How Much Knowledge Can You Pack into a LoRA Adapter without Harming LLM?

Sergey Pletenev[1,2,†] **Maria Marina**[1,2,†] **Daniil Moskovskiy**[1,2]
Vasily Konovalov[1,3] **Pavel Braslavski**[4] **Alexander Panchenko**[2,1] **Mikhail Salnikov**[1,2]
[1]AIRI [2]Skoltech [3]Moscow Institute of Physics and Technology [4]Nazarbayev University
{S.Pletenev, Maria.Marina, A.Panchenko, Mikhail.Salnikov}@skol.tech

Abstract

The performance of Large Language Models (LLMs) on many tasks is greatly limited by the knowledge learned during pre-training and stored in the model's parameters. Low-rank adaptation (LoRA) is a popular and efficient training technique for updating or domain-specific adaptation of LLMs. In this study, we investigate how new facts can be incorporated into the LLM using LoRA without compromising the previously learned knowledge. We fine-tuned Llama-3.1-8B-instruct using LoRA with varying amounts of new knowledge. Our experiments have shown that the best results are obtained when the training data contains a mixture of known and new facts. However, this approach is still potentially harmful because the model's performance on external question-answering benchmarks declines after such fine-tuning. When the training data is biased towards certain entities, the model tends to regress to few overrepresented answers. In addition, we found that the model becomes more confident and refuses to provide an answer in only few cases. These findings highlight the potential pitfalls of LoRA-based LLM updates and underscore the importance of training data composition and tuning parameters to balance new knowledge integration and general model capabilities.

1 Introduction

Large Language Models (LLMs) have been widely adopted in many applications due to their ability to produce human-like responses to user queries. This is made possible by their ability to generalize information and accumulate a large amount of knowledge during the pre-training phase (Chen et al., 2024). These models can solve various problems, such as summarization, reasoning, and question answering, among others (Bhakthavatsalam et al., 2021; Lin et al., 2022; Hendrycks et al., 2021; Moskovskiy et al., 2024).

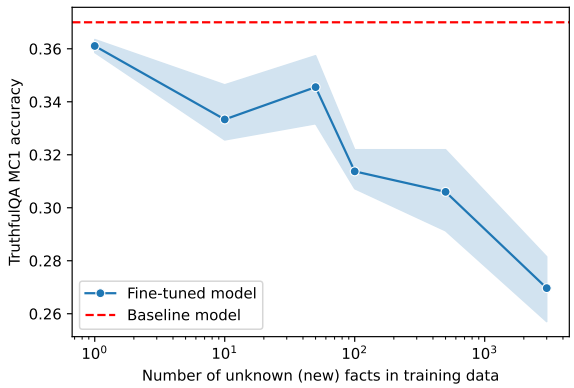

Figure 1: Decrease in quality with increase of new facts learned by the model: results of the fine-tuned Llama-3.1-8B-Instruct on TruthfulQA (solid line corresponds to the mean score, error margin – to the min/max scores of three runs with different random seeds).

Despite the general capabilities of LLMs, there are still situations that require the incorporation of new knowledge to better meet specific needs. This could be due to changes in general knowledge that occur after the LLM training period, or possibly due to specific knowledge that was intentionally omitted during the training period. To address these issues techniques such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Belikova et al., 2024) or few-shot learning (Brown et al., 2020) can be applied. In general, RAG requires access to an external knowledge base, which may be undesirable in some contexts. With respect to in-context learning, the resulting generation is strongly dependent on the selected few-shot samples (Rubin et al., 2022). In this work, we revisit fine-tuning as one of the most popular approaches for integrating new knowledge into LLMs.

Fine-tuning LLMs, which often have hundreds of billions of parameters, is a computationally expensive and time-consuming process. To address these challenges, Parameter-Efficient Fine-Tuning (PEFT) techniques have gained popular-

† These authors contributed equally to this work.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4309–4322

ity (Han et al., 2024), with Low-Rank Adaptation (LoRA) (Hu et al., 2022) being one of the most effective methods. However, these modified LLMs may suffer from drawbacks, such as catastrophic forgetting (Aleixo et al., 2024; Kirkpatrick et al., 2017) or less severe but still notable loss of previously learned associations (Hoelscher-Obermaier et al., 2023). As shown in Figure 1, an increased amount of new data during fine-tuning with LoRA can degrade the model's pre-existing world knowledge, as evidenced by declining performance of the fine-tuned Llama-3.1 model on the TruthfulQA benchmark.

We investigate the extent to which additional knowledge can be integrated into LLMs via the LoRA adapter while preserving its general capabilities. We seek to identify the underlying reasons for any performance drops when new information is introduced, and explore strategies to effectively minimize these adverse effects.

Our contributions are as follows:

- We conducted a series of extensive experiments incorporating into the LoRA model 1, 10, 50, 100, 500 and 3000 facts unknown to the model tracking how the model degrades intrinsically (via positive and negative shits) and extrinsically (by tracking the degradation of reasoning abilities on the external benchmarks, such as MMLU and TruthfulQA).
- We introduced two fine-tuning techniques to mitigate negative shifts and degradation of the model's reasoning abilities: (1) adding paraphrased new facts, and (2) adding facts the model already knows – and conducted a careful analysis of the results obtained.
- Despite the possible degradation of the model, we found positive shifts – the cases where the model learned new knowledge for which it was not trained, and explained the nature of these shifts.

We release code and data for further usage.[1]

2 Related Work

Although LLMs are highly effective in performing various natural language processing (NLP) tasks, they often struggle with tasks that require extensive real-world knowledge, particularly when dealing with long-tail facts, i.e., facts associated with less common entities (Sun et al., 2024). This limitation highlights the need for augmenting LLMs with non-parametric knowledge or integrating this knowledge into the model's parameters.

Non-parametric knowledge can significantly enhance LLM performance, particularly when the required knowledge is rare or involves multiple relationships between entities (Huang et al., 2024b). However, external knowledge sources can also potentially mislead LLMs when answering questions about well-known entities, as powerful LLMs have already internalized this information within their parameters (Mallen et al., 2023).

RAG is also not a universal solution. On the one hand, models have been shown to rely more on nonparametric knowledge (Farahani and Johansson, 2024). However, LLMs still face challenges in discriminating highly semantically related information and can easily be distracted by this irrelevant and misleading content (Wu et al., 2024). Furthermore, RAG methods introduce latency in response time, as retrieval must be performed, particularly in the case of iterative RAG approaches (Li et al., 2024). These involve a multi-step process, including retrieval followed by augmentation (Krayko et al., 2024).

Recent studies have revealed that LLMs acquire their knowledge predominantly during the pre-training phase (Allen-Zhu and Li, 2024). However, pre-training a model each time you want to incorporate new information appears to be excessive. It has been shown that attempting to acquire new knowledge through supervised fine-tuning can actually lead to an increase in hallucinations relative to the existing knowledge. Furthermore, LLMs tend to struggle when trying to integrate new knowledge via fine-tuning, instead primarily learning how to make better use of their pre-existing knowledge (Gekhman et al., 2024).

Low-Rank Adaptation, or LoRA (Hu et al., 2022) freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, thus greatly reducing the number of trainable parameters for downstream tasks.

There have been several papers on the continuous knowledge editing through various training and evaluation mechanisms. Wang et al. (2024a) have shown that almost all SoTA editing methods have a trade-off between accuracy, recall, and hallucination.

[1] https://github.com/AIRI-Institute/knowledge-packing

3 Study Design

To evaluate the ability of the LoRA adapter to incorporate new knowledge and its overall effect on the model, we define what constitutes new knowledge. We consider a knowledge fact as the combination of a question q and its corresponding answer a. The model's ability to accurately or inaccurately respond to a question determines whether it possesses or lacks this specific knowledge (q, a). We also delve into the process of fine-tuning a language model using LoRA and the methods to quantify any residual consequences of this fine-tuning procedure.

3.1 Low-Rank Adaptation

The popularity of LoRA as a method for fine-tuning LLMs resides in its time and cost effectiveness. This approach has allowed researchers and engineers to achieve results comparable to those obtained through vanilla fine-tuning on many tasks (Hu et al., 2022).

During LoRA training, each model's weight matrix $W_0 \in \mathbb{R}^{d\times k}$ is updated by the low-rank decomposition of ΔW:

$$W = W_0 + \Delta W = W_0 + BA, \qquad (1)$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, rank $r \ll \min(d, k)$. Hu et al. (2022) showed that even low-rank adaptations ($r = 1...4$) produce acceptable results on various tasks. Despite all the advantages of using Low-Rank Adaptation, LoRA-tuned LMs suffer hallucinations (Wang et al., 2024b).

3.2 Knowledge Categories

In order to fully explore the implications of fine-tuning an LLM, it is critical to recognize what knowledge it currently possesses and where its limitations lie. To determine which facts are genuinely new to a LLM, we adopted the taxonomy similar to SliCK (Sampling-based Categorisation of Knowledge) introduced by Gekhman et al. (2024).

Our approach considers that a model is knowledgeable about the answer a to a particular question q if, upon receiving q as input, it produces a as its response. However, language models can produce different responses for the same query q depending on the sampling method or prompt used. We categorize knowledge into three groups using the definition of $\mathbf{P}_{\text{correct}}(q, a, F)$ as an estimate of the probability that the language model is able to accurately generate the correct answer a to q when using different few-shot prompts F.

These knowledge types are: `HighlyKnown`, `MaybeKnown`, and `Unknown` (see Table 1). If the language model never predicts the correct answer to the question for different few-shot prompts, i.e. $\mathbf{P}_{\text{correct}}(q, a, F) = 0$, it means this fact is `Unknown` to the model. If the model generates the correct answer sometimes ($\mathbf{P}_{\text{correct}}(q, a, F) > 0$), we define this fact as `MaybeKnown`. Finally, we define a fact as `HighlyKnown` if the LLM always predicts the correct answer for all few-shot prompts, i.e. $\mathbf{P}_{\text{correct}}(q, a, F) = 1$.

3.3 Model's Reliability

Reliability (Yao et al., 2023) is the model's ability to remember both current and previous edits after sequential editing. Unlike simple accuracy or exact match, we count the occurrence of the correct answer in the model-generated response sub-string. For each question-answer pair, we have several aliases for answers.

3.4 Undesirable Effects

The primary objective of this work is to refine LoRA adapters in a way that avoids substantial degradation of the LM's performance. We employ both intrinsic and extrinsic evaluation methods to monitor the effectiveness of different LoRA-based fine-tuning configurations.

Leveraging introduced knowledge categories (Table 1), we intrinsically assess what facts the model learns (e.g., a fact shifts from the `Unknown` to `HighlyKnown` category, UK $\rightarrow$ HK) or forgets (e.g., HK $\rightarrow$ UK). No or only a few 'negative' shifts after fine-tuning mean adding new knowledge does not harm the model.

As for the extrinsic approach, we additionally evaluate all the models we train on two well-established benchmarks: MMLU (Hendrycks et al., 2021) and TruthfulQA (Lin et al., 2022). MMLU is a benchmark for knowledge and reasoning (Guo et al., 2023), is used as a proxy for measuring the model's reasoning abilities. TruthfulQA was chosen as an additional proxy for thruthfullnes, this benchmark includes the set of tricky questions that even some humans would answer falsely. We use `lm-evaluation-harness`[2] (Gao et al., 2024) for all evaluation experiments. For MMLU we use 5-shot prompting, for TruthfulQA – 0-shot prompting. For both benchmarks the final metric is accuracy

[2] https://github.com/EleutherAI/lm-evaluation-harness

Category	Definition	Explanation	# Facts
`Unknown (UK)`	$\mathbf{P}_{\text{correct}}(q, a, F) = 0$	LLM **never** returns the correct answer	14,373
`MaybeKnown (MK)`	$\mathbf{P}_{\text{correct}}(q, a, F) > 0$	LLM returns the correct answer **occasionally**	3,931
`HighlyKnown (HK)`	$\mathbf{P}_{\text{correct}}(q, a, F) = 1$	LLM **always** returns the correct answer	2,732

Table 1: **Fact categories** based on the probability of providing the correct answer to a corresponding question and number of fact (q, a) from each category.

of the answers. For TruthfulQA there are two accuracy metrics (Lin et al., 2022). In the MC1 mode, the correct answer is chosen from 4 or 5 options. This mode focuses on identifying the single truth among the choices. MC2 (multi-true) mode, on the other hand, requires the identification of multiple correct answers from a set. Both MC1 and MC2 are multiple-choice assessments.

4 Experiments

To evaluate the harmful effects of tuning the model with new knowledge, we conducted comprehensive experiments using the Llama-3.1-8B-Instruct model, which we tuned with the LoRA adapter on `Unknown` (q, a) facts.

4.1 Data

To conduct reproducible and reliable experiments, we created datasets that were not included in the pre-training set of any LLM. We are confident in this, because these data were collected in accordance with the methodology described in Sun et al. (2024), which did not provide precomputed data.

The questions in our dataset are based on Knowledge Graph (KG) entities, which are stored as triples of the form `<subject, relation, object>`. Entities are divided into three categories: head, torso, and tail, according to their popularity. This popularity is determined by the density, which is the number of relational triples that contain the entity in the KG. By including entities of varying popularity, the dataset is balanced in terms of the complexity of the questions posed to LLMs. Questions with popular entities are easier to answer, questions with torso entities are the most difficult. The KG-based structure of the dataset is attractive not only because it allows for the creation of question-answer pairs that are not seen in training data, but also because it allows analysis of shifts within the same relational domain.

We use DBpedia[3] to extract our own collection of triples and generate our own set of (q, a) pairs based on templates. In addition, we used TriviaQA (Joshi et al., 2017) as an additional source of training data to generate extra `HighlyKnown` samples. Interestingly, most of the questions in this dataset are `HighlyKnown`, even though they are complicated from human perspective. This further supports our intention to conduct experiments with data that have not been seen by any LM. Table 1 provides an analysis for each category of knowledge facts (q, a).

4.2 Fine-tuning

As a default, the model Llama-3.1-8B-Instruct[4] was chosen. It is an auto-regressive language model that uses supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

We opted for the instructed variant due to its enhanced capacity to follow instructions, and a lightweight version for streamlined implementation across a series of experimental trials.

The model is trained with 1, 10, 50, 100, 500, and $3,000$ `Unknown` (q, a) pairs. `Unknown` is the set of questions that were not answered by the default Llama-3.1-8B-Instruct model. The answer is considered correct if the answer from the triple of the question or one of its aliases is inside the LM response.

We fine-tuned the model in a zero-shot mode. The system prompt is: *"Answer the following question."* The user prompt: *"Question: "* + `question text`; the assistant prompt: *"Answer: "* + `answer text`. The data for questions and respective answers is taken from DBpedia.

Simply fine-tuning LoRA on new knowledge is challenging (Huang et al., 2024a); we augment the training dataset with synthetic data, including *paraphrases* and `HighlyKnown` facts.

When a model learns new singular knowledge as a simple sentence, it learns it without "inner structure". But if we augment it with *paraphrases*

[3] https://databus.dbpedia.org/dbpedia/mappings/mappingbased-objects

[4] https://hf.co/meta-llama/Meta-Llama-3-8B-Instruct

or HighlyKnown the model retains the new knowledge structurally, since the HighlyKnown elements model knows not as simple sentences *"Paris is a capital of a France"*, but as models' "inner space of capitals" and models' "inner space of countries". Adding new knowledge in this way is less disruptive than simply learning singular knowledge (Allen-Zhu and Li, 2024).

Paraphrasing By paraphrasing, we mean augmenting initial training data with the UK question paraphrases. For generating paraphrases we use Llama-3-70B-Instruct[5] with the system prompt: *"Please, rephrase the question 200 times differently"*. The models are trained with 0, 1, and 10 paraphrases per question. For instance, the training configuration with 10 UK + 10 paraphrases means that for each of the 10 unknown questions we take 10 paraphrases and take initial 10 questions, thus having 110 training samples.

HighlyKnown In the highly known mode, in addition to Unknown samples are added HighlyKnown samples. The sample is considered to be HighlyKnown by the default Llama-3.1-8B-Instruct, these samples are taken from DBpedia or TriviaQA.

LoRA training LoRA models were trained for 10 epochs, with learning rate $1e-3$, batch size 16, lora rank 1, lora alpha 2, lora dropout 0.1, and lora layers `down_proj, gate_proj, up_proj`.

Train-test splits We have a total of 21,036 question-answer pairs based on DBpedia triples. Using Llama-3.1-8B-instruct responses, we categorized them into – Unknown, MaybeKnown and HighlyKnown. Samples of each category are presented in Appendix C. After that we randomly take n questions, whose category is Unknown for the training part. Then, depending on the configuration, each of the n questions is augmented with k paraphrases or k HighlyKnown samples. For example, in case of 10 HK+10 UK configuration we randomly sample 10 Unknown examples out of 21,036 questions and additionally 10 HighlyKnown (out of 21,036 as well) samples for each of the Unknown questions, making 110 training samples in total. These 110 questions are part of the train. The initial 21,036 questions are the test part. Since examples for the train part are taken from the test part, we are sure that the model has indeed learned the new knowledge. Although the intersection of training and test data is not welcomed in conventional ML settings, it is crucial for our setting to check that the model has learned what it has been trained for.

Evaluation LoRA models are inferenced 10 times with 4-shot prompts. The example of a 4-shot prompt is presented in the Appendix B. Each of these 10 prompts has four distinct few shot prompts. These four few-shot examples are question-answer pairs from TriviaQA.

5 Analysis

This section includes a step-by-step analysis of the trained models. In Subsection 5.1, we analyze whether LoRA adapters can learn all the knowledge samples for which they have been trained. Subsection 5.2 provides an in-depth analysis of knowledge shifts. In Subsection 5.3 we check how reasoning abilities and truthfulness change on the external data. Finally, Subsection 5.4 sheds light on the reasons of knowledge shifts: why could the model have learned additional facts it was not trained on and forgot something it was completely sure about?

5.1 Accuracy

	Highly Known			Paraphrase		
UK	0	1	10	0	1	10
1	1.0	1.0	1.0	1.0	1.0	1.0
10	1.0	1.0	1.0	1.0	1.0	1.0
50	1.0	1.0	1.0	1.0	1.0	1.0
100	0.98	1.0	1.0	0.98	0.99	1.0
500	1.0	0.99	0.97	1.0	0.99	1.0
3,000	0.98	0.92	0.48	0.98	0.97	0.99

Table 2: **Reliability** on test for models trained on HighlyKnown and Paraphrase. Almost all UK facts are learned except for 3,000 UK trained with HK.

In this section, we consider the possibility of the model to simply learn the acquired knowledge, along with not forgetting previously known knowledge. As it can be seen from Table 2, models can learn up to 500 unknown samples with 100% reliability score. For 3,000 unknown samples 10 epochs is not enough for model to learn all samples. The same can be seen in Figure 2: with additional paraphrases for each 1 unknown sample model converges faster. But adding HighlyKnown data can be harmful to the training process. At best it has a neutral effect on the convergence, and at worst it slows down.

[5] https://hf.co/meta-llama/Meta-Llama-3-70B-Instruct

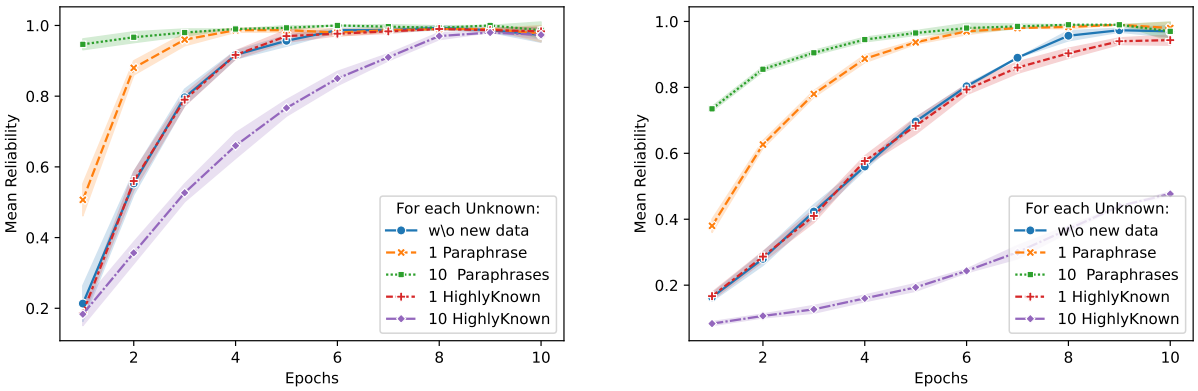

Figure 2: **Dynamics of the reliability score** during training on 500 (left) and 3,000 (right) `Unknown` items along with paraphrases and `HighlyKnown` facts. Error bar is min-max for 3 seed run.

For a paraphrase, this result is not surprising. Increasing training data with different augmentation methods shows an increase in the speed and quality of model training (Zhou et al., 2024; Wang et al., 2024c; Voznyuk and Konovalov, 2024). Also, `HighlyKnown` variant is only harmful for convergence speed, not the model's ability to learn new knowledge.

5.2 Knowledge Shifts

In Table 3, knowledge shifts of the trained models are grouped by the number of `Unknown` samples trained. In the columns we differentiate the training modes: either there was 0, 1 or 10 additional paraphrases or `HighlyKnown` facts per sample. We single out positive shifts and negative shifts. *Positive shifts* are those in which the knowledge category of the (q, a) pair improves. These are the shifts from UK $\rightarrow$ HK, UK $\rightarrow$ MK and MK $\rightarrow$ HK. In contrast, a *negative shift* is a shift, where the knowledge category of the (q, a) fact gets worse. These are the shifts from HK $\rightarrow$ UK, HK $\rightarrow$ MK and MK $\rightarrow$ UK.

If in the previous sections adding paraphrased facts seems to give better results in terms of convergence and maximum amount of knowledge learned, it is clear from Table 3 that training with `HighlyKnown` samples is a winning strategy: it both maximizes positive and minimizes negative shifts.

Although we lose more than we win in almost all training modes, since the negative shift is higher than the positive one, we see that with the increase of the `Unknown` data learned the difference between positive and negative shifts shrinks. However, this observation is true only for the small amount of extra `HighlyKnown` or paraphrased data.

	0	1 HK	1 Paraphrase	10 HK	10 Paraphrase
1 UK					
positive shift	0.034	0.036	0.029	**0.056**	0.045
negative shift	**0.053**	0.054	0.056	0.118	0.067
10 UK					
positive shift	0.021	0.051	0.049	**0.068**	0.038
negative shift	0.245	0.181	**0.154**	0.158	0.187
50 UK					
positive shift	0.06	0.071	0.069	**0.078**	0.07
negative shift	0.148	0.138	0.159	**0.16**	0.174
100 UK					
positive shift	0.067	**0.083**	0.078	0.071	0.064
negative shift	0.172	**0.151**	0.154	0.181	0.204
500 UK					
positive shift	0.096	0.1	**0.105**	0.089	0.076
negative shift	0.195	**0.191**	0.194	0.213	0.25
3,000 UK					
positive shift	0.222	**0.229**	0.222	0.163	0.213
negative shift	0.235	0.202	0.23	**0.149**	0.253

Table 3: **Positive and negative shifts.** Each minitable compares positive and negative shifts for amount of unknown facts learned. Columns represent extra training data used: either HK or paraphrasing. **Green** numbers indicate maximum positive shift for the amount of UK learned, while **red** numbers indicate minimum negative shift for UK learned.

5.3 Benchmarks

We only check the quality of trained LoRA adapters against our test data, so it is not clear whether some of LM's key capabilities have been broken. To fill this gap, we have checked the reasoning abilities of the models on the MMLU benchmark (see Figure 3). Adding 10 `HighlyKnown` or paraphrased samples to train leads to a significant drop in accuracy. On the other side, checking truthfulness on TruftfulQA we see that MC1 and MC2 accu-

racy scores are significantly higher for the training mode with extra paraphrased samples. For a detailed overview of the metrics for external benchmarks, see Table 7 in the Appendix A. An additional evaluation for the ARC (Clark et al., 2018) and LogiQA (Liu et al., 2020) benchmarks is presented in Table 8 in the Appendix E.

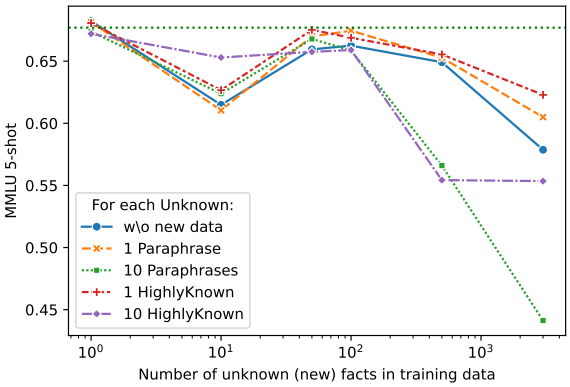

Figure 3: **MMLU:** Accuracy dependent on the amount of Unknown learned. Pointed horizontal line indicates the baseline. Models trained with less additional data tend to disrupt reasoning less.

5.4 Interpretation of Knowledge Shifts

In this section, we analyze the reasons for knowledge shifts. In particular, the reasons of shifting from `HighlKnown` to `Unknown` and from `Unknown` to `HighlyKnown` are considered. The first shift is highly undesirable since the model starts forgetting the knowledge it was completely sure about. The second type of shift is unexpected, since there are a number of shifts from `Unknown` to `Highly Known` in cases for which the model was not initially trained. However, this shift is of particular importance since it makes possible knowledge transfer from training examples to previously unknown facts.

Two general trends in training models have been observed and are presented in Table 4. First, the default `Llama-3.1-8B-Instruct` refuses to answer the questions in 15% of cases. As an answer to the question it produces the following patterns: *"I couldn't find any information ..."*, or *"I cannot verify the ..."*. However, Table 4 shows that almost all models trained in the `Unknown + HighlyKnown` mode lose this ability.

Second, in the test data there are a number of questions for which the answer is the same. For example, for both questions *"Which borough does Aldersgate belong to?"* and *"In which city is the location of General Motors Diesel?"* the answer is *London*. For the default model, the number of unique answers is 48,136 (with respect to repetitions of the same question in greedy search). For some LoRA configurations, there is a significant increase in the number of unique answers (like 100 UK + 10 HK) while for others there is a significant drop. Consider the configurations 10 UK + 0 HK or 10 UK + 1 HK, there is a twofold decrease in the number of unique answers. Although the number of unique answers has decreased dramatically, the number of questions remains the same, which means the model has suggested the same variant for the larger number of questions. For the default model, the answer with the largest number of cases where it is true is *Animalia*: 661 cases. For the aforementioned configurations, it is *Alençon*: 9,393 cases. The large variance in some configurations indicates that trained LoRA models are beginning to converge on some answers (*'exploded'* variants).

	# Refused	# Unique	Mean (Variance)
Default	**3,189**	48,136	3.72 (10.96)
1 UK + 0 HK	758	48,084	4.17 (13.61)
+ 1 HK	816	46,966	4.31 (12.91)
+ 10 HK	**0**	43,766	4.81 (14.26)
10 UK + 0 HK	**0**	**22,804**	**9.22 (96.52)**
+ 1 HK	**0**	**22,148**	**9.38 (166.57)**
+ 10 HK	5	36,798	5.71 (38.26)
50 UK + 0 HK	**0**	37,394	5.62 (38.26)
+ 1 HK	**0**	52,253	4.02 (14.72)
+ 10 HK	**0**	47,734	4.40 (15.14)
100 UK + 0 HK	1	49,403	4.26 (16.58)
+ 1 HK	1	53,576	3.92 (11.74)
+ 10 HK	**0**	59,914	3.51 (12.02)
500 UK + 1 HK	1	48,446	4.34 (16.17)
+ 10 HK	**0**	57,114	3.68 (12.97)

Table 4: **Trends** for the answers' **refusal** and **diversity** in trained models. Default model refuses to answer in 15% cases, while trained models almost never. Some models converge to answers that become overrepresented.

Metrics Table 5 summarizes the absolute amount of shifts from `Unknown` to `HighlyKnown` ($UK \rightarrow HK$) and from `HighlyKnown` to `Unknown` ($HK \rightarrow UK$). The numbers for the shifts are relative and reflect the percentage of cases that fall into this or that reason. The reasons we consider are as follows:

1. *non-refusal* – the percentage of cases where the default model refused to answer while the trained model produced correct answer
2. *explosion* – the percentage of cases where the model predicted one of the exploded answers.

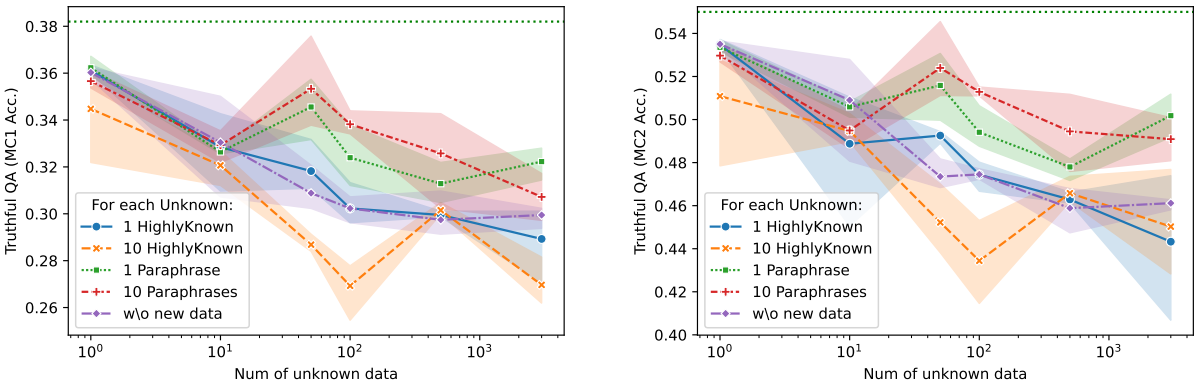

Figure 4: **TruthfulQA:** MC1 and MC2 accuracy metrics dependent on the amount of Unknown learned. Horizontal dotted lines indicate the baselines. Models trained with paraphrases tend to disrupt truthfulness less. Error bar is min-max for 3 seed run.

	Positive shifts						Negative shifts				
Model	*UK* → *HK*	non-refusion	expl-osion	target-based	domain shift	shift explained	*HK* → *UK*	expl-osion	target-based	domain shift	shift explained
1 UK + 0 HK	4	0.25	0.00	**0.50**	**0.50**	0.75	5	0.00	0.00	0.00	0.00
+ 1 HK	4	**0.75**	0.00	0.00	0.00	0.75	5	0.00	0.00	0.00	0.00
+ 10 HK	45	**0.20**	0.00	0.07	0.09	0.29	153	0.01	0.00	**0.02**	0.03
10 UK + 0 HK	111	0.21	0.27	0.24	**0.35**	0.54	409	**0.29**	0.04	0.27	0.48
+ 1 HK	140	0.12	0.14	0.17	**0.32**	0.65	709	0.25	0.10	**0.30**	0.61
+ 10 HK	203	0.10	0.10	0.15	**0.27**	0.33	512	0.08	0.09	**0.25**	0.28
50 UK + 0 HK	241	0.12	0.15	0.40	**0.57**	0.63	392	0.07	0.15	**0.45**	0.50
+ 1 HK	153	0.19	0.05	0.32	**0.60**	0.71	275	0.03	0.00	**0.43**	0.43
+ 10 HK	255	0.14	0.03	0.30	**0.58**	0.63	501	0.01	0.04	**0.43**	0.44
100 UK + 0 HK	185	0.20	0.08	0.36	**0.69**	0.77	314	0.05	0.11	**0.51**	0.54
+ 1 HK	264	0.14	0.00	0.39	**0.73**	0.78	425	0.00	0.08	**0.53**	0.55
+ 10 HK	213	0.12	0.01	0.41	**0.73**	0.79	618	0.01	0.06	**0.49**	0.52
500 UK + 1 HK	748	0.12	0.07	0.79	**0.84**	0.95	802	0.01	0.23	**0.81**	0.85
+ 10 HK	568	0.11	0.01	0.84	**0.86**	0.97	1,134	0.01	0.22	**0.80**	0.83

Table 5: **Reasons for knowledge shifts.** UK → HK and HK → UK indicate absolute amount of shifts. Each reason reflects the relative contribution to understanding the nature of shifts. Bold numbers reflect a main reason of shift, underlined numbers - the second best reason of shift.

The exploded answer is the one to which the model converges in percentage of cases considerably higher than expected by the default model

3. *target-based* – the percentage of cases where the model predicts one of the answers from the unknown examples it was trained on
4. *domain shift* – all questions in the dataset were constructed by templates in a form of `<subject, relation, object>`. Each relation in DBpedia has either its range or domain. Both give insight into the way the relation links a subject to an object. In case of domain, the subject qualifies as a type of thing specified in the domain. The range works exactly like the domain but applies to the object. 390 relations from the test data fall into 92 relation domain categories. For example, the relation domain "PopulatedPlace" includes 24 relations, allowing us to analyze shifts of knowledge of a large amount of similar relations instead of looking at the specific one. *domain shift* shows the percentage of cases where the shift goes from the same domain or range of the relation for the cases the model was trained on
5. *shift explained* – the percentage of shift explained by all aforementioned reasons. Note that the sum of the rates for all reasons does not necessarily sum to the percentage of the shift explained, since some of the reasons overlap.

Except for the *non-refusal* all reasons fall into both positive and negative shifts. For example, we see that both positive and negative shifts occur inside the domain. Specific examples of positive and negative shifts are presented in Appendix D.

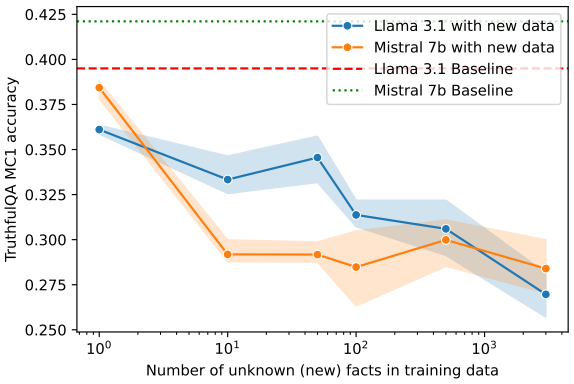

Figure 5: Difference in quality drop for the two models: Llama 3.1 8B and Mistral 7B v0.3.

Discussion If we train the model just for 1 new unknown fact, the positive shift occurs mainly because the model starts answering questions it has not answered previously.

From the exploded answers suffer the models (10 UK+ 1 HK and 10 UK + 0 HK) that have a low amount of training data with a large proportion of new trained unknown facts relative to the size of the training data.

Finally, we can see that with the increase in the amount of learned unknown facts, there is an increase of the rate of domain shift and target spillover for positive shifts. On the other hand, for the negative shifts, there is only the tendency for the increase of the rate of the domain shift explaining the proportion of the shift explained. Besides, if we compare target-based percentages for positive and negative shifts, we can see that for all models this percentage is higher for positive shifts. The same is true for the negative shifts and for the domain shift as well. It means that out of these two reasons the positive effect of learning from the same domain and target is higher than negative.

6 Additional Results for Mistral

Similar effects to those described above for the LLaMa model may be observed in other models, such as GPT-3 and Mistral. We have conducted a set of experiments for the Mistral-7B-Instruct-v0.3 model, to verify if the results obtained for the Llama-3.1-8B-Instruct model are generalizable to other decoder-only models.

As we can see in Table 6, the number of Paraphrases and `HighlyKnown` increases, so does the positive shift. This corroborates our findings on Llama-3.1-8B-Instruct.

Method	**Accuracy shift**	
	Pos. shift	Neg. shift
1 Unknown + 1 Paraphrase	0.159	0.106
1 Unknown + 10 Paraphrases	0.196	0.114
1 Unknown + 1 HighlyKnown	0.160	0.114
1 Unknown + 10 HighlyKnown	0.174	0.146

Table 6: Accuracy shift for Mistral-7B-Instruct-v0.3.

As can be seen in Figure 5, the quality of the model decreases as the number of new knowledge increases. Models are not strongly correlated, but we see similar features, such as a significant drop from 1 new knowledge to 10, and a small recovery of quality at 100 new knowledge. Thus, we believe that generalizing our research to other decoder-only models is possible.

7 Conclusion

Our findings revealed that the most significant increase in the acquisition of additional knowledge is achieved when a blend of `Unknown` and `HighlyKnown` data is incorporated into the training phase. However, this approach comes with a trade-off: it compromises the model's capability to accurately answer complex or nuanced questions, as assessed by TruthfulQA.

Additionally, it has been observed that the incorporation of a limited number of unknown facts within a vast set of `HighlyKnown` examples or paraphrases can substantially impair the model's reasoning abilities as measured by MMLU.

In an additional observation, it was noted that upon being fine-tuned with LoRA adapters, LLMs experience a reduction in their ability to express uncertainty when formulating answers. In certain scenarios, this leads to the models disproportionately favoring responses that are statistically overrepresented.

Acknowledgement

This research was conducted within the framework of a joint MTS-Skoltech laboratory on AI.

Limitations

Since we only used one LLM for our study (namely, Llama-3.1-8B-Instruct), it's unknown if the outcomes will change when utilizing different LLMs. However, because our work requires a lot of computation, it is difficult to repeat on several LLMs.

First, we experimented a lot with different learning rates, LoRA ranks and used two data augmentation techniques with 0, 1, 10 extra examples per sample. This results in a great number of experiments. Second, and perhaps most crucially, we had to annotate a sizable dataset with the knowledge categories to make our analysis easier. It was essential to precisely evaluate the model's knowledge with regard to a particular fine-tuning case in order to draw trustworthy findings. Beyond this, detecting knowledge categories is a machine-intensive task not only for defining the initial categories, but at each inference step.

In addition, there are a few crucial directions for future research left unexplored in this paper. Namely, how early stopping could have influenced the distribution of knowledge categories and models' behavior on external benchmarks. Adding training examples from the same relation domain or range as well as few-shot prompts from this category could have also boosted the performance. We leave these questions for further exploration.

Ethical Considerations

We have carefully curated the generated dataset, and we have not encountered any inappropriate or offensive content within it.

References

Everton Lima Aleixo, Juan Gabriel Colonna, Marco Cristo, and Everlandio Fernandes. 2024. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *J. Braz. Comput. Soc.*, 30(1).

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.2, knowledge manipulation. *Preprint*, arXiv:2309.14402.

Julia Belikova, Evegeniy Beliakin, and Vasily Konovalov. 2024. JellyBell at TextGraphs-17 shared task: Fusing large language models with external knowledge for enhanced question answering. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 154–160, Bangkok, Thailand. Association for Computational Linguistics.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge. *CoRR*, abs/2102.03315.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yiren Chen, Mengjiao Cui, Ding Wang, Yiyang Cao, Peian Yang, Bo Jiang, Zhigang Lu, and Baoxu Liu. 2024. A survey of large language models for cyber threat detection. *Comput. Secur.*, 145:104016.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Mehrdad Farahani and Richard Johansson. 2024. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 16966–16977. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

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? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7765–7784. Association for Computational Linguistics.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *CoRR*, abs/2403.14608.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting edit failures in large language models: An improved specificity benchmark. *CoRR*, abs/2305.17553.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024a. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1416–1428. Association for Computational Linguistics.

Wenyu Huang, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, and Jeff Z Pan. 2024b. Prompting large language models with knowledge graphs for question answering involving long-tail facts. *arXiv preprint arXiv:2405.06524*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Nikita Krayko, Ivan Sidorov, Fedor Laputin, Daria Galimzianova, and Vasily Konovalov. 2024. Efficient answer retrieval system (EARS): Combining local DB search and web search for generative QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1584–1594, Miami, Florida, US. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dongyang Li, Junbing Yan, Taolin Zhang, Chengyu Wang, Xiaofeng He, Longtao Huang, Hui Xue, and Jun Huang. 2024. On the role of long-tail knowledge in retrieval augmented large language models. *Preprint*, arXiv:2406.16367.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Daniil Moskovskiy, Sergey Pletenev, and Alexander Panchenko. 2024. Llms to replace crowdsourcing for parallel data creation? the case of text detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14361–14373. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (llms)? A.K.A. will llms replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 311–325. Association for Computational Linguistics.

Anastasia Voznyuk and Vasily Konovalov. 2024. DeepPavlov at SemEval-2024 task 8: Leveraging transfer learning for detecting boundaries of machine-generated texts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1821–1829, Mexico City, Mexico. Association for Computational Linguistics.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024a. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Preprint*, arXiv:2405.14768.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024b. Knowledge editing for large language models: A survey. *ACM Comput. Surv.* Just Accepted.

Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. 2024c. A comprehensive survey on data augmentation. *CoRR*, abs/2405.09591.

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

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10222–10240. Association for Computational Linguistics.

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024. A survey on data augmentation in large model era. *ArXiv*, abs/2401.15422.

A MMLU Performance

			MMLU		ThruthfulQA					
			Accuracy	**Std**	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**MC1 Acc.**	**MC2 Acc.**
Llama-3.1-8B-Instruct			0.677	0.004	35.780	60.325	47.789	58.512	0.382	0.552
1 UK	+	0 HK	0.682	0.004	33.741	57.414	44.176	55.374	0.359	0.533
	+	1 HK	0.681	0.004	34.920	58.537	45.814	56.587	0.361	0.538
	+	10 HK	0.672	0.004	25.367	44.604	30.873	42.816	0.322	0.478
10 UK	+	0 HK	0.615	0.004	12.524	22.778	11.987	21.519	0.305	0.481
	+	1 HK	0.627	0.004	13.251	27.477	15.078	25.530	0.310	0.451
	+	10 HK	0.653	0.004	21.659	39.459	25.678	37.562	0.319	0.490
50 UK	+	0 HK	0.660	0.004	18.906	34.491	19.259	32.569	0.306	0.482
	+	1 HK	0.675	0.004	18.740	35.436	19.318	33.398	0.313	0.494
	+	10 HK	0.658	0.004	5.592	20.859	9.420	19.321	0.282	0.448
100 UK	+	0 HK	0.663	0.004	16.987	34.884	19.598	32.935	0.304	0.474
	+	1 HK	0.669	0.004	19.586	37.386	22.375	35.500	0.315	0.482
	+	10 HK	0.659	0.004	9.609	27.236	13.829	25.363	0.277	0.452
500 UK	+	0 HK	0.649	0.004	10.115	25.991	12.541	23.909	0.290	0.447
	+	1 HK	0.655	0.004	7.561	23.507	11.480	21.498	0.297	0.460
	+	10 HK	0.554	0.004	6.829	23.143	9.583	21.519	0.296	0.463
3000 UK	+	0 HK	0.579	0.004	11.415	27.783	14.884	25.363	0.294	0.461
	+	1 HK	0.623	0.004	5.561	19.906	7.422	18.280	0.257	0.420
	+	10 HK	0.554	0.004	9.239	23.447	11.558	21.415	0.263	0.445
1 UK	+	0 Paraphrase	0.682	0.004	33.741	57.414	44.176	55.374	0.359	0.533
	+	1 Paraphrase	0.681	0.004	36.582	60.991	48.642	59.052	0.365	0.537
	+	10 Paraphrase	0.674	0.004	35.930	59.606	46.787	57.851	0.356	0.535
10 UK	+	0 Paraphrase	0.615	0.004	12.524	22.778	11.987	21.519	0.305	0.481
	+	1 Paraphrase	0.610	0.004	18.518	36.578	23.061	34.481	0.343	0.527
	+	10 Paraphrase	0.624	0.004	15.420	31.984	20.046	29.987	0.332	0.495
50 UK	+	0 Paraphrase	0.660	0.004	18.906	34.491	19.259	32.569	0.306	0.482
	+	1 Paraphrase	0.670	0.004	17.795	37.690	21.159	35.887	0.344	0.517
	+	10 Paraphrase	0.668	0.004	17.944	38.240	22.340	35.891	0.337	0.511
100 UK	+	0 Paraphrase	0.663	0.004	16.987	34.884	19.598	32.935	0.304	0.474
	+	1 Paraphrase	0.675	0.004	19.882	40.465	24.316	38.299	0.334	0.507
	+	10 Paraphrase	0.659	0.004	25.621	47.628	32.787	45.445	0.337	0.511
500 UK	+	0 Paraphrase	0.649	0.004	10.115	25.991	12.541	23.909	0.290	0.447
	+	1 Paraphrase	0.653	0.004	14.245	33.977	17.747	31.160	0.322	0.472
	+	10 Paraphrase	0.566	0.004	10.514	26.604	12.961	24.337	0.343	0.512
3000 UK	+	0 Paraphrase	0.579	0.004	11.415	27.783	14.884	25.363	0.294	0.461
	+	1 Paraphrase	0.605	0.004	17.107	38.457	22.222	35.300	0.319	0.502
	+	10 Paraphrase	0.441	0.004	29.352	50.871	40.383	49.400	0.305	0.491

Table 7: Accuracy for MMLU and a range of metrics for ThruthfulQA for all trained LoRA adapters

B Few-shot Prompts

Here is an example of the 4-shot prompt. The prompt includes 4 question-answer examples, 5th question is the question of interest.

Answer the following question.
Question: Who wrote the novel Evening Class? Answer: maeve binchy
Question: Which country does the airline Air Pacific come from? Answer: fidji
Question: In which branch of the arts does Allegra Kent work? Answer: balletti
Question: Who had a 70s No 1 hit with Billy, Don't Be A Hero? Answer: bo donaldson and heywoods
Question: 12th Brigade (Australia) fought in what battle?

C Samples of Knowledge Categories

`HighlyKnown`. *Question: where is Alfa Romeo MiTo assembled? Answer: Turin*
`MaybeKnown`. *Question: Daredevil (TV series) is on which channel? Answer: Netflix*
`Unknown`. *Question: Can you name a band member of Ashes of Ares? Answer: Matthew Barlow*

D Examples of Positive and Negative Shifts

Let's illustrate the concept of positive shifts through examples. This occurs when a model begins to accurately respond to queries for which it has not been specifically trained (A_{before} – an answer before fine-tuning, A_{after} – an answer after fine-tuning).

Positive Shifts:

Q: *Batata vada is located in which region?* A_{before}: *India* A_{after}: *Maharashtra*
Q: *Esslingen am Neckar is in which administrative?* A_{before}: *Esslingen* A_{after}: *Stuttgart Government Region*
Q: *What is the home arena of Anaheim Storm?* A_{before}: *I couldn't find any information* A_{after}: *Anaheim Arena*

Negative Shifts:

Q: *Where did John Bigger pass away?* A_{before}: *I cannot verify where* A_{after}: *London*
Q: *In which canton is Gachnang located?* A_{before}: *Thurgovia* A_{after}: *aargau*

E ARC and LogiQA Benchmarks

Method	**ARC**		**LogiQA**
Llama-3.1-8B-Instruct	ARC-E	ARC-C	Acc
1 Unknown + 0 Paraphrase	0.7967	0.5520	0.3210
1 Unknown + 1 Paraphrase	0.7942	0.5512	0.3226
1 Unknown + 10 Paraphrase	0.7875	0.5546	0.3164
10 Unknown + 0 Paraphrase	0.7252	0.5213	0.3272
10 Unknown + 1 Paraphrase	0.7386	0.5350	0.3149
10 Unknown + 10 Paraphrase	0.7142	0.5299	0.3088
1 Unknown + 0 HighKnown	0.7967	0.5520	0.3210
1 Unknown + 1 HighKnown	0.7988	0.5538	0.3149
1 Unknown + 10 HighKnown	0.7723	0.5333	0.3287
10 Unknown + 0 HighKnown	0.7252	0.5213	0.3272
10 Unknown + 1 HighKnown	0.7218	0.5247	0.3041
10 Unknown + 10 HighKnown	0.7517	0.5324	0.2980

Table 8: Metrics for ARC and LogiQA benchmarks for trained LoRA adapters