# Conditional Language Learning with Context

**Xiao Zhang** [1]  **Miao Li** [1]  **Ji Wu** [1][2]

## Abstract

Language models can learn sophisticated language understanding skills from fitting raw text. They also unselectively learn useless corpus statistics and biases, especially during finetuning on domain-specific corpora. In this paper, we propose a simple modification to causal language modeling called conditional finetuning, which performs language modeling conditioned on a context. We show that a context can "explain away" certain corpus statistics and make the model avoid learning them. In this fashion, conditional finetuning achieves selective learning from a corpus, learning knowledge useful for downstream tasks while avoiding learning useless corpus statistics like topic biases. This selective learning effect leads to less forgetting and better stability-plasticity tradeoff in domain finetuning, potentially benefitting lifelong learning with language models.

## 1. Introduction

Language models pretrained on large-scale corpus have shown impressive performance on a wide variety of downstream tasks (Chung et al., 2022; Touvron et al., 2023a; OpenAI, 2023). It is impressive that these language models learn sophisticated knowledge and reasoning abilities solely from training on raw text with a causal language modeling objective. The objective is also used when adapting the pretrained general-purpose language models to specific domains, via finetuning on a domain corpus (also called "continual pretraining") (Chen et al., 2021; Lewkowycz et al., 2022; Singhal et al., 2023).

Although finetuning effectively improves the model's domain knowledge and performance on domain tasks, it can also lead to forgetting of existing knowledge (Chen et al.,

2020; Jang et al., 2022) due to modifying the pretrained model. It is also observed that finetuning can lead to over-adaptation to the statistical properties of the domain corpus, causing the model to be biased heavily towards certain topics and styles (Zhang & Wu, 2024).

In domain finetuning, it would be desirable to improve the model's domain knowledge without learning useless statistics and biases from the corpus. The causal language modeling objective maximizes likelihood of all the tokens in the corpus, and is therefore unselective in what kind of information to learn from the corpus. In this paper, we propose a simple enhancement to causal language modeling called conditional finetuning, that uses contexts to achieve selective learning of useful information from the corpus.

It is well-known that the behavior of pretrained language models is sensitive to contextual information in the input during inference. For example, few-shot prompting can let models learn to perform tasks based on examples in the context (Dong et al., 2023). Specific instructions like chain-of-thought (Wei et al., 2022) and self-verification (Weng et al., 2023) could guide the model towards certain behaviors like multi-step reasoning. For dialog and assistant use cases, language models can be further finetuned on instruction-following data to make them more sensitive to instructions in the context (Ouyang et al., 2022; Chung et al., 2022; Sanh et al., 2022). While the effect of context during inference has been extensively studied, its role during the pretraining phase is less explored. In this paper, we investigate how adding a context to language modeling could affect the model's learning behavior during pretraining and domain finetuning.

The two main contributions of the paper are:

We propose conditional finetuning, a domain finetuning method for language models that adds a context to causal language modeling. We reveal how adding a context affects the language modeling objective. In conditional finetuning, we use a piece of text as context and prepend it to corpus text during finetuning. The idea of the method is illustrated in Figure 1. We show that the context can "explain away" statistical properties of the corpus so that the model would ignore them and avoid learning them in finetuning. For example, when finetuning on a domain corpus with a domain hint as context, the model can keep its topic prior almost

---

[1]Department of Electronics Engineering, Tsinghua University [2]College of AI, Tsinghua University. Correspondence to: Ji Wu <wuji_ee@mail.tsinghua.edu.cn>.
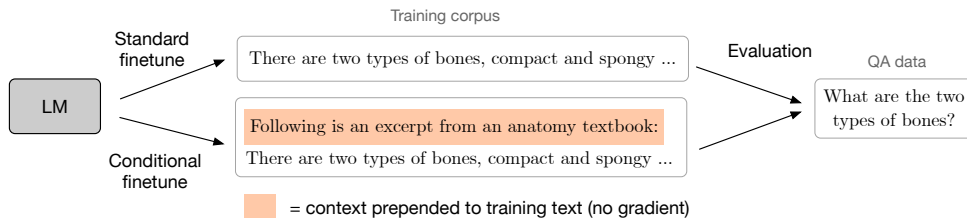
*Figure 1.* Illustration of conditional finetuning a language model on domain corpus. Compared to standard finetuning, conditional finetuning prepends a context to each document and only learns information conditioned on the context.

unchanged, without adapting to the topic distribution of the domain corpus, as typically observed in conventional finetuning.

We also show that conditional finetuning achieves selective learning of useful information. Knowledge useful for downstream tasks is learned without compromise, while learning of corpus statistics is reduced, which leads to significantly less modification to the pretrained model during finetuning. This selective learning leads to better stability-plasticity tradeoff and less forgetting in one-time finetuning (transfer learning) and multiple-time finetuning (continual learning) scenarios, making it a better alternative to conventional finetuning for lifelong learning with language models.

We release our code implementation along with the original part of the data used in the paper[1].

## 2. Related Work

**Domain finetuning of language models.** Finetuning pretrained language models on domain corpus is a common approach to enhances domain knowledge and help language models perform better on domain tasks, for example, in mathematics (Lewkowycz et al., 2022), coding (Chen et al., 2021) and medicine (Singhal et al., 2023). Finetuning on multiple corpora can help language models continually learn knowledge from multiple domains (Gupta et al., 2023; Jin et al., 2022; Ke et al., 2023), as a form of continual or lifelong learning (Thrun, 1998). Besides learning knowledge, finetuning can also lead to over-adaptation to the statistics of domain corpus like topic and style, as an unwanted side effect (Du et al., 2024; Zhang & Wu, 2024).

**Inferencing with context: in-context learning.** With the extensive use of autoregressive language models for question answering and problem-solving, contexts (prompts) are often used to provide extra information or guidance to the model. For example, In-context learning (Dong et al., 2023) provide a few examples as context for the model to learn from during inference. Chain-of-thought (Wei et al., 2022)

---

[1] https://github.com/xiaozeroone/conditional_finetune

and tree-of-thought (Yao et al., 2023) use special instructions to guide the model to perform step-wise reasoning. Self-verification (Weng et al., 2023) also uses the model's own prediction as context to perform an extra verification step to improve the accuracy of reasoning.

**Learning context for inference: prompt learning.** Because context can significantly affect language model's performance in inference, prompt learning methods are used to learn optimal contexts for a target task. Prompt tuning (Lester et al., 2021) and prefix tuning (Li & Liang, 2021) optimize soft prompts for target tasks as a parameter-efficient tuning alternative to full finetuning. For better expressivity, soft prompts can be individually given to each layer of the model (Liu et al., 2022). Gradient-free black-box optimization (Sun et al., 2022) can also be used to learn prompts where models weights are inaccessible, such as when using commercial models.

**Training on context: instruction finetuning.** To enhance language model's ability to use context, instruction finetuning (Ouyang et al., 2022) finetunes model on instruction-response pairs to make model better at following instructions. Training with diverse instructions, e.g., FLAN (Chung et al., 2022) and T0 (Sanh et al., 2022), helps model generalize to new instructions and tasks. Training with few-shot and chain-of-thought examples also helps model better utilize those kinds of context in inference (Longpre et al., 2023).

Such methods usually train with loss on the "response" part of the instruction-response pairs, using the instruction as context. The goal is to learn the relationship between various instructions and their corresponding responses for better instruction following. By contrast, conditional finetuning is used in continual pretraining where the goal is to learn knowledge from corpora. Conditional finetuning does not learn information from the context, but instead uses the context to "explain away" corpus statistics thus reduce learning of those useless statistics.

**Selective learning.** Generative models are trained to directly maximize the likelihood of data, so they tend to indiscriminately learn all patterns within the data. To make

2

models selectively learn certain patterns from data, one can perform data selection (Jain et al., 2023; Xie et al., 2023), choosing subsets of training data that contains the desired patterns, or perform soft example selection by importance sampling (Katharopoulos & Fleuret, 2018). For language models, one can also use loss re-weighting at token level to selectively learn from informative tokens (Hu et al., 2023). Attention guidance is another approach that leverages the mechanistic interpretability of attention to make model "focus" more on certain features in the input, thus selectively learn certain features more (Chrysostomou & Aletras, 2021; Feng et al., 2022; Shi et al., 2023). More related to our approach, it is possible to learn an ensemble of models in order to factor learned patterns into different models. This is successfully used to de-bias models in natural language tasks (Clark et al., 2019; 2020; Sanh et al., 2021).

## 3. Conditional Learning

We first use the language of probabilistic modeling to illustrate the idea of conditional learning.

Consider a probabilistic model $p$ of some data. Suppose an example $x$ has a property $c$ that can be inferred from $x$, i.e., $p(c|x) = 1$. Then the probability of $x$ can be decomposed as

$$p(x) = p(x, c) = p(x|c)p(c). \qquad (1)$$

If we want to fit the model on data, increasing the likelihood of $x$ under $p$, we can either increase $p(c)$ or $p(x|c)$. The former is fitting to the property $c$. The latter is leaning the regularities in $x$ *besides* the property $c$, which we refer to as *conditional learning* in this paper.

If a set of examples $\{x_i\}_{i=1}^N$ all have the same property $c$, then the average log-likelihood of the dataset is

$$\frac{1}{N} \sum_i \log p(x_i) = \frac{1}{N} \sum_i \log p(x_i|c) + \log p(c). \qquad (2)$$

When fitting model $p$ to the dataset by maximizing data likelihood, according to Equation (2), increasing $\log p(c)$ will likely increase data likelihood faster than increasing $\log p(x_i|c)$ for certain individual examples. This implies that the model could be biased towards adapting to the common property $c$ of the data if such property is present. Moreover, if the property $c$ is simple, the model may also adapt $p(c)$ faster and earlier than learning $p(x|c)$ (Geirhos et al., 2020; Du et al., 2024).

In this paper, we specifically explore this situation in language modeling. When finetuning a general-purpose language model on a specialized domain corpus, the model can exhibit a noticeable bias towards the domain. This bias arises because the domain acts as a common property among the corpus documents, leading the model to significantly

adapt its topic prior in favor of the domain topic (Zhang & Wu, 2024). Nonetheless, the ultimate goal of finetuning language models is to enhance their domain-specific knowledge without compromising their general knowledge and ability (Chen et al., 2020; Jang et al., 2022).

Knowledge is often embedded in text in the form of conditional token probabilities like $p(x_{[k...n]}|x_{[1...k]})$. For instance, $p(\text{"}London\text{"}|\text{"}The\ capital\ city\ of\ England\ is\text{"})$ can represent the factual knowledge within the sentence "The capital city of England is London".

In this case, learning the conditional probability $p(x_{[k...n]}|x_{[1...k]}, c)$ conditioned on corpus-level properties $c$ (such as the topic of the corpus) is enough for the purpose of knowledge learning. i.e., learning $p(x|c)$ in Equation (1) is sufficient in domain finetuning of language models. Learning $p(\text{topic})$ and other corpus properties are not necessary and may be harmful in case of lifelong learning (Thrun, 1998) because they introduce unnecessary bias.

Luckily, it is straightforward to perform conditional learning in causal language modeling, where $p(x)$ is decomposed as probabilities of each token given the previous tokens:

$$p(x) = \prod_{i=1}^n p(x_i|x_{<i}). \qquad (3)$$

Now we can explicitly prepend the property $c$ (in text form) to the original text $x$. Here, we use notation $\langle, \rangle$ to indicate concatenation of text. Under causal language modeling, applying Equation (3) gives the decomposition

$$p(\langle c, x \rangle) = p(x|c)p(c). \qquad (4)$$

To only learn the conditional probability $p(x|c)$, we feed the concatenated sequence $\langle c, x \rangle$ into the model and use the negative log-likelihood on tokens in $x$ as loss, which directly corresponds to $p(x|c)$. It can be implemented on standard causal language modeling by simply masking out the loss on the first few tokens corresponding to context $c$.

In theory, we could prepend any context $a$ to the text, not necessarily those representing properties of the corpus. In this case, we will have

$$p(\langle a, x \rangle) = p(x|a)p(a) = p(x|c, a)p(c|a)p(a) \qquad (5)$$

we will see later from experiments that supervising $p(x|a)$ will lead to learning both the conditional probability $p(x|c, a)$ and a "conditional prior" $p(c|a)$, such as a conditional topic prior $p(\text{topic}|a)$ activating domain topics when the context $a$ is given.

**Selective learning.** Conditional learning is a form of selective learning: it factorizes the objective in Equation (2) into two parts and learns one part of it. It turns the unselective learning objective of language modeling into a selective

learning objective that only learns knowledge useful for downstream tasks and avoids learning corpus statistics, reducing side effects such as over-adaptation and forgetting.

# 4. Conditional Language Modeling with Context

In this section, we apply conditional learning to language modeling and finetune a pretrained language model on a domain corpus. We analyze the effect of conditional learning and show that it reduces the learning of the topic prior $p(\text{topic})$. Instead, the model uses a conditional prior $p(\text{topic}|c)$ to fit the topic distribution of the domain corpus.

**Data and model.** We use the medical textbooks provided with the MedQA dataset (Jin et al., 2021) as a domain corpus to finetune LLaMA-2 (Touvron et al., 2023b), a series of state-of-the-art language models pretrained on large-scale general text. The medical textbooks collection contains 18 textbooks on various medical subjects and has a total of 25.7M tokens. The textbooks are dense in medical knowledge and are a good candidate for studying domain-specific knowledge learning.

We finetune the model with the AdamW optimizer (Loshchilov & Hutter, 2019), a learning rate of 3e-5, and a batch size of 16. The maximum sequence length is set to 2048. We use the Transformers library (Wolf et al., 2020) and an NVIDIA A100 GPU for the experiments.

**Conditional finetuning.** In conditional finetuning, we prepend a context $c$ to each document $x$ and use $\langle c, x \rangle$ instead of $x$ as training examples. The training objective is the same as conventional causal language modeling, except that the loss from the first $|c|$ tokens are ignored ($|c|$ = the number of tokens in $c$). Conditional finetuning uses the same hyperparameters as standard finetuning.

The context $c$ in only used in training. During inference, conditionally finetuned models are used *without* the context just as normal language models.

## 4.1. Conditional Finetuning Reduces Learning of the Topic Prior $p(\text{topic})$

To examine the effect of conditional finetuning, we first use a short sentence hinting the topic as the context, i.e., $c =$ "*Following is an excerpt from a medical textbook.*"

To measure the topic prior of a language model, we use a simple topic likelihood probe. The probe is a sentence "*The text is about [topic].*" where [topic] is replaced by a topic word. The likelihood of the topic word $p([\text{topic}]|\text{"The text is about"})$ given by the model indicates the model's topic prior.
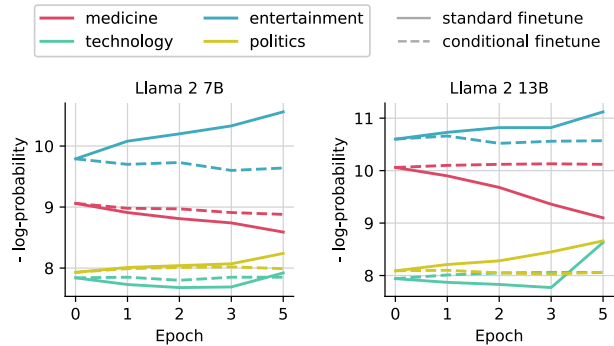


*Figure 2.* Topic likelihood changes during finetuning. Unlike standard finetuning, conditional finetuning does not significantly change topic likelihoods.

Figure 2 shows that topic likelihoods changes significantly during standard finetuning. Finetuning typically increases the likelihood of the domain topic and decreases the likelihood of other topics. Conditional finetuning, on the other hand, keeps the topic likelihoods stable even after several epochs of training. This suggests that when topic information is already given as context, the model no longer need to fit its topic prior to the training text, confirming the conditional learning hypothesis in Section 3.

Next, we show how conditional finetuning affects the modeling of a longer piece of text. We calculate the finetuned models' average language modeling loss on medical textbooks and C4 (Raffel et al., 2020), a corpus of general web text. We calculated average loss separately at each token position, in a similar fashion as in Zhang & Wu (2024).

Figure 3 (left) shows that standard finetuning significantly changes loss on the first few tokens of text. Loss decreases on the training corpus and significantly increases on C4. This suggests that the finetuned model becomes highly predisposed to medical texts from the beginning, likely due to an over-adaptation of the topic prior. On the other hand, conditional finetuning results in a negligible change in loss across all positions on the C4 corpus. This ability to keep loss stable on general corpus suggests a potential to mitigate issues related to over-adaptation and forgetting in finetuning.

## 4.2. Conditional Finetuning Learns a Conditional Topic Prior $p(\text{topic}|a)$ regardless of Context $a$

Interestingly, we found that conditional finetuning always learns a conditional topic prior $p(\text{topic}|a)$ instead of an unconditional topic prior $p(\text{topic})$, regardless of the context $a$ given in training. To see this, we experiment with three types of context:

***Domain hint*** (informative context): A short sentence hinting the topic: "*Following is an excerpt from a medical textbook.*"
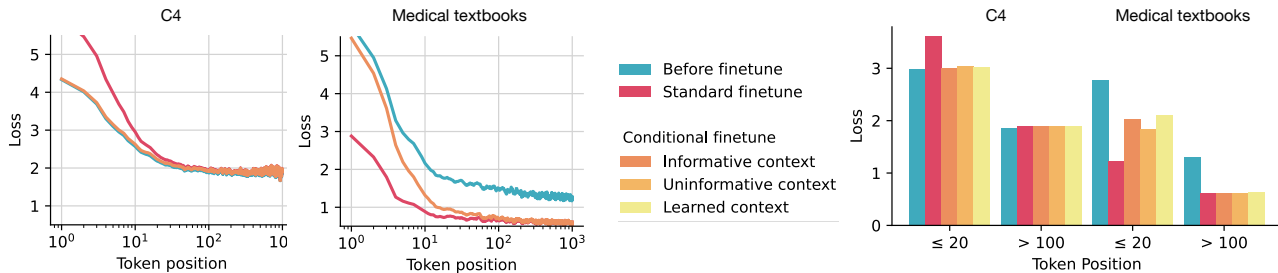
*Figure 3.* Comparing language modeling loss at different token positions, for models finetuned with standard finetuning and conditional finetuning with three types of context. Unlike standard finetuning, conditional finetuning barely increases loss on general text (C4), regardless of the type of context used in training. (model: LLaMA-2 7B)

***Random*** (uninformative context): A randomly generated Universally Unique Identifier (UUID) string: "*7a1d64b1-fa43-47a8-9389-60406eb96778.*"

***Learned*** (learned context): A soft prompt of 10 vectors learned using the prompt tuning method (Lester et al., 2021). The soft prompt is learned by fitting the training corpus while keeping the language model fixed. The soft prompt is then used as a fixed context during finetuning the whole model. Detailed procedure is described in Appendix B.

Specifically, the learned context is a context that maximizes the likelihood of the training corpus. Therefore, it likely encodes some overall statistics of the corpus, such as the main topic. The context is unlikely to encode detailed knowledge due to the limited capacity of the soft prompt.

Figure 3 (right) compares the loss at different token positions of models finetuned with different context types. All types of contexts are similarly effective at keeping loss unchanged on general corpus (C4). This shows that the presence of any context would reduce learning of the topic prior $p(\text{topic})$ regardless of the content of the context.

The reason is that even when provided with a non-informative context $a$, the model learns a conditional topic prior $p(\text{topic}|a)$ that makes $a$ function similarly as an informative context. We can clearly see this from some examples.

Figure 4 shows an example of the context's effect on the model's prediction. A medical text and a technological text are used as examples. On the original pretrained model, domain hint increases the likelihood of medical terms and decreases the likelihood of technological terms. The random UUID string has the opposite effect, likely because UUID strings are more commonly associated with technological contexts. Despite behaving differently on the pretrained model, all kinds of contexts have a remarkably similar effect on their own conditionally finetuned models. They all make the model strongly favor medical over technological words. This indicates that the model learns a conditional topic prior $p(\text{topic}|a)$ that makes $a$ function similarly as a domain hint. If $a$ is already informative about the topic, $p(\text{topic}|a)$ may
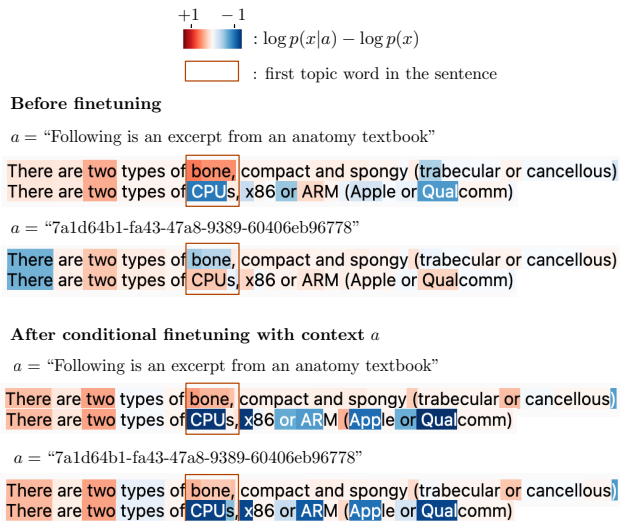


*Figure 4.* Examples showing the loss change $\log p(x) - \log p(x|a)$ caused by the context. Before finetuning, domain hint makes the model favor the medical term "bone" while the random UUID string favors technological term "CPU". After finetuning, both contexts have similar effect of favoring medical terms.

simply be reinforced during conditional finetuning.

## 4.3. Conditional Finetuning does not Affect Knowledge Learning

Figure 3 shows that, for later token positions of the training corpus (e.g., >100), conditional finetuning achieves the same loss reduction as standard finetuning. This suggests that conditional finetuning likely will not affect the learning of factual knowledge, which is mostly in the form of $p(x_{[k...n]}|x_{[1...k]})$ (discussed in Section 3). The hypothesis is verified in the next section where we evaluate performance on downstream tasks. These findings indicate that conditional training is effective at selective learning, learning knowledge useful for future tasks while avoiding learning corpus statistics that are not useful.

5

# 5. Less Forgetting through Selective Learning

Evidence from language modeling suggests that conditional finetuning performs selective learning. In this section, we further elucidate this selective learning effect. Conditional finetuning modifies the pretrained model less than standard finetuning, and therefore helps achieve less forgetting in transfer learning and continual learning scenarios. At the same time, knowledge learning is not affected.

## 5.1. Conditional Finetuning Modifies Model Less

We use two metrics to measure how much the pretrained language model is modified during finetuning. To estimate the influence of the training objectives on the model, we calculate the gradient norm of standard and conditional finetuning objectives on the pretrained model. L2-norm of the gradient is calculated over the training corpus by flattening all parameters of the model into a single vector[2]. As shown in Table 1, conditional training objective has a significantly smaller gradient norm than the standard finetuning objective, even though they both use the same cross-entropy loss on the same training tokens. This indicates that the conditional finetuning objective requires less modification to the model parameters, likely by removing the gradient for fitting corpus statistics.

| Gradient norm of objective | Standard finetune | Conditional finetune w/ domain hint |
|---|---|---|
| *On medical textbooks* | | |
| LLaMA-2 7B | 1.93 | 1.13 |
| LLaMA-2 13B | 1.56 | 1.06 |

*Table 1.* Gradient norm of standard finetune and conditional finetune objectives. The conditional finetune objective has a significantly smaller gradient norm.

To see how much the model changes during finetuning, we can measure the similarity of the model before and after finetuning. We calculate the KL-divergence between the output probability distribution of the pretrained model and the finetuned model, averaged over all tokens. The models are all finetuned for 5 epochs. As shown in Table 2, on C4, models finetuned with conditional finetuning have significantly smaller KL-divergence to the pretrained model than models finetuned with standard finetuning. This confirms that conditional training modifies the model less than standard finetuning.

---

[2]Scaling parameters in layer normalization (Ba et al., 2016) are excluded as they can have large gradients and dominate the L2 norm when included.

| KL-divergence to pretrained model | Standard finetune | Conditional finetune w/ domain hint |
|---|---|---|
| *On C4* | | |
| LLaMA-2 7B | 0.082 | 0.036 |
| LLaMA-2 13B | 0.116 | 0.071 |

*Table 2.* KL-divergence from the finetuned model to the pretrained model. Conditional finetuning results in a significantly smaller KL-divergence than standard finetuning.

## 5.2. Conditional Finetuning Reduces Forgetting and Maintains Knowledge Learning in Transfer Learning

We next show that because conditional finetuning modifies the model less, it achieves less forgetting in transfer learning (one-time finetuning). Also, knowledge learning is uncompromised in selective learning. As a result, conditional training achieves better stability-plasticity tradeoff over learning new knowledge and retaining existing knowledge, the perennial dilemma in lifelong learning (Parisi et al., 2019; Biesialska et al., 2020).

We evaluate knowledge learning in finetuned language models with question answering tasks, a common approach in previous work (Hendrycks et al., 2021; Singhal et al., 2023). We finetune language models on two kinds of domain text, one specific domain in medicine (medical textbook) and one general domain (Wikipedia). The finetuned models are then evaluated on the corresponding question answering tasks. The two scenarios are described below:

- **Anatomy.** Training corpus: the anatomy textbook from the MedQA dataset (Jin et al., 2021). QA data: 500 multiple choice quiz questions on core anatomy concepts in the textbook. Quiz questions are generated automatically with GPT-4 (OpenAI, 2023). The procedure to generate quiz questions, including prompts examples are described in Appendix A.1.2. QA performance is evaluated with standard 5-shot prompting.

- **SQuAD (closed-book).** Training corpus: Wikipedia excerpts from the SQuAD dataset (Rajpurkar et al., 2016). QA data: questions about facts in the Wikipedia excerpts, also from the SQuAD dataset. We turned the reading comprehension dataset of SQuAD into a closed-book QA, by first finetuning the model on the Wikipedia excerpts and then evaluate it on question answering without giving the excerpts. This closed-book QA setting was previously used to evaluate knowledge learning in language models (Hu et al., 2023). QA performance is measured using normalized F1 score.

  Evaluation details are described in Appendix C. Results on more datasets are given in Figure 7 in appendix.
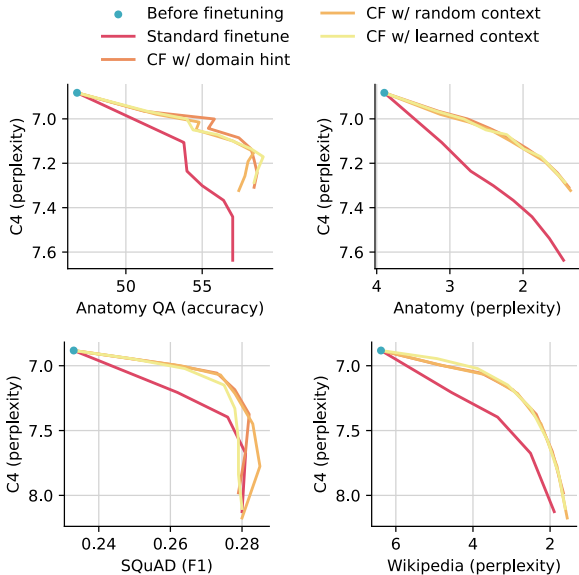
*Figure 5.* Performance-forgetting tradeoff curve of standard finetuning and conditional finetuning on Anatomy and SQuAD (closed-book). Conditional finetuning has less forgetting at similar levels of performance on downstream tasks, achieving significantly better tradeoff than standard finetuning.

We plot the stability-plasticity tradeoff curve (or "performance-forgetting" curve) of standard finetuning and conditional finetuning in Figure 5. The curves show forgetting as a function of learning. Curves on the top-right represent better tradeoff. To obtain the curve, we finetune the model on domain corpus for different numbers of epochs (from 1 to 8). We evaluate the model on the QA task as a measure of knowledge learning, and use perplexity on C4 as a measure of forgetting of existing information.

Figure 5 shows that conditional finetuning achieves significantly better tradeoff than standard finetuning. Furthermore, the tradeoff is relatively insensitive to the type of context. The maximum achievable performance on QA task is similar for the two training methods, while conditional finetuning has significantly less forgetting at each levels of learning. This shows that by selective learning, conditional finetuning poses less disruption to the information in pretrained model and achieves better stability-plasticity tradeoff.

## 5.3. Conditional Finetuning Reduces Forgetting and Improves Knowledge Learning in Continual Learning

When continually finetuning on multiple corpora, language models can continually learn new domain knowledge to integrate with existing knowledge. We show that conditional finetuning can also reduce forgetting of previously learned knowledge in a continual learning setting. It results

in improved cumulative knowledge learning over the entire course of continual finetuning.

Similar to the transfer learning scenario, we finetune language models on a medical domain and a general domain:

- **Medical textbooks** (13 corpora). Training corpus: 13 medical textbooks from the MedQA dataset (Jin et al., 2021). Details are described in Appendix A.1.1. We continually finetune the model on 13 textbooks in the following order: anatomy, biochemistry, cell biology, gynecology, histology, immunology, neurology, obstetrics, pathology, pediatrics, pharmacology, physiology, and psychiatry. QA data: 500 multiple choice quiz questions for each subject, similar to the transfer learning setting (Appendix A.1.2).

- **MRQA (closed-book**, 6 corpora). Training corpus: 6 corpora of Wikipedia pages, web text, and news articles from the 6 reading comprehension datasets in the MRQA benchmark (Fisch et al., 2019): SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019). We continually finetune the model on corpora provided by each dataset in that order. QA data: questions about facts in the corresponding corpus, also from each of the 6 datasets. Questions are turned into a closed-book format, and performance is evaluated using the evaluation protocol of SQuAD for consistency (Appendix C).

We use the Average Forgetting metric from Chaudhry et al. (2019) to evaluate forgetting in continual learning:

$$F_k = \frac{1}{k-1} \sum_{i=1}^{k-1} \max_{j \in \{1,...,k-1\}} a_{j,i} - a_{k,i} \qquad (6)$$

$F_k$ measures the average forgetting on previous QA tasks after training on the $k$-th corpus. $a_{k,i}$ is the accuracy on the $i$-th QA task after training on the $k$-th corpus.

We also use Cumulative Accuracy as a measure of the total knowledge learned over the course of continual finetuning:

$$C_k = \frac{1}{n} \sum_{i=1}^{n} a_{k,i} \qquad (7)$$

$C_k$ measures the average accuracy on all QA tasks after training on the $k$-th corpus[3]. $n$ is the total number of tasks.

We adapted the three types of context in Section 4 to use in continual learning: for **domain hint**, we use "*Following*

---

[3]Compared to Average Accuracy in Chaudhry et al. (2019), Cumulative Accuracy takes into account the initial performance of pretrained models and is more suited to measure learning on pretrained models.

| **Performance** | Average Forgetting ($\downarrow$) | | Cumulative Accuracy ($\uparrow$) | |
|---|---|---|---|---|
| | LLaMA-2 7B | 13B | 7B | 13B |
| *Medical textbooks* | | | | |
| Pretrained | - | - | 53.5 | 59.2 |
| Standard finetune | 2.5 | 2.6 | 60.3 | 65.3 |
| CF (w/ domain hint) | 2.3 | 2.5 | 60.5 | 65.2 |
| CF (w/ random) | 2.3 | 2.6 | 60.1 | 65.2 |
| CF (w/ learned) | **2.1** | **2.3** | **60.7** | **65.6** |
| *MRQA (closed-book)* | | | | |
| Pretrained | - | - | 0.390 | 0.431 |
| Standard finetune | 0.026 | 0.014 | 0.390 | 0.449 |
| CF (w/ random) | 0.022 | 0.014 | 0.382 | 0.450 |
| CF (w/ learned) | **0.019** | 0.015 | **0.395** | 0.450 |

*Table 3.* Continual learning performance of standard finetuning and conditional finetuning (at the last episode, $k=n$). CF = Conditional finetune. Conditional finetuning has less forgetting and achieves better cumulative accuracy. Learned context has better performance than other types of contexts. Note that the metric is F1 instead of accuracy for MRQA.

*is an excerpt from a [subject] textbook*" as context, where [subject] is replaced by the subject of each textbook. For ***random***, we use a different random UUID string for each corpus. For ***learned***, we learn soft prompts for each corpus.

Table 3 shows that conditional finetuning has less forgetting and achieves better cumulative accuracy than standard finetuning, especially with learned context. Figure 6 shows that conditional finetuning consistently has less forgetting of knowledge learned from previous corpora, over the entire course of continual learning.
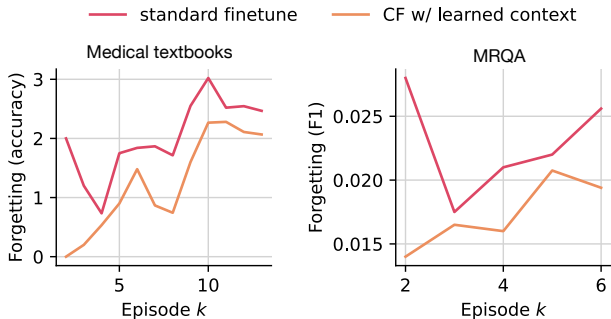


*Figure 6.* Average forgetting $F_k$ over the entire course of continual learning. Conditional finetuning has consistent less forgetting than standard finetuning. (LLaMA-2 7B)

In transfer learning where only a single context is used, the choice of context seems not affect performance much. In continual learning, the choice of context for each corpus can have a significant impact on performance. As we have shown in Section 4, the model learns a conditional topic prior $p(\text{topic}|a)$ in conditional finetuning. If a similar con-

text is later used in training on a different corpus, the context will activate the previous learned topic prior which provides wrong information about the current corpus. As a result, the model may need to unlearn the conditional topic prior of the previous corpus before learning from the new corpus, hindering the learning of new knowledge.

Therefore, to achieve the maximum "selectivity" in selective learning and reduce useless learning as much as possible, it is preferable to use contexts that provide information specific to each corpus when continually finetuning on multiple corpora. To verify this relationship between context choice and performance, we measure the similarity between contexts under the three context types. We calculate the average pairwise KL-divergence between conditional data distributions given context $a_i$ ($a_i$ is used on the $i$-th corpus):

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} D_{KL}(p(x|a_i)||p(x|a_j)) \quad (8)$$

The *random* contexts have a average KL-divergence of 0.006, the *domain hint* contexts 0.025, and the *learned* contexts 0.033. This shows that the random UUIDs are semantically very similar, while the learned contexts are the most dissimilar. The learned contexts also provide most specific information about each corpus as they are optimized for each corpus. The similarity between contexts is inversely correlated with performance observed in Table 3. The use of context specific to each corpus in conditional training leads to better selective learning and less forgetting.

## 6. Discussion

In this paper, we explored the effect of conditional finetuning, specifically on language modeling conditioned on a context. We found that by utilizing context to explain away corpus-level statistics, conditional training allows for selective learning from a corpus. It allows learning knowledge useful for downstream tasks while minimizing the learning of useless corpus statistics, such as topic biases. As a result, conditional training reduces side effects of domain finetuning and achieves less forgetting.

Selective learning gives finer-grained control over what the model learns in language modeling, and could be utilized for multiple purposes beyond discussed here. For example, it could help keeping the language model unbiased and better retain its general-purpose ability in continual finetuning and lifelong learning scenarios. For statistically biased corpora, conditional training may reduce the model's learning of the biases in sensitive attributes like race and gender, like in previous approaches (Clark et al., 2019; Sanh et al., 2021).

**Limitations.** We studied the effect of conditional finetuning on language models with a limited-size corpus due to

computational limitations. It might require further investigation to check whether the effect of conditional finetuning scales to large-scale training (e.g., billions of tokens).

In terms of evaluation, we evaluate model's performance with QA tasks on the main concepts in the corpus, which evaluates model's memorization and basic understanding of the concepts in the corpus. We did not verify whether the model can apply the learned knowledge in more complex reasoning scenarios, which seems challenging for current language models (Zhong et al., 2023; Berglund et al., 2024), and how conditional training affects such abilities.

Although conditional finetuning is observed to reduce forgetting, it is not proposed as a solution to catastrophic forgetting. Over-adaptation and unnecessary learning of corpus statistics is likely only one of the many factors that cause forgetting. We mainly aim to understand the effect of conditional learning in this paper, and leave the development of more effective methods to reduce forgetting to future work.

## Impact Statement

Our work mainly explores the effect of training language models with a context, and the results indicate that contexts can be used to exclude simple corpus statistics from learning by the model. This may be used to reduce social bias and improve the fairness of language models, because social bias is often simple bias caused by learning on a biased dataset. We have not yet foreseen any potential negative ethical consequences requiring particular discussion here.

## Acknowledgements

## References

Ba, L. J., Kiros, J. R., and Hinton, G. E. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL http://arxiv.org/abs/1607.06450.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. URL https://openreview.net/forum?id=GPKTIktA0k.

Biesialska, M., Biesialska, K., and Costa-jussà, M. R. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, pp. 6523–6541. International Committee on Computational

Linguistics, 2020. doi: 10.18653/V1/2020.COLING-MAIN.574. URL https://doi.org/10.18653/v1/2020.coling-main.574.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with A-GEM. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., and Yu, X. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pp. 7870–7881. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.634. URL https://doi.org/10.18653/v1/2020.emnlp-main.634.

Chrysostomou, G. and Aletras, N. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pp. 8189–8200. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.645. URL https://doi.org/10.18653/v1/2021.emnlp-main.645.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V. Y., Huang, Y., Dai, A. M., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/arXiv.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

Clark, C., Yatskar, M., and Zettlemoyer, L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4069–4082. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1418. URL https://aclanthology.org/D19-1418.

Clark, C., Yatskar, M., and Zettlemoyer, L. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 3031–3045. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020. FINDINGS-EMNLP.272. URL https://doi.org/10.18653/v1/2020.findings-emnlp.272.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., and Sui, Z. A survey on in-context learning. *CoRR*, abs/2301.00234, 2023. URL https://arxiv.org/abs/2301.00234.

Du, M., He, F., Zou, N., Tao, D., and Hu, X. Shortcut learning of large language models in natural language understanding. *Commun. ACM*, 67(1):110–120, 2024. doi: 10.1145/3596490. URL https://doi.org/10.1145/3596490.

Dunn, M., Sagun, L., Higgins, M., Güney, V. U., Cirik, V., and Cho, K. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179, 2017. URL http://arxiv.org/abs/1704.05179.

Feng, A., Zhang, X., and Song, X. Unrestricted attention may not be all you need-masked attention mechanism focuses better on relevant parts in aspect-based sentiment analysis. *IEEE Access*, 10:8518–8528, 2022. doi: 10.1109/ACCESS.2022.3142178. URL https://doi.org/10.1109/ACCESS.2022.3142178.

Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., and Chen, D. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019*, pp. 1–13. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-5801. URL https://doi.org/10.18653/v1/D19-5801.

Gao, L., Tow, J., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., McDonell, K., Muennighoff, N., Phang, J., Reynolds, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation. https://github.com/EleutherAI/lm-evaluation-harness, September 2021. URL https://doi.org/10.5281/zenodo.5371628.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2 (11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Gupta, K., Thérien, B., Ibrahim, A., Richter, M. L., Anthony, Q. G., Belilovsky, E., Rish, I., and Lesort, T. Continual pre-training of large language models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023. URL https://openreview.net/forum?id=pg7PUJe0Tl.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hu, N., Mitchell, E., Manning, C. D., and Finn, C. Meta-learning online adaptation of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pp. 4418–4432. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.268.

Jain, S., Salman, H., Khaddaj, A., Wong, E., Park, S. M., and Madry, A. A data-based perspective on transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pp. 3613–3622. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00352. URL https://doi.org/10.1109/CVPR52729.2023.00352.

Jang, J., Ye, S., Yang, S., Shin, J., Han, J., Kim, G., Choi, S. J., and Seo, M. Towards continual knowledge learning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=vfsRB5MImo9.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Jin, X., Zhang, D., Zhu, H., Xiao, W., Li, S., Wei, X., Arnold, A. O., and Ren, X. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pp. 4764–4780. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.351. URL https://doi.org/10.18653/v1/2022.naacl-main.351.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. 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, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Katharopoulos, A. and Fleuret, F. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2530–2539. PMLR, 2018. URL http://proceedings.mlr.press/v80/katharopoulos18a.html.

Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., and Liu, B. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=m_GDIItaI3o.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A. P., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL\_A\_00276. URL https://doi.org/10.1162/tacl_a_00276.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.243. URL https://doi.org/10.18653/v1/2021.emnlp-main.243.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V. V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving quantitative reasoning problems with language models. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., and Tang, J. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., and Roberts, A. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730, 2018. URL http://arxiv.org/abs/1806.08730.

OpenAI. Gpt-4 technical report. Technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training

language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012. URL https://doi.org/10.1016/j.neunet.2019.01.012.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264. URL http://aclweb.org/anthology/D16-1264.

Sanh, V., Wolf, T., Belinkov, Y., and Rush, A. M. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. URL https://openreview.net/forum?id=Hf3qXoiNkR.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., Datta, D., Chang, J., Jiang, M. T., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Févry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Shi, B., Darrell, T., and Wang, X. Top-down visual attention from analysis by synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pp. 2102–2112. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00209. URL https://doi.org/10.1109/CVPR52729.2023.00209.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S.,

Mansfield, P. A., Prakash, S., Green, B., Dominowska, E., y Arcas, B. A., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J. K., Webster, D. R., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617, 2023. doi: 10.48550/arXiv.2305.09617. URL https://doi.org/10.48550/arXiv.2305.09617.

Sun, T., Shao, Y., Qian, H., Huang, X., and Qiu, X. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20841–20855. PMLR, 2022. URL https://proceedings.mlr.press/v162/sun22e.html.

Thrun, S. Lifelong learning algorithms. In *Learning to Learn*, pp. 181–209. Springer, 1998. doi: 10.1007/978-1-4615-5529-2\_8. URL https://doi.org/10.1007/978-1-4615-5529-2_8.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b. doi: 10.48550/arXiv.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017*, pp. 191–200. Association for Computational Lin-

guistics, 2017. doi: 10.18653/V1/W17-2623. URL https://doi.org/10.18653/v1/w17-2623.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K., and Zhao, J. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-emnlp.167.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Xie, S. M., Santurkar, S., Ma, T., and Liang, P. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=uPSQv0leAu.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. R. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=5Xc1ecxO1h.

Zhang, X. and Wu, J. Dissecting learning and forgetting in language model finetuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. URL https://openreview.net/forum?id=tmsqb6WpLz.

Zhong, Z., Wu, Z., Manning, C. D., Potts, C., and Chen, D. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pp. 15686–15702. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.971.

# A. Data

## A.1. Medical domain

### A.1.1. CORPUS

We use the medical textbooks provided with the MedQA dataset (Jin et al., 2021) as a knowledge-rich corpus in the medical domain. To avoid varying the corpus size too much in continual learning setting, we use the 13 textbooks (1 subject each) that have size in the range of 1-10MB. Table 4 shows a statistics of the medical textbooks used. Number of tokens is measured with the tokenization scheme of LLaMA-2 (Touvron et al., 2023b) model.

### A.1.2. QA TASK

We use GPT-4 to generate multiple-choice quiz questions on each medical subject. Given a medical subject, we use the procedure to generate questions:

1. Split the textbook material of the subject into excerpts of 2048 tokens long, then randomly sample 50 excerpts (to help reduce the cost of GPT-4 usage). For each excerpt,

2. Instruct GPT-4 to generate 10 multiple-choice quiz questions examining the key concepts covered in the excerpt. The prompt is as follows:

    *Here is an excerpt from a subject textbook:*

    *⟨excerpt⟩*
    *{input}*
    *⟨/excerpt⟩*

    *Please write 10 multiple-choice quiz questions to examine whether a student remembers the key concepts from the above excerpt, after they studied the entire textbook.*

    *Requirements on content:*
    *- each question should have four choices, one choice must be definitely correct, the other three choices must be definitely wrong*
    *- the choices should be short and simple*
    *- each question should examine different key concepts in the material*
    *- provide enough context in the question so that it is answerable unambiguously, but do not refer to the particular excerpt, the figures, or the textbook*
    *- do not use negation (e.g., "not", "except") in the question, and do not use combination (e.g., "all of the above", "both A and B") in the choices*

    *Requirements on format:*
    *- please provide questions and answers in the following format:*
    *Question: ⟨question⟩*
    *A) ⟨choice 1⟩*
    *B) ⟨choice 2⟩*
    *C) ⟨choice 3⟩*
    *D) ⟨choice 4⟩*
    *Answer: ⟨the answer (a single letter)⟩*
    *- please directly give output without comments*

    where {subject} is replaced by the subject and {input} is replaced by the excerpt.

Table 5 shows examples of the generated questions for some subjects.

| Subject | Corpus length (tokens) | # Questions |
|---|---|---|
| Anatomy | 661K | 500 |
| Biochemistry | 404K | 500 |
| Cell biology | 1296K | 500 |
| Gynecology | 1768K | 500 |
| Histology | 841K | 500 |
| Immunology | 969K | 500 |
| Neurology | 2272K | 500 |
| Obstetrics | 2156K | 500 |
| Pathology | 1122K | 500 |
| Pediatrics | 842K | 500 |
| Pharmacology | 1467K | 500 |
| Physiology | 889K | 500 |
| Psychiatry | 821K | 500 |

*Table 4.* Statistics of the medical textbook corpus for each subject.

### A.2. General domain

We use the MRQA benchmark (Fisch et al., 2019) for evaluating knowledge learning on general domain. We extract all the documents from the 6 reading comprehension datasets as training corpora. The datasets are SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019). To balance the size of the corpora and reduce computation cost, we sample 1,000 questions and the associating documents from each dataset. The questions are turned into a closed-book format QA, and the documents are used as the training corpus for finetuning the language models.

## B. Training details

### Language model finetuning

We finetune the model with the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 3e-5. A linear learning rate decay is used with a warm-up of 10% of the total number of steps. We use a gradient clipping at 1.0. Batch size is set to 16. The maximum sequence length is set to 2048.

### Prompt tuning

To learn a soft prompt, we train the 10 embedding vectors on the training corpus for 3 epochs with a learning rate of 1e-1. The learning rate is chosen from a range search that minimizes loss. The training objective is conventional causal language modeling loss. The 10 embedding vectors have the same dimensionality as the language model's token embeddings. Prompt tuning is performed with the PEFT (Mangrulkar et al., 2022) library. The learned soft prompts are fixed when used as a context in conditional finetuning.

## C. Evaluation details

All QA tasks are evaluated using EleutherAI's Language Model Evaluation Harness framework (Gao et al., 2021).

The evaluation format for multiple choice-style QA tasks (anatomy, medical textbooks) follows the format of the MMLU (Hendrycks et al., 2021) benchmark.

The evaluation format for completion-style QA tasks (SQuAD, MRQA) follows the evaluation protocol of the SQuAD (Rajpurkar et al., 2016) benchmark, which uses normalized F1 scores as metric. Answers and groudtruths are normalized and have articles and punctuation removed before word-level F1 score is calculated. The protocol is often used in a combined evaluation on multiple QA tasks (McCann et al., 2018; Fisch et al., 2019).

All QA tasks are evaluated with 5-shot prompting.

For language modeling perplexity, when evaluating on C4, we randomly sampled 10,000 documents from the validation split of the English part of C4 corpus (C4/en) due to the large size of C4.
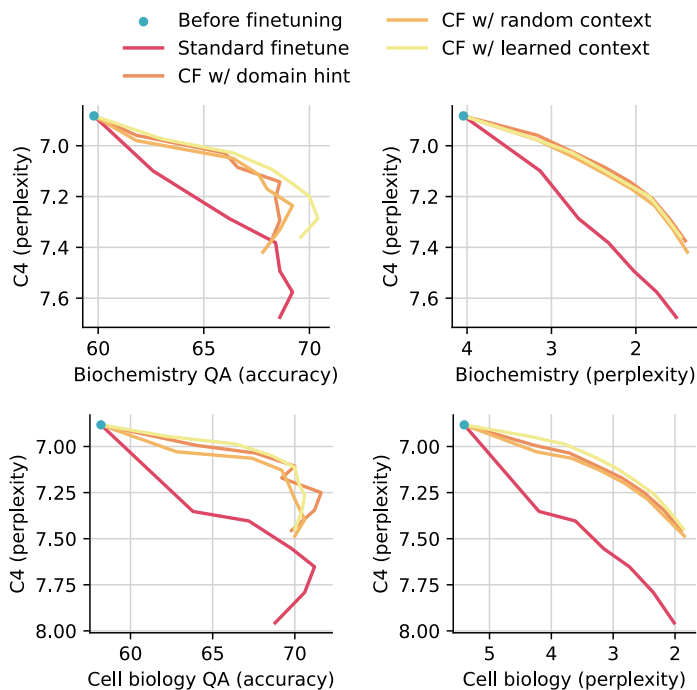


*Figure 7.* Performance-forgetting tradeoff curve of standard finetuning and conditional finetuning on Biochemistry and Cell biology, the second and third subjects in medical textbooks. Conditional finetuning has consistently less forgetting at similar levels of performance on downstream tasks, achieving significantly better tradeoff than standard finetuning.

| Subject |
| --- |
| *Anatomy* <br> Question: The inferior gluteal nerve innervates which of the following muscles? <br> A) Tensor fasciae latae <br> B) Gluteus medius <br> C) Gluteus maximus <br> D) Obturator internus <br> Answer: C |
| *Biochemistry* <br> Question: Which compound is an allosteric inhibitor of glutamate dehydrogenase (GDH)? <br> A) Adenosine triphosphate <br> B) Adenosine diphosphate <br> C) Guanosine diphosphate <br> D) Guanosine triphosphate <br> Answer: D |
| *Cell biology* <br> Question: In bacterial transcription, what helps the core enzyme break free from its interactions with promoter DNA? <br> A) The binding of ribonucleotides <br> B) Sigma factor reassociation <br> C) Transcription bubble contraction <br> D) Stress generated by scrunching <br> Answer: D |
| *Gyneacology* <br> Question: What is the recommended diagnostic step for premenarcheal patients with a pelvic mass? <br> A) MRI scan <br> B) Karyotype determination <br> C) Pelvic ultrasound <br> D) Hormone level testing <br> Answer: B |
| *Histology* <br> Question: In hepatocytes, where are lysosomes typically concentrated? <br> A) Near the bile canaliculus <br> B) Throughout the cytoplasm evenly <br> C) Inside the nucleus <br> D) At the cell periphery <br> Answer: A |

*Table 5.* Examples of the generated questions, for the first 5 subjects in medical textbooks.