

# LoRA Soups: Merging LoRAs for Practical Skill Composition Tasks

**Akshara Prabhakar\***

Department of Computer Science  
Princeton University

**Yuanzhi Li**

Microsoft Research

**Karthik Narasimhan**

Department of Computer Science  
Princeton University

**Sham Kakade, Eran Malach**

Kempner Institute  
Harvard University

**Samy Jelassi**

Center of Mathematical Sciences and Applications  
Harvard University

## Abstract

Low-Rank Adaptation (LoRA) is a popular technique for parameter-efficient fine-tuning of Large Language Models (LLMs). We study how different LoRA modules can be merged to achieve *skill composition*—testing the performance of the merged model on a target task that involves combining multiple skills, each skill coming from a single LoRA. This setup is favorable when it is difficult to obtain training data for the target task and when it can be decomposed into multiple skills. First, we identify practically occurring use-cases that can be studied under the realm of skill composition, e.g. solving hard math-word problems with code, creating a bot to answer questions on proprietary manuals or about domain-specialized corpora. Our main contribution is to show that concatenation of LoRAs (CAT), which optimally weights LoRAs that were individually trained on different skills, outperforms existing model- and data- merging techniques; for instance on math-word problems, CAT beats these methods by an average of 43% and 12% respectively. Thus, this paper advocates model merging as an efficient way to solve compositional tasks and underscores CAT as a simple, compute-friendly and effective procedure. To our knowledge, this is the first work demonstrating the superiority of model merging over data mixing for binary skill composition tasks.<sup>†</sup>

## 1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in conversational tasks and general-purpose applications, such as writing emails or answering common questions. However, these general purpose LLMs may have restricted performance on tasks where specific skills and knowledge is required. We primarily focus on skill composition tasks, that necessitate the integration of multiple skills.

Many industrial applications fit in this framework. Consider a company that manufactures ovens and is trying to design a chatbot to answer customer queries about its working and specifics. Directly using a frontier LLM (like gpt-4o) would fail since it lacks knowledge about the company’s product. The ideal solution here would be to design an instruction dataset consisting of question-answer pairs about this product and fine-tuning an LLM on it. However, such a data collection and annotation procedure is expensive. Another possible solution is to fine-tune an LLM on a collection of product manuals and then impart it chat abilities by further fine-tuning on an instruction-tuning dataset like

\*Corresponding author <akshblr555@gmail.com>

<sup>†</sup>Code and data are available at <https://github.com/aksh555/LoRA-Soups>.

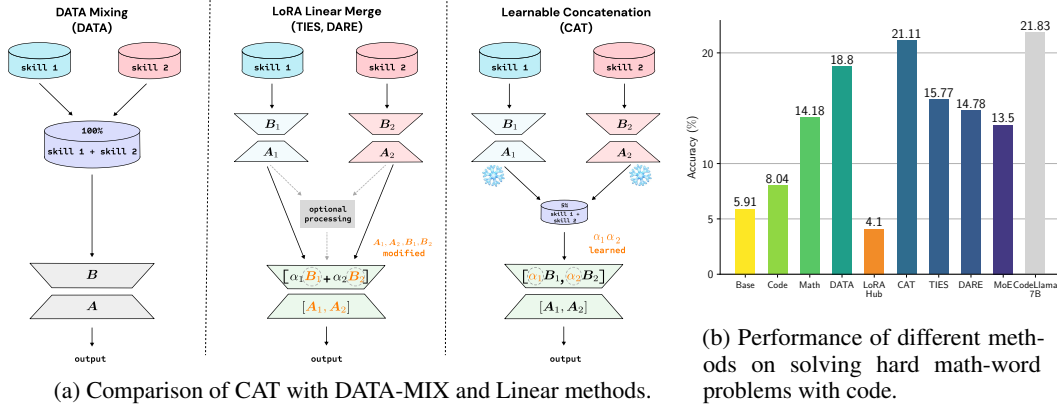


Figure 1: Method and performance overview of Learnable Concatenation (CAT).

Alpaca [55]. We refer to this method as DATA-MIX. Besides this approach being sequential, it suffers from catastrophic forgetting [28]. Whenever the company creates a new product, they need to redo fine-tuning on this data mixture or create a new question-answer dataset for the former method.

In this paper, we study *model merging* as an alternative approach. Given a model that is fine-tuned on the manuals and one that possesses question-answering capabilities, we optimally combine their weights to obtain a model that can answer product-specific questions. This approach is more efficient since we merge skill-specific fine-tuned models without any additional data collection or training from scratch. Among the multiple techniques to perform model merging, our framework specifically builds on LoRA [23] (a review of the LoRA method is in Appendix A), a fine-tuning technique that consists of adding a low-rank update to a few layers in the model. In this context, model merging consists of combining the LoRA weights from different models.

Prior works [19, 49, 70] have investigated LoRA merging in computer vision where each skill is a visual concept or style and the objective is image generation. On the other hand, natural language tasks are more challenging since identifying the skills needed for solving a task is not always clear. On natural language tasks, most prior works [5, 14, 37, 43, 62] merged LoRAs with the objective of multitask learning. In this setting, the individual LoRA modules are trained on (potentially) independent tasks and the merged model is tested on the original tasks. A successful model retains the skills of each individual LoRA. Differently, Huang et al. [24] devise LoraHub, a strategy to merge LoRAs for cross-task learning. By finetuning LoRAs on FLAN [35], they achieve performance equivalent to few-shot prompting on some Big-Bench Hard (BBH) tasks [54]. Closer to our work, Akiba et al. [2] merge specialized LMs in Japanese and in math [11] to solve word-math problems in Japanese [50]. Though this can be viewed as a skill composition task, they focus on studying only one such task, and their contribution is an evolutionary algorithm for merging models.

*Given LoRAs trained on specialized domains (biology, math, code, reading comprehension, question-answering), is it possible to merge them to effectively solve a new problem that requires a combination of these domains?*

We underline that most settings we consider are *out-of-domain* since the specialized LoRAs have been trained on datasets that are very different from the target task. To our knowledge, this is the first paper exhibiting model merging is superior to data mixing for binary skill composition problems. Our key contributions are summarized as follows:

- In Section 3, we define the *skill composition* problem and show some practical applications in domains spanning code, science, robustness, and in-house use cases.
- We introduce Learnable Concatenation (CAT), a LoRA merging technique that involves a simple weighted average of the encompassed skill LoRAs. While concatenating LoRAs as a technique already exists, our contribution lies in cheaply learning the weights layer-wise and applying this to compositional natural language tasks.
- In Section 5, we perform a comprehensive evaluation of several baselines and demonstrate that CAT achieves better binary skill composition than both existing merging methods and data mixing.

- In Section 6, we discuss some limitations of our approach. Our main limitation is that we focus on skill composition problems that require 2 skills and an interesting direction would be to extend this to more skills.

## 2 Related Work

**Merging methods.** The traditional approach to learning multiple skills/tasks simultaneously is joint training on a mixture of task datasets [7]. Over the years, several works have improved this multi-task learning approach [41, 53, 48, 20, 38, 9, 48]. As data collection for specialized tasks, and training large models from scratch get more expensive; coupled with the rapid expansion in the availability of well-trained open-source models – model merging has emerged as a convenient way of building powerful models from existing ones [17, 32]. The richly studied simplest way of merging by averaging model weights [57, 52, 16, 26] paved the way to linear weight averaging [60]. Some methods like Fisher Merging [39] and RegMean [27] need training data based pre-computations to measure individual parameter importance but these are highly memory and data intensive. Expanding on weight averaging, Task Arithmetic [25] involving the creation and combination of task vectors facilitated multi-task learning. While this weight interpolation was heavily used for merging image generation models, recent methods like TIES [67] and DARE [2] reset redundant parameters, resolve sign conflicts, and exclusively merge parameters that exhibit sign-consistency, and SLERP [59] by spherical linear interpolation build upon this for language models. In all these methods, the coefficients governing the model merging are determined by trial-error; while works in the vision domain demonstrate the pivotal role played by these coefficients [68]. In contrast to the cumbersome grid-searching, our method CAT learns these coefficients layer-wise with access to very few examples in the natural language domain.

**LoRA merging methods.** PEM composition [73] adopts the task arithmetic framework to incorporate the merging of LoRAs. Recently, the vision community witnessed the widespread application LoRAs [5, 14, 37, 43, 62, 75, 69] as an effective approach to multi-task learning and composing styles and subjects [49]. Many of these utilize Mixture of Experts (MoE) [5, 14, 37, 62] based architectures having input-dependent learnable routers. These models have been primarily used in the context of multitask learning. Apart from multi-task learning [25, 33, 66, 62, 14, 68], merging models have served various use cases – ensembling to improve generalization on the target task [26, 3, 21], federated learning [40], model compression [34]. In this work, we study LoRA model merging for *skill composition* tasks.

## 3 Setting

### 3.1 Skill composition

Most of the downstream tasks used to evaluate LLMs require mastering multiple skills to be solved. Skill here refers to specific capabilities that the LLM needs for customization to downstream use cases. These skills can be acquired from knowledge source like textbooks and manuals or from foundational datasets designed for arithmetic, coding, etc. For instance, achieving high score in the GSM8k benchmark [11] requires good commonsense reasoning and arithmetic skills. In this paper, we focus on downstream tasks where composition can be ensured and isolated, and we mainly focus on tasks that require two skills. Skill composition is challenging because, not only does the model need to “know” the different skills, but it also needs to understand the appropriate context for applying each skill. We now present some skill composition examples of practical interest and Table 1 summarizes them.

**Hard math-word problems.** Prior works noticed that when fine-tuning a large language model on GSM-8k [11], the resulting model performs poorly on GSM-Hard [15], a dataset that gathers similar problems but with more complex arithmetic operations. The program-aided approach by Gao et al. [15] is one solution to address this issue: the LLM reads the GSM problem and generates programs as the intermediate reasoning steps, but offloads the solution

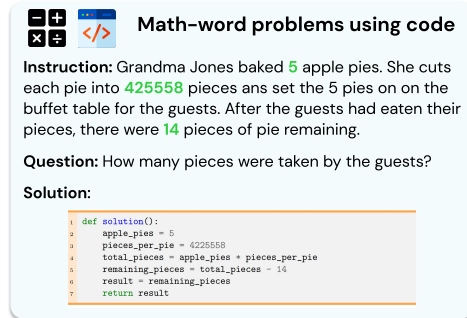


Figure 2: An example of solving hard word-math problem using code.

step to a Python interpreter. In this case, the model must excel in both mathematical skills to first reason about the problem and also coding skills to translate its reasoning to code. Therefore, in order to improve the accuracy on GSM-hard, we set the first skill to be MetaMathQA [71] – an augmented version of the GSM-8k dataset and the second skill to be Code-Alpaca [8].

**QABot on proprietary manuals.** Chat models are language models that aim to facilitate conversational interactions. Similarly to LLMs, they are first pre-trained on large language datasets. Then, they are finetuned on instruction-tuning datasets [10, 12, 30, 55, 63, 64, 76]. While general-purpose chatbots are useful, an institution or corporation may desire to have a *specialized* question-answering bot/assistant that addresses domain-specific questions. Examples of this case may be a university that wants a bot that answers questions about quantum physics or a refrigerator retailer that wants a bot to answer questions about their last device. The aforementioned approach may fall short since when fine-tuning a chat model on a specialized document, it may lose its conversational abilities. Another solution would be to obtain an instruction-tuning dataset that covers the specialized material but this may be too costly. On the other hand, model merging seems to be a convenient solution to address this problem: we train one LoRA model on a general instruction-tuning dataset and another one on the specialized document. Therefore, we set the first skill to be the Alpaca dataset [55] and study two cases where the second skill is: (1) a molecular biology university textbook [4] from <https://bio.libretexts.org/> and (2) a manual of the "Dungeons and Dragons" game [22].

**Reading comprehension on technical documents.** Medical reports and law contracts are very challenging to read for non-specialist users: they are usually very long and involve a myriad of technical jargon. Having a language model that can read these documents and answer questions would be very useful. For this reason, we train one LoRA model on SQuAD [45] to acquire the reading comprehension skill and another LoRA on open-source biology textbooks [18, 29, 44, 51] collected from <https://libretexts.org>. We test our models on the BioASQ-QA dataset [31], a reading comprehension dataset on biomedical engineering.

**DUNGEONS & DRAGONS Closed-Book question answering bot**  
**Question:** What are the proficiencies of a Barbarian in terms of armor, weapons, and saving throws?  
**Solution:** A Barbarian is proficient in light armor, medium armor, shields, simple weapons, martial weapons, and has saving throws for Strength and Constitution.

Figure 3: An example of question-answering task based on *Dungeons & Dragons* game manual.

**Open-Book reading comprehension**  
**Context:** Recent functional genetic studies have pointed out that LKB1 plays a physiological role in **controlling the Wnt-signaling pathway** and activation of the pathway as a consequence of LKB1 haploinsufficiency might be responsible for the development of hamartomatous polyps. PJS polyps from all patients showed generalized membrane and cytoplasmic localizations of beta-catenin along the mucosal endothelium.  
**Question:** What is the physiological role of LKB1 involved in Peutz-Jeghers syndrome?  
**Solution:** controlling the Wnt-signaling pathway

Figure 4: An example of the reading comprehension task in the biomedical domain.

### 3.2 Merging methods

Figure 1a details the various merging methods in the literature and the proposed CAT method.

**Data mixing (DATA-MIX).** The naive approach for solving these tasks is to train on a mixture of datasets containing different skills. This method does not merge models but merges datasets. We fine-tune a *single* model with LoRA weights  $(A, B)$  where  $A, B \in \mathbb{R}^{d \times kr}$  on the concatenation of datasets and thus simultaneously teach the  $k$  skills to the model. Data mixing is a very expensive approach in practice. Indeed, whenever we want to add a new skill to the model, we need to retrain the model on an updated mixture of datasets. Model merging offers a solution to this problem. We train one model per skill and then merge all of them to solve the compositional task. We set the rank to  $kr$  (and not  $r$ ) in order to ensure a fair comparison with the other merging methods.

Next, we review the different merging schemes and report their skill composition performance in the next sections. We distinguish two main classes of LoRA merging techniques: the concatenation of LoRAs and the linear merging of LoRAs. These two methods differ in whether we linearly combine the parameters or the LoRA updates.

**Concatenation of LoRAs.** This method consists of a linear combination of the LoRA updates i.e.  $\Delta W^l = \alpha_1^l B_1 A_1^\top + \alpha_2^l B_2 A_2^\top$ , where  $\alpha_1^l, \alpha_2^l \in [0, 1]$  are merging coefficients for layer  $l$ . We refer to this method as concatenation of LoRAs. In what follows, we study three variants of this method that differ by their merging coefficients definition:

- **Learnable concatenation (CAT)** (introduced in this paper): in this variant, we set  $\alpha_1^l, \alpha_2^l$  as trainable parameters. This distinguishes us from other methods like TIES, DARE, LoRA Hub that learn static values for every layer  $l$  of the network. Once the LoRA modules are separately trained, we add a step where we only train the merging coefficients on a small mixture of the datasets.
- **Mixture of Experts (MoE)** [5, 14, 37, 43, 62]: the main difference of this method compared to the previous ones is that the merging coefficients are *input-dependent*. Indeed, we have a trainable router parameter  $W_r^l \in \mathbb{R}^{d \times 2}$  which computes the logits  $h^l(x) = W_r^{l\top} x$ . Then, the merging coefficients are defined as:

$$\alpha_1^l = \frac{e^{h^l(x)_1}}{e^{h^l(x)_1} + e^{h^l(x)_2}}, \alpha_2^l = \frac{e^{h^l(x)_2}}{e^{h^l(x)_1} + e^{h^l(x)_2}}$$

**Linear merging of LoRAs.** This method consists of a linear combination of the LoRA parameters i.e.  $\Delta W^l = (\alpha_1^l B_1 + \alpha_2^l B_2)(\alpha_1^l A_1 + \alpha_2^l A_2)^\top$ . **We note that all these methods use the same static weights  $(\alpha_1, \alpha_2)$  for every layer.** Compared to the concatenation of LoRAs, this method involves additional cross-terms. In what follows, we study three variants of this method that apply a series of preprocessing steps on the LoRA parameters before applying the update.

- **TIES** [67]: we prune the smallest values of  $(A_k, B_k)$  for  $k \in \{1, 2\}$  and retain the top values based on the specified density fraction  $\lambda \in [0, 1]$ . Then, we calculate the majority sign mask from the pruned LoRA weights by summing all their parameter values and storing the sign of this sum. Lastly, the LoRA weights are multiplied by weights  $(\alpha_1, \alpha_2)$ . Finally, we apply the linear merging update based on the stored majority sign. Here,  $(\lambda, \alpha_1, \alpha_2)$  are hyper-parameters.
- **DARE** [72]: we first randomly prune the values of the LoRA parameters  $(A_k, B_k)$  for  $k \in \{1, 2\}$  based on a density  $\lambda \in [0, 1]$ . Then, we rescale the pruned LoRA weights by  $1/\lambda$ . Finally, we apply the linear merging update. Here,  $(\lambda, \alpha_1, \alpha_2)$  are hyper-parameters.
- **LoRA Hub** [24]: this method was primarily proposed to select and assign weights to the constituent LoRA modules that would help solve an unseen test task. Here  $(\alpha_1, \alpha_2)$  are learned from a few (5) examples from the target task in a gradient-free manner.

Most of these aforementioned methods have been introduced in the literature for solving the multitask problem, i.e. to perform well simultaneously on  $k$  skills, when tested independently. In this paper, we study how these methods perform in *skill composition*.

## 4 Experimental details

**Skill Finetuning.** Our base model is a Llama-7b [56] and set the LoRA rank  $r = 32$ , LoRA alpha = 64, LoRA dropout = 0.05 and the target modules to be {"q\_proj", "v\_proj", "k\_proj", "up\_proj", "down\_proj"}. We finetune the individual LoRAs using AdamW [36] for 3 epochs, with a learning rate  $3e-4$ , 100 warmup steps, a linear decaying schedule, batch size in {4, 8} and gradient accumulation 4. We set the precision to float16. Our loss is the standard next-token prediction loss. When finetuning on the instruction datasets, we mask the instruction so that we only penalize wrong token predictions at the level of answers. When finetuning on textbooks, we do not apply any masking. We followed the implementation of *mergo* [32] for MoE which is based on recent MoE for LoRAs [14, 6].

**DATA-MIX.** For DATA-MIX, we train LoRA modules with a rank  $2r = 64$ , to ensure an equal number of trainable parameters with the merging methods. For LoRA Hub, we used at least 5 few-shot examples for learning the weights. Training a model on a mixture of datasets in Subsection 5.3, Subsection 5.2 is not straightforward, as different examples have different masking schemes, which makes standard data mixing fail. When training with alternating between one batch of each dataset



Task	Skills Composed	Test Dataset	Evaluation Metric
Hard math-word	{math, code}	GSM-Hard [15] (1319)	Exec. Accuracy
Question-answering	{biology/game, closed-book QA}	Open-source book/manual §B.2 (95)	Accuracy (GPT-4)
Reading comprehension	{biomedical, open-book QA}	BioASQ-QA [31] (200)	Elo Rating (GPT-4)
Prompt robustness	{ $prompt_i, prompt_j$ }	SuperNaturalInstructions [58] (417)	Accuracy

Table 1: Rundown of our skill-composition tasks, source of test datasets, and evaluation criterion.

type we observed poor performance. Hence, we perform continual training by fine-tuning for 3 epochs on the chapter/manual, merge the weights with the pre-trained model weights, followed by 1 epoch of fine-tuning on the instruction-following dataset (SQuAD/Alpaca) similar to [61].

**CAT.** We freeze the trained LoRA skill modules and train  $\alpha_1^l, \alpha_2^l$  on a dataset made by selecting the minimum of 5% of the data points from both skill 1 and skill 2. This additional step only runs for 1 epoch with a learning rate of  $1e-4$ . The best model is chosen as the one that does best on the validation set.

**Hyperparameters.** The  $(\lambda, \alpha_1, \alpha_2)$  hyperparameter values for TIES, DARE are chosen by doing a sweep over  $\lambda \in [0, 1], \alpha_1 \in [1, 2], \alpha_2 \in [1, 2]$  in increments of 0.2 and we report the best results. At inference time, we generate answers in an autoregressive fashion setting temperature to 0.01, max\_new\_tokens to 200, with nucleus sampling probability top\_p as 0.95.

Table 1 summarizes our skill composition tasks, dataset sources, and evaluation criterion. Further training details and compute used are presented in Subsection B.1.

## 5 Experiments

### 5.1 Super-linear improvement when solving GSM-Hard with code

**Evaluation setup.** We evaluate the ability to solve hard math-word problems in GSM-Hard using code. The model takes in a math word problem, mathematically reasons, and then outputs a corresponding Python function whose return value is the answer to the problem. We run this solution in the Python interpreter and use it to test. This setup is motivated by the fact that while finetuning a single LoRA on MetaMathQA yields 55% accuracy on the standard GSM-8k, the performance on GSM-Hard is much lower at 14.18%. Though GSM-Hard and GSM-8k problems share the same structure, they differ in the more complex arithmetic operations involved in GSM-Hard, due to which access to a calculator-like tool aids performance.

**Baselines.** Base (Llama-7b with 8-shot PAL); Skill LoRAs: Math (trained on MetaMath) with 8-shot PAL, Code (trained on Code Alpaca). DATA-MIX (trained on [MetaMath; Code Alpaca]); LoRA Merging: TIES, DARE, MoE.

**Results.** Figure 1b illustrates that finetuning on the concatenation of MetaMathQA and Code-Alpaca – which corresponds to the DATA strategy is effective since the accuracy increases by 32% over Math (from 14.18%  $\rightarrow$  18.8%) and the model effectively exploits the synergies between natural and programming language [65]. However, we note that DATA does not always conform to outputting code despite 8-shot PAL prompting, while model merging methods do so  $\geq 99\%$  times.

Regarding model merging, Yadav et al. [67] and Yu et al. [72] report that TIES and DARE are worse than DATA in the multitask setting. We observe that this finding also holds in the skill composition of math and code. Lastly, CAT is the best method on GSM-Hard; Figure 5 shows a qualitative

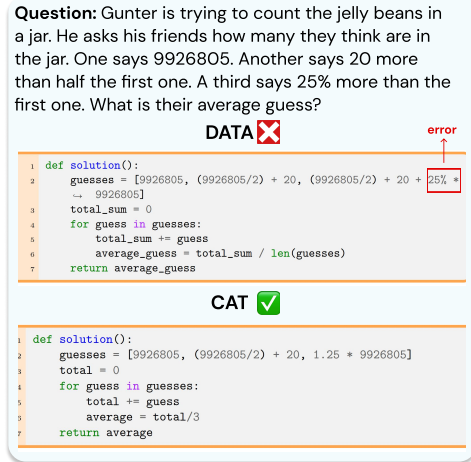


Figure 5: CAT vs. DATA solving a GSM-Hard problem. DATA makes frequent coding errors.

example comparing CAT to the next best method DATA. Additionally, despite being fine-tuned on a relatively much smaller general code dataset i.e. Code Alpaca, it matches the performance of Python specialized Code Llama - Python 7b [46], which highlights the effectiveness of linearly combining LoRA updates.

**Super-linear improvement.** Table 2 demonstrates that with CAT, we achieve a super-linear improvement on GSM-Hard. The performance improvement when finetuning on both math and code with respect to the base model (257%) is superior to the sum of the improvements on code only (36%) and math only (140%). We note that such super-linear improvement essentially proves that the merged model is

	No Code	Code
<b>No Math</b>	5.91%	8.04% (+36%)
<b>Math</b>	14.18% (+140%)	21.11% (+257%)

Table 2: Super-linear improvement with CAT.

able to leverage the capabilities of the individual models when solving some of the problems. Indeed, if Math solves 8% more problems than the base model, and Code solves 2% more problems, then the total number of problems solved by both models is at most 10% more than the original model, with a total of 16% solved problems<sup>‡</sup>. The merged model, however, solves 21% of the problems, meaning that at least 5% of the problems it solves are not solved by either of the individual models. The solution to these problems must therefore arise from combining the knowledge of both models.

## 5.2 Building specialized question-answering bots (QABots)

**Evaluation setup.** We test the closed-book question-answering capability (i.e. no access to the book/manual about which questions are asked) by considering two settings: (1) simple – university biology textbook chapter, (2) hard – nuanced *Dungeons & Dragons* game manual. Details on the preparation of the dataset and judging are discussed in Subsection B.2, Subsection C.2. In contrast to Subsection 5.3, this is closed-book QA as we have access to the context during inference.

**Baselines.** Skill LoRAs: Biology/Manual (trained on textbook/manual), Alpaca (trained on Alpaca). Apart from the usual LoRA merging baselines, we include two additional relevant ones: retrieval using langchain based RetrievalQAChain retrieval using (1) Llama-7b and (2) Llama2-7b-chat models. Here the documents are embedded using sentence transformer sentence-transformers/all-MiniLM-L6-v2 and stored in a FAISS vector store. These are indeed open-book but are included to benchmark merging methods and get an upper bound.

**Results.** Figure 6 shows the performance on the harder game setting. We observe that CAT beats most merging and data mixing but we note the scope for improvement compared to retrieval methods. Here as well, we note the greater need for having the instruction-following skill than the game knowledge due to Alpaca’s superior performance; as well as from Llama-chat beating base Llama. Again this highlights how composing the two skills boosts performance. The results on the simpler biology chapter setting are provided in (Subsection C.2, Figure 10).

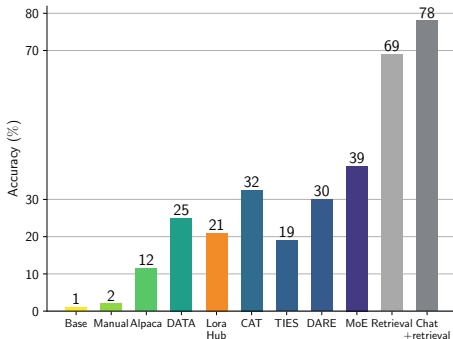


Figure 6: Performance on closed-book game QA.

## 5.3 Reading comprehension on technical documents

**Evaluation setup.** We test the open-book question-answering ability (i.e. context is accessible to the model) on BioASQ by choosing the “factoid” subset of questions to align with the format of questions seen in SQuAD and restricting to examples whose “Context” field is less than 500 tokens. Indeed, we observed that the model suffers from length generalization: when the context length is much larger than the one seen during training, the model does not output any answer. Since the ground truth answers in BioASQ are very long and sometimes contain more details than what is provided in the context, we observe low results in exact matching and F1 scores (see Subsection C.3

<sup>‡</sup>We make the reasonable assumption that the finetuned models are solving new problems, but not degrading the performance on the problems already solved by the base model.

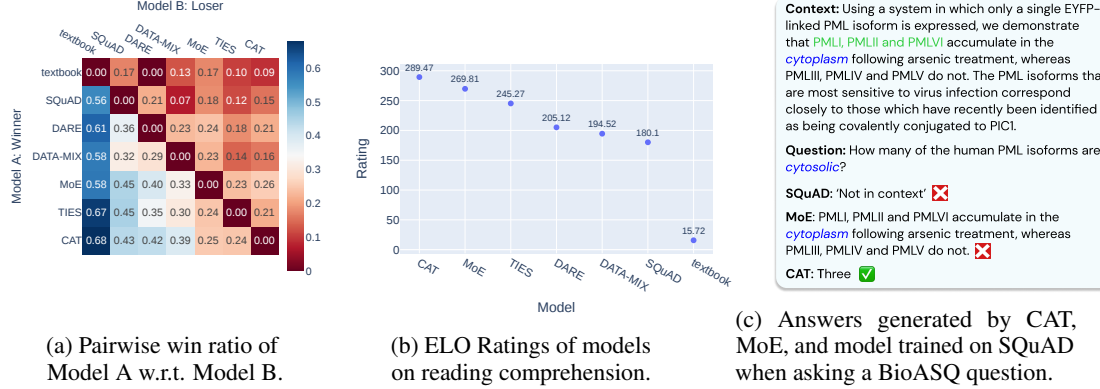


Figure 7: Quantitative and qualitative results on reading comprehension task.

Table 3). To alleviate this issue, we use GPT-4 as a judge [74]: given the answers generated by a pair of models, we ask it to score the two in terms of relative correctness to the gold reference answer. This gives a table that records all the pairs of methods, the number of ties, and wins by each model. We report the pairwise win fractions (the ratio of a model’s wins over 200 questions) and the ELO rating [13] (see Subsection C.3 for prompts and more details).

**Results.** Figure 7 reports the pairwise win fractions and ELO ratings obtained by the different methods on BioASQ. Again, we observe that CAT is the best method both in terms of pairwise win fraction and ELO ratings. Evidently, the question-answering skill is more useful than having domain knowledge as SQuAD fails 6× better compared to textbook. From Figure 7c we see the lack of biomedical knowledge hurting SQuAD; how CAT gives well-formed answers compared to MoE which is indeed able to understand the context but just copies relevant text. The other methods are worse and the merging methods beat DATA-MIX. As in the previous experiment, we use continual learning for the DATA-MIX experiment.

#### 5.4 LoRA merging improves robustness to prompt format changes

**Problem.** Recently, a few works [42, 47] reported the sensitivity of language models to prompt formatting. These works report performance differences of up to 76 accuracy points due to subtle changes in prompt formatting when using LLaMA-2-13B [56]. This sensitivity remains when increasing model size, the number of few-shot examples, or performing instruction tuning. We investigate whether model or data mixing is a solution to this problem. Therefore, we set the first skill to be the dataset with format  $i$  and the second skill to be the dataset with format  $j$  and evaluate the merging methods on a dataset with format  $k$ , where  $i, j, k \in \{1, \dots, 10\}$  and  $i \neq j \neq k$ .

**Evaluation setup.** We choose task\_131 of counting the number of words having a particular letter in the given sentence from Super-NaturalInstructions [58]. Following the grammar over *descriptors*, *separators*, and *spaces* defined in [47], we sample 7 prompt formats (see Figure 8 for some examples of format variations – as simple as spacing, casing). The *descriptors* refer to the format of {Sentence, Answer}, *separators* refer to the separator between the *descriptor* and the example, and *space* referees to the dividing character between the *descriptors*. We train on 6000 examples and test on 417.

**Results.** Figure 11 reports the performance of each model when evaluating on 4 out-of-distribution formats (Formats 3, 7, 8 and 10). Figures 11a, 11b and 11c respectively display the single and merged model when trained on Formats 1&4, Formats 1&2 and Formats 2&4. A well-performing merged model should perform at least as well as the individual models i.e. the models trained on a single

<b>Format 1</b>
Sentence: 'a person working alone in a large kitchen'.
How many words contain the letter 'o' in the sentence?
Answer:
<b>Format 2</b>
SENTENCE- 'a person working alone in a large kitchen'.
How many words contain the letter 'o' in the sentence? ANSWER-
<b>Format 3</b>
SENTENCE: 'a person working alone in a large kitchen'.
How many words contain the letter 'o' in the sentence? <sep> ANSWER
<b>Format 4</b>
SENTENCE\n
'a person working alone in a large kitchen'.
How many words contain the letter 'o' in the sentence?
ANSWER\n

Figure 8: Different prompt formats – varying *descriptors*, *separators*, and *spaces* for the same task example.



format. Remarkably, data mixing performs poorly, even worse than the individual models. Regarding the merged models, the results are variable. When finetuning on Formats 1&4 (Figure 11a), the performance of CAT does not vary across target formats and remains high. TIES performance is also high contrary to DARE. On Formats 1&2 (Figure 11b), TIES is the best model since it performs as the single model when evaluated on Formats 3 and 10. CAT and DARE perform worse and only obtain a decent performance when evaluated on Format 3. Lastly, on Formats 2 & 4, merging fails since all the models perform poorly. Thus, we show that model merging is an approach to attaining robustness to prompt formatting changes.

## 5.5 Ablations

**Learning the weights of CAT.** We analyze the impact of learning the weights to be assigned to each skill. As discussed in Section 4, we learn the merging coefficients. In this section, we demonstrate that this learning entails a significant performance improvement. Figure 9 shows how that “learned” CAT beats “static” CAT in the GSM-Hard experiment and the QAbot experiment (on the game manual). In this case, we simply average – set  $\alpha_1^l, \alpha_2^l = 0.5$ .

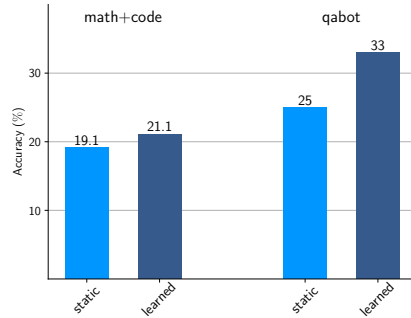


Figure 9: Performance of the learned vs static approach in CAT.

## 6 Limitations & Future Directions

Our work is a first step towards using model merging for solving skill composition tasks. Our study is limited to settings where the number of needed skills  $k = 2$ . Extension to a higher number of skills is challenging for two reasons: (1) Identifying practical settings that need a large number of skills that can be clearly identified is difficult. (2) Even if we find such settings, it remains unclear whether LoRA merging is effective. To test a hypothetical situation, we consider the prompt robustness experiment from Subsection 5.4 and report the performance of the different baselines when trained on 3 different prompt formats. This time, we observe that all the LoRA merging baselines perform worse than DATA-MIX (refer Subsection C.4 Figure 12). This suggests that when the number of skills is large, DATA-MIX is still the best strategy for skill composition. Improving current model merging schemes is an important future direction for getting an efficient composition of many skills.

## 7 Conclusion

We conclude that when obtaining training data is challenging, decomposing a task into its underlying skills and model merging is a promising approach. We demonstrate several practical use cases that can be treated as binary skill composition problems. We show that methods based on the concatenation of LoRAs achieve the best skill composition and outperform the popular data mixing method. Another highlight of our analysis is the super-linear modularity results from Subsection 5.1 which shows that the model is effectively able to compose the math and code skills to solve new problems. As mentioned in Section 6, an exciting future direction is to design merging methods that would enable to composition of more than two skills. This would give an interesting alternative to the current paradigm where we train large-scale models on curated datasets; instead we could merge specialized LoRAs to obtain an equivalently good model.

## References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] T. Akiba, M. Shing, Y. Tang, Q. Sun, and D. Ha. Evolutionary optimization of model merging recipes, 2024.

- [3] D. Arpit, H. Wang, Y. Zhou, and C. Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022.
- [4] G. Bergtrom. Basic cell and molecular biology 5e: What we know and how we find out. 2022.
- [5] E. L. Buehler and M. J. Buehler. X-lora: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and design. *arXiv preprint arXiv:2402.07148*, 2024.
- [6] E. L. Buehler and M. J. Buehler. X-lora: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design, 2024.
- [7] R. Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.
- [8] S. Chaudhary. Code alpaca: An instruction-following llama model for code generation. <https://github.com/sahil280114/codealpaca>, 2023.
- [9] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.
- [10] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [11] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [12] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm. *Company Blog of Databricks*, 2023.
- [13] A. E. Elo and S. Sloan. The rating of chessplayers: Past and present. (*No Title*), 1978.
- [14] W. Feng, C. Hao, Y. Zhang, Y. Han, and H. Wang. Mixture-of-loras: An efficient multitask tuning for large language models. *arXiv preprint arXiv:2403.03432*, 2024.
- [15] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR, 2023.
- [16] T. Garipov, P. Izmailov, D. Podoprikin, D. P. Vetrov, and A. G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.
- [17] C. Goddard, S. Siriwardhana, M. Ehghaghi, L. Meyers, V. Karpukhin, B. Benedict, M. McQuade, and J. Solawetz. Arcee’s mergekit: A toolkit for merging large language models, 2024.
- [18] M. Grewal and S. Wakim. Human biology. 2023.
- [19] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [20] P. Guo, C.-Y. Lee, and D. Ulbricht. Learning to branch for multi-task learning. In *International conference on machine learning*, pages 3854–3863. PMLR, 2020.
- [21] V. Gupta, S. A. Serrano, and D. DeCoste. Stochastic weight averaging in parallel: Large-batch training that generalizes well. *arXiv preprint arXiv:2001.02312*, 2020.
- [22] G. Gygax and D. Arneson. *dungeons & dragons*, volume 19. Tactical Studies Rules Lake Geneva, WI, 1974.

- [23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [24] C. Huang, Q. Liu, B. Y. Lin, T. Pang, C. Du, and M. Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.
- [25] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.
- [26] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.
- [27] X. Jin, X. Ren, D. Preotiuc-Pietro, and P. Cheng. Dataless knowledge fusion by merging weights of language models, 2023.
- [28] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1611835114>.
- [29] M. Klymkowsky, M. Cooper, and O. T. Library. *Biofundamentals 2.0*. Open textbook library. Michael Klymkowsky, Melanie Cooper. URL <https://books.google.com/books?id=q837wQEACAAJ>.
- [30] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.
- [31] A. Krithara, A. Nentidis, K. Bougiatiotis, and G. Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.
- [32] Leroo-AI. *Leeroo-AI/mergoo*. 2024. URL <https://github.com/Leeroo-AI/mergoo>.
- [33] M. Li, S. Gururangan, T. Dettmers, M. Lewis, T. Althoff, N. A. Smith, and L. Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*, 2022.
- [34] P. Li, Z. Zhang, P. Yadav, Y.-L. Sung, Y. Cheng, M. Bansal, and T. Chen. Merge, then compress: Demystify efficient smoe with hints from its routing policy. *arXiv preprint arXiv:2310.01334*, 2023.
- [35] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023.
- [36] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [37] T. Luo, J. Lei, F. Lei, W. Liu, S. He, J. Zhao, and K. Liu. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models, 2024.
- [38] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018.
- [39] M. S. Matena and C. A. Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.
- [40] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

- [41] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016.
- [42] M. Mizrahi, G. Kaplan, D. Malkin, R. Dror, D. Shahaf, and G. Stanovsky. State of what art? a call for multi-prompt llm evaluation, 2024.
- [43] M. Muqeeth, H. Liu, Y. Liu, and C. Raffel. Learning to route among specialized experts for zero-shot generalization, 2024.
- [44] A. Nelson and K. Greene. Medical terminology for healthcare professions. 2021.
- [45] P. Rajpurkar, R. Jia, and P. Liang. Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.
- [46] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- [47] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=RIu5lyNXjT>.
- [48] O. Sener and V. Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.
- [49] V. Shah, N. Ruiz, F. Cole, E. Lu, S. Lazebnik, Y. Li, and V. Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023.
- [50] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.
- [51] V. Singh. *Genetics*. Elsevier Health Sciences, 2014.
- [52] J. Smith and M. Gashler. An investigation of how neural networks learn from the experiences of peers through periodic weight averaging. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 731–736. IEEE, 2017.
- [53] X. Sun, R. Panda, R. Feris, and K. Saenko. Adashare: Learning what to share for efficient deep multi-task learning. *Advances in Neural Information Processing Systems*, 33:8728–8740, 2020.
- [54] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- [55] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- [56] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [57] J. Utans. Weight averaging for neural networks and local resampling schemes. In *Proc. AAAI-96 Workshop on Integrating Multiple Learned Models*. AAAI Press, pages 133–138. Citeseer, 1996.
- [58] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022.
- [59] T. White. Sampling generative networks, 2017. URL <https://openreview.net/forum?id=SypU8101e>.

- [60] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.
- [61] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.
- [62] X. Wu, S. Huang, and F. Wei. Mole: Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*, 2023.
- [63] C. Xu, D. Guo, N. Duan, and J. McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.
- [64] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.
- [65] Y. Xu, H. Su, C. Xing, B. Mi, Q. Liu, W. Shi, B. Hui, F. Zhou, Y. Liu, T. Xie, et al. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830*, 2023.
- [66] P. Yadav, D. Tam, L. Choshen, C. Raffel, and M. Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=xtaX3WyCj1>.
- [67] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [68] E. Yang, Z. Wang, L. Shen, S. Liu, G. Guo, X. Wang, and D. Tao. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*, 2023.
- [69] Y. Yang, W. Wang, L. Peng, C. Song, Y. Chen, H. Li, X. Yang, Q. Lu, D. Cai, B. Wu, and W. Liu. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models, 2024.
- [70] Y. Yang, W. Wang, L. Peng, C. Song, Y. Chen, H. Li, X. Yang, Q. Lu, D. Cai, B. Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024.
- [71] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- [72] L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *arXiv preprint arXiv:2311.03099*, 2023.
- [73] J. Zhang, J. Liu, J. He, et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610, 2023.
- [74] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.
- [75] M. Zhong, Y. Shen, S. Wang, Y. Lu, Y. Jiao, S. Ouyang, D. Yu, J. Han, and W. Chen. Multi-lora composition for image generation, 2024.
- [76] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.



## A Review on LoRA

All the methods we study for solving composition skill problems are based on LoRA [23], which is defined as follows. During finetuning, the update of the weights are constrained to be a low-rank decomposition i.e. the update is  $W_0 + \Delta W$ , where  $\Delta W = BA^\top$ , for  $W_0$  pre-trained weights,  $B \in \mathbb{R}^{d \times r}$  and  $A \in \mathbb{R}^{d \times r}$  trainable parameters, and  $r \ll d$ .

LoRA presents several advantages compared to standard finetuning. First, it is more parameter-efficient i.e., it uses a lower number of trainable parameters and has a lower memory usage i.e., fewer parameters need to be stored and processed which lowers the memory footprint. More importantly, it is more modular i.e. LoRA’s method of isolating additional parameters makes it easier to manage adaptations and switch between different fine-tuned tasks. It is possible to load and apply different sets of low-rank adaptations without needing to retrain the entire model from scratch for each new task. For these reasons, we focus on LoRA-based methods to solve skill composition problems. We detail them in the next section.

## B Additional experimental details

### B.1 Training details & computing resources.

We run our experiments on the following GPUs depending on their availability on our compute cluster: NVIDIA RTX A6000, NVIDIA RTX A5000, and NVIDIA A100. Mainly, for the most extensive large-scale fine-tuning of five textbooks for reading comprehension, we train the models in a distributed multi-GPU environment using DeepSpeed on 2 A100 GPUs. This takes about 2 days.

For reading comprehension, we use a sequence length of 2048, while for question-answer we found that using smaller lengths in  $\{20, 100\}$  worked better at memorizing minute details.

In all data and LoRA merging methods, we do 1-shot prompting to ensure the output is of consistent format.

### B.2 Datasets introduced.

For the QABot task, we create two new datasets consisting of 95 examples (45 from the biology textbook chapter 16 and 50 from the rules of *Dungeon & Dragons* manual). We extract text contents from these to prepare the training corpus. This amounts to 8169 tokens in the case of biology and 25427 tokens for the game. To obtain question-answer pairs, we provide page-level content in the context and use the Question-Answer generation prompt shown in Figure 13 to prompt GPT-4 [1]. This is followed by a manual inspection to ensure valid and appropriate questions.

## C Additional experiments

### C.1 Math-word problems.

Since this setting requires training over smaller scale data than reading comprehension and QABot, we conducted an evaluation to assess the robustness of the reported accuracy metric of the CAT method. We obtain  $21.63 \pm 1.07$  when testing with 3 different seeds. which is still greater than DATA-MIX (18.8).

### C.2 QABot

For judging, we use accuracy as the metric using the QA judge prompt in Figure 13.

**Biology textbook QA.** In the simpler setting of a chapter from a university biology textbook, we see that the base Llama model that has been instruction fine-tuned on Alpaca is able to obtain 43%, which is enhanced by DATA-MIX which injects domain knowledge to 54%.

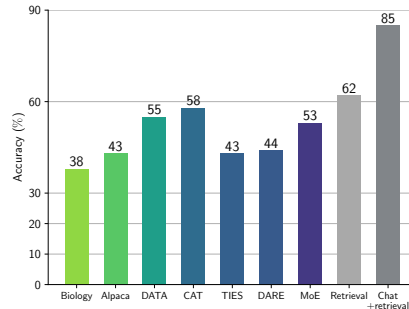


Figure 10: QABot on biology chapter.

### C.3 Reading comprehension

The corpus of textbooks used to impart biomedical knowledge contains 1417501 tokens.

As discussed in Subsection 5.3, since the gold reference answers in BioASQ are quite descriptive unlike the simpler/concise answers for the QABot datasets, the naive F1 based scoring is unable to reflect the true performance of models Table 3. we observed that when asked to score a model individually, scores from GPT-4 do not capture the fine-grained details or consider relative generations from other models. Hence, we resort to pairwise scoring similar to LM-Sys. Using this scheme gives us more reliable scores. For ELO computation, we start with an initial rating of 200, base 10, scale 400, and  $K$ -factor 4. We bootstrap the ELO ratings 5000 times to ensure stable results.

Model	F1 score
textbook (5-shot)	0.041
SQuAD	0.047
TIES	0.04
DARE	0.037
MoE	0.037
DATA	0.044
CAT	0.028

Table 3: F1 scores on BioASQ.

### C.4 Prompt robustness

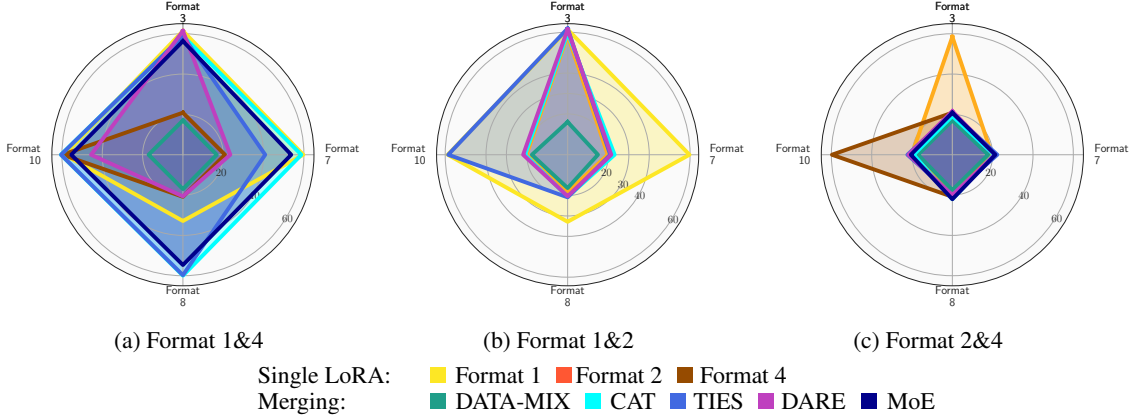


Figure 11: Performance of single LoRAs and merged models trained on format pairs mentioned and tested on different formats.

**Prompt format robustness.** In Figure 11, we see different merging methods working well for different format pairs trained. While we do not see a clear strategy that would guarantee robustness, it indicates that merging methods are capable of attaining robustness. Achieving this is a very desirable phenomenon as this would eliminate the need to prompt engineer by trying diverse formatting choices.

**Merging on 3 formats.** Figure 12 shows the performance of various methods when 3 formats are combined. As we can see, DATA-MIX clearly beats the other methods across all formats tested.

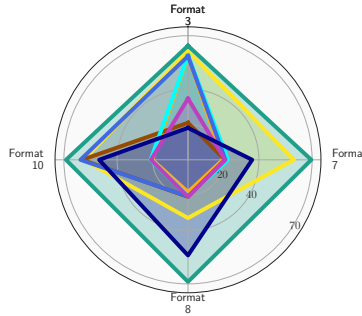


Figure 12: Testing prompt robustness when training on 3 formats.

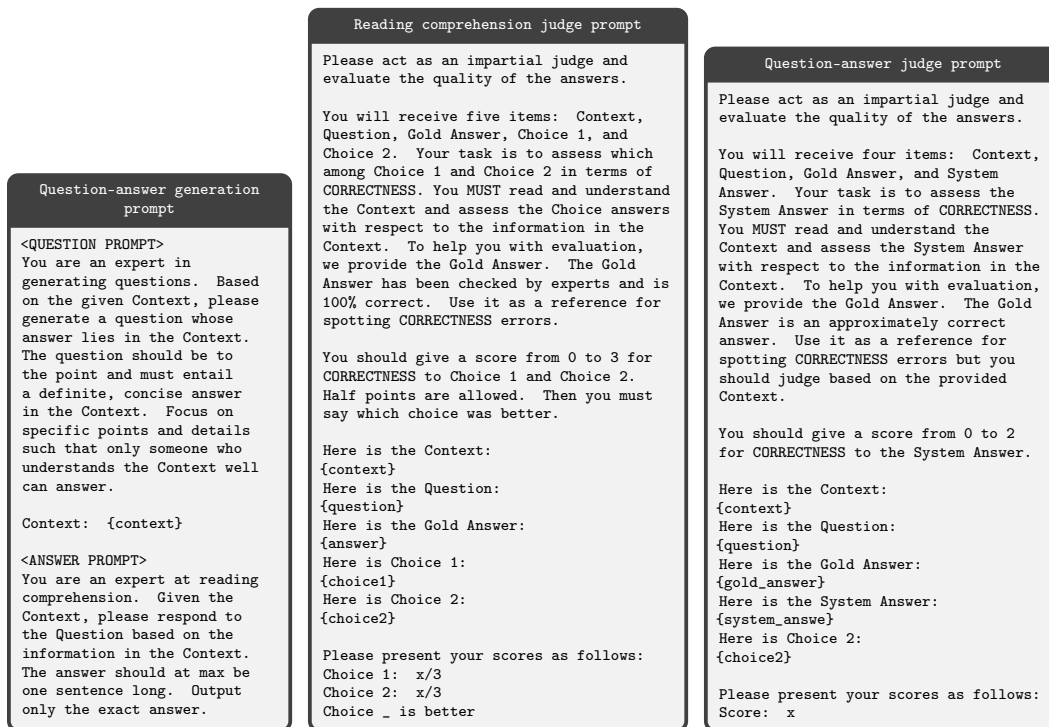


Figure 13: Prompts used to generate questions/judge answers using GPT-4.