson-Long is a professional football player. Options: -Yes; -It’s impossible to say; -No
Answer: No
Question: Based on the paragraph above can we conclude that <Alec Sawyer Gipson-Long> Sawyer Gipson-Long plays for the Detroit Tigers in Major League Baseball. Options: -Yes; -It’s impossible to say; -No
Answer: Yes
Question: Based on the paragraph above can we conclude that <Alec Sawyer Gipson-Long> Sawyer Gipson-Long made his MLB debut in 2020. Options: -Yes; -It’s impossible to say; -No
Answer: No

Dataset: Wiki-Newpages-2023-10-Multi

Document: <2023 Astana Open 2̌013 Singles - Wikipedia> Novak Djokovic was the reigning champion, but chose not to compete this year.Seeds.

Question: Based on the paragraph above can we conclude that <2023 Astana Open 2̌013 Singles> Novak Djokovic won the previous Astana Open singles tournament.Options: -Yes; -It’s impossible to say; -No
Answer: Yes
Question: Based on the paragraph above can we conclude that <2023 Astana Open 2̌013 Singles> Novak Djokovic is participating in the 2023 Astana Open singles tournament.Options: -Yes; -It’s impossible to say; -No
Answer: No
Question: Based on the paragraph above can we conclude that <2023 Astana Open 2̌013 Singles> The 2023 Astana Open is a tennis tournament.Options: -Yes; -It’s impossible to say; -No
Answer: It’s impossible to say
Question: Based on the paragraph above can we conclude that <2023 Astana Open 2̌013 Singles> Novak Djokovic was injured and could not compete in the 2023 Astana Open singles tournament.Options: -Yes; -It’s impossible to say; -No
Answer: It’s impossible to say

Dataset: Wiki-Newpages-2023-(9)10-Film

Document: <Unstoppable (2023 film) - Wikipedia> Unstoppable is a 2023 comedy-drama film directed by Diamond Ratnababu and produced by Rajith Rao under AB2 Productions. The film was released theatrically worldwide on 9 June 2023.

Question: Based on the paragraph above can we conclude that<Unstoppable (2023 film)> Unstoppable is a film that combines elements of comedy and drama.Options: -Yes; -It’s impossible to say; -No
Answer: Yes
Question: Based on the paragraph above can we conclude that<Unstoppable (2023 film)> Diamond Ratnababu is the producer of the film Unstoppable.Options: -Yes; -It’s impossible to say; -No
Answer: No
Question: Based on the paragraph above can we conclude that<Unstoppable (2023 film)> Unstoppable was released in theaters worldwide.Options: -Yes; -It’s impossible to say; -No
Answer: Yes
Question: Based on the paragraph above can we conclude that<Unstoppable (2023 film)> The film Unstoppable was released before June 2023.Options: -Yes; -It’s impossible to say; -No
Answer: No
Question: Based on the paragraph above can we conclude that<Unstoppable (2023 film)> The film Unstoppable was distributed by Diamond Ratnababu.Options: -Yes; -It’s impossible to say; -No
Answer: It’s impossible to say

Table 13: Sample document and associated QA pairs for natural language inference tasks in Wiki-Newpages-2023-10-Bio, Wiki-Newpages-2023-10-Multi, and Wiki-Newpages-2023-(9)10-Film test datasets.

Dataset	QA Type Instances	QA Types		QA Types w/ Multiple Facts	
		Statistics	**Top-5 Types**	**Statistics**	**Top-5 Types**
Wiki-Newpages-2023-10-Bio (Single-domain)					
Train	Birth Date, Achievements, Position, *etc.*	2014 (# Types); 6073 (# Counts)	Birth Date (11.24%) Nationality (5.37%) Profession (5.15%) Team/Affiliation (3.05%) Role/Position (2.56%)	158 (# Types); 265 (# Counts)	Birth & Death Dates (0.93%) Birth Date & Place (0.44%) Death Date & Place (0.12%) Nationality & Profession (0.10%) Current Position & Tenure (0.08%)
Test	Full Name, Affiliation, Residence, *etc.*	281 (# Types); 655 (# Counts)	Birth Date (13.11%) Profession (6.18%) Nationality (5.62%) Team/Affiliation (4.49%) Role/Position (3.00%)	16 (# Types); 30 (# Counts)	Birth Date & Place (1.31%) Birth & Death Dates (1.12%) Death Date & Place (0.56%) Car Number & Manufacturer (0.37%) Current Club & League (0.19%)
Within the train and test sets, there are 63 and 8 answers labeled as "Information not provided/missing," respectively.					
Wiki-Newpages-2023-10-Multi (Multi-domain)					
Train	Album Source, Location, Season Number, *etc.*	4813 (# Types); 9973 (# Counts)	Birth Date (3.37%) Profession (1.76%) Nationality (1.47%) Location (1.39%) Release Date (1.27%)	303 (# Types); 371 (# Counts)	Birth & Death Dates (0.32%) Birth Date & Place (0.14%) Event Date & Location (0.06%) Death Date & Place (0.06%) Nationality & Profession (0.05%)
Test	Legacy/Impact, Purpose, Leadership, *etc.*	924 (# Types); 1498 (# Counts)	Birth Date (3.06%) Release Date (1.80%) Profession (1.57%) Nationality (1.25%) Team/Affiliation (1.02%)	57 (# Types); 66 (# Counts)	Birth & Death Dates (0.31%) Birth Date & Place (0.31%) Death Date & Place (0.16%) Job Titles & Affiliations (0.16%) Language & Genre (0.16%)
Within the train and test sets, there are 31 and 4 answers labeled as "Information not provided/missing," respectively.					
Wiki-Newpages-2023-(9)10-Film (Single-domain)					
Test	Director, Actor, Music Composer, *etc.*	339 (# Types); 955 (# Counts)	Director (9.07%) Release Date (7.23%) Genre (6.96%) Cast (3.55%) Language (2.76%)	13 (# Types); 15 (# Counts)	Title & Release Year (0.39%) Milestone & Historical Comparison (0.13%) Profession & Industry (0.13%) Cast & Roles (0.13%) Producer & Production Banner (0.13%)

Table 14: A comprehensive analysis of QA types related to factual information in open-ended generation QA datasets from Wiki-Newpages-2023-10-Bio (Wiki-Bio), Wiki-Newpages-2023-10-Multi (Wiki-Multi), and Wiki-Newpages-2023-(9)10-Film (Wiki-Film).

Type	Task	Template
Memorization		
Next-Token Prediction	Text-to-Text	<Document>
Comprehension		
① Summarization	Text-to-Topic	**Question**: Write a title: <Document>. **Answer**: <Title>.
② Gist Identification	Text-to-Word	**Question**: Highlight the key information within the article: <Document>. **Answer**: <Entity1>, <Entity2>, *etc.*
③ Natural Language Inference	Text-to-Option	**Question**: <Document> Based on the article above can we conclude that <Sentence>. Options: -Yes; -It's impossible to say; -No. **Answer**: Yes/It's impossible to say/No.
Self-Reflection		
① "Teaching"	Topic-to-Text	**Question**: Tell me about <Title>. **Answer**: <Document>.
② "Flashcards"	Word-to-Text	**Question**: Generate a concrete description about <Title>. based on the following keywords: <Entity>, *etc.* **Answer**: <Document>.
③ Fill-in-the-Blank	Cloze Sentence-to-Entity	**Question**: <Title> <Sentence_Part1> – <Sentence_Part2> (w/o <Entity>). **Answer**: <Entity>.
④ Multi-Choice QA	Cloze Sentence (w/ options)-to-Entity	**Question**: <Title> <Sentence_Part1> – <Sentence_Part2> (w/o <Entity>) Options: - <Entity1>; - <Entity2>, *etc.* **Answer**: <Entity>.
⑤ Sentence Completion	Text-to-Text	**Question**: <Title> <Sentence_Part1>: **Answer**: <Sentence_Part2>.

Table 15: The detailed templates for each task used in the SELF-TEACHING learning strategy.

Type	Example
Memorization	
Next-Token Prediction	<Robert Anderson (artist) - Wikipedia> Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps.
Comprehension	
① Summarization	**Question**: Write a title: <Robert Anderson (artist) ... stamps. **Answer**: Robert Anderson (artist).
② Gist Identification	**Question**: Highlight the key information within the article: <Robert Anderson (artist) ... stamps. **Answer**: United States; American; Alan Greenspan; George W. Bush; Robert Alexander Anderson; 1946
③ Natural Language Inference	**Question**: <Robert Anderson (artist) ... stamps. Based on the article above can we conclude that <Robert Anderson (artist)> Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Options: - Yes - It's impossible to say - No **Answer**: Yes
Self-Reflection	
① "Teaching"	**Question**: Tell me about Robert Anderson (artist). **Answer**: Robert Alexander Anderson (born 1946) is ... stamps.
② "Flashcards"	**Question**: Generate a concrete description about Robert Anderson (artist), based on the following keywords: United States; American; Alan Greenspan; George W. Bush; Robert Alexander Anderson; 1946 **Answer**: Robert Alexander Anderson (born 1946) is ... stamps.
③ Fill-in-the-Blank	**Question**: <Robert Anderson (artist)> Robert Alexander Anderson (born 1946) is an American – known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. **Answer**: Portrait artist.
④ Multi-Choice QA	**Question**:<Robert Anderson (artist)> - (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Options: - Alan Greenspan - 1946 - Robert Alexander Anderson - George W. Bush **Answer**: Robert Alexander Anderson.
⑤ Sentence Completion	**Question**: <Robert Anderson (artist)> Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as: **Answer**: Designing United States postage stamps.

Table 16: An example of a training document from the Wiki-Newpages-2023-10-Bio train set, accompanied by related self-teaching tasks.

The prompt utilized by GPT-4 for building QA datasets for open-ended generation tasks
Below is a paragraph about the 51st International Emmy Awards ceremony. Your task is to formulate a detailed list of questions and corresponding answers that encompass all the information within the paragraph. To ensure clarity, each question should explicitly mention the 51st International Emmy Awards ceremony. Answers should be concise, consisting of a few short phrases separated by commas. For instance: Paragraph: The 51st International Emmy Awards ceremony, presented by the International Academy of Television Arts and Sciences (IATAS), occurred on November 20, 2023, at the New York Hilton Midtown in New York City. It was held to acknowledge the best television programs initially produced and aired outside the United States in 2022. Nominations were announced on September 26, 2023. Question: When was the 51st International Emmy Awards ceremony held? Answer: November 20, 2023. Question: Who was responsible for presenting the 51st International Emmy Awards ceremony? Answer: The International Academy of Television Arts and Sciences (IATAS). Question: Where was the 51st International Emmy Awards ceremony held? Answer: The New York Hilton Midtown in New York City. Question: What was the purpose of the 51st International Emmy Awards ceremony? Answer: To recognize the best television programs initially produced and aired outside the United States in 2022. Question: When were the nominations for the 51st International Emmy Awards announced? Answer: September 26, 2023. Below is a paragraph about `{topic}`. Your task is to formulate a detailed list of questions and corresponding answers that encompass all the information within the paragraph. To ensure clarity, each question should explicitly mention `{topic}`. Answers should be concise, consisting of a few short phrases separated by commas. For instance: Paragraph: `{paragraph}` Question:

Table 17: The prompt utilized by GPT-4 for building QA datasets for open-ended generation tasks based on the gathered Wiki-Newpages documents.

The prompt utilized by GPT-4 for building QA datasets for natural language inference tasks
Below is a paragraph about Luis Hugo Hernán Palma Pérez. Your task is to formulate a detailed list of natural language inference tasks with questions and corresponding answers based on the paragraph. For instance: Paragraph: Luis Hugo Hernán Palma Pérez (born November 3, 1958) is a Chilean surgeon and politician, founding member of the Humanist Party of Chile. He is a deputy for the period 2022-2026, after being elected in the 2021 Chilean parliamentary elections. Question: Based on the paragraph above can we conclude that Luis Hugo Hernán Palma Pérez was born in November. Options: - Yes - It's impossible to say - No Answer: Yes Question: Based on the paragraph above can we conclude that Luis Hugo Hernán Palma Pérez is a deputy for the period 2020-2024. Options: - Yes - It's impossible to say - No Answer: No Question: Based on the paragraph above can we conclude that The Humanist Party of Chile is a political party in Chile. Options: - Yes - It's impossible to say - No Answer: Yes Question: Based on the paragraph above can we conclude that Luis Hugo Hernán Palma Pérez is a dentist. Options: - Yes - It's impossible to say - No Answer: No Question: Based on the paragraph above can we conclude that Luis Hugo Hernán Palma Pérez was elected in the 2021 Chilean parliamentary elections. Options: - Yes - It's impossible to say - No Answer: Yes Below is a paragraph about `{topic}`. Your task is to formulate a detailed list of natural language inference tasks with questions and corresponding answers based on the paragraph. For instance: Paragraph: `{paragraph}` Question:

Table 18: The prompt utilized by GPT-4 for building QA datasets for natural language inference tasks based on the gathered Wiki-Newpages documents.

The five-shot prompt used for assessing open-ended generation tasks
Question: Which animated film is included in the list of characters in the Zootopia franchise? Answer: The animated film "Zootopia" (2016). Question: Who were the coaches in The Voice Generations (Philippine TV series)? Answer: Billy Crawford, Chito Miranda, Julie Anne San Jose, and Stell of SB19. Question: Who is Cyrelle Saut? Answer: A futsal and football player who has been associated with Tuloy Foundation and the Azkals Development team. Question: What team does the 2023 Southern Miss Golden Eagles football team represent? Answer: The University of Southern Mississippi. Question: When was Kenneth Mitchell (basketball) born? Answer: October 1, 1975.

Table 19: The five-shot prompt used for assessing open-ended generation tasks, which is derived from the gathered Wiki-Newpages-2024-03 documents.

The prompt used by GPT-4 for annotating QA types in the open-ended generation tasks of the Wiki-Newpages-2023-QA datasets
Below is a paragraph along with corresponding question and answer pairs. Your task is to analyze the paragraph and the question-answer pairs by categorizing the type of information they inquire about or provide. Use concise phrases to describe each category. For example: Paragraph: <Andrew Turner (rugby union, born 2002) - Wikipedia> Andrew Turner (born 16 February 2002) is an English rugby union player, currently playing for the and . His preferred position is prop. Question: When was Andrew Turner (rugby union, born 2002) born? Answer: February 16, 2002. Question: What nationality is Andrew Turner (rugby union, born 2002)? Answer: English. Question: What sport does Andrew Turner (rugby union, born 2002) play? Answer: Rugby union. Analysis: Types of question-answer pairs: (1) Birth date, (2) Nationality, (3) Sport/Profession. Types of the paragraph: Biography - Biographical information about Andrew Turner, a rugby union player born in 2002, including his birth date, nationality, sport, and preferred position. Below is a paragraph along with corresponding question and answer pairs. Your task is to analyze the paragraph and the question-answer pairs by categorizing the type of information they inquire about or provide. Use concise phrases to describe each category. For example: Paragraph: `{paragraph}` `{QA}` Analysis:

Table 20: The prompt used by GPT-4 for annotating QA types in the open-ended generation tasks of the Wiki-Newpages-2023-QA datasets.

A training document example in the reading-comprehension format
<Robert Anderson (artist) - Wikipedia> Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Answer the questions based on the article: Question: Write a title: Answer:Robert Anderson (artist) Question: Highlight the key information within the article: Answer:United States; American; Alan Greenspan; George W. Bush; Robert Alexander Anderson; 1946 Question: Based on the article above can we conclude that <Robert Anderson (artist)> Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Options: - Yes - It's impossible to say - No Answer: Yes Question: Tell me about Robert Anderson (artist). Answer:Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Question: Generate a concrete description about Robert Anderson (artist) based on the following keywords: United States; American; Alan Greenspan; George W. Bush; Robert Alexander Anderson; 1946 Answer:Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Question: <Robert Anderson (artist)> Robert Alexander Anderson (born 1946) is an American – known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Answer: Portrait artist. Question: <Robert Anderson (artist)> - (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as designing United States postage stamps. Options: - Alan Greenspan - 1946 - Robert Alexander Anderson - George W. Bush Answer:Robert Alexander Anderson Question: <Robert Anderson (artist)> Robert Alexander Anderson (born 1946) is an American portrait artist known for painting the official portraits of George W. Bush and Alan Greenspan as well as: Answer:designing United States postage stamps

Table 21: An example of a training document from the Wiki-Newpages-2023-10-Bio train set, presented in a reading-comprehension format.