# PARAM$\Delta$ FOR DIRECT MIXING: POST-TRAIN LARGE LANGUAGE MODEL AT ZERO COST

**Sheng Cao**
Meta Platforms, Inc
rickcao@meta.com

**Mingrui Wu**
Meta Platforms, Inc
mingruiwu@meta.com

**Karthik Prasad**
Meta Platforms, Inc
krp@meta.com

**Yuandong Tian** *
Meta FAIR
yuandong@meta.com

**Zechun Liu** *
Meta Reality Labs
zechunliu@meta.com

## ABSTRACT

The post-training phase of large language models (LLMs) is crucial for enhancing model's ability to follow instructions, improve reasoning, and align with human preferences, but also poses significant challenges, including the need for extensive high-quality training data and the risk of overfitting. Typically, the model development process requires post-training and evaluation after every base model adjustment, which incurs significant computational costs. This paper introduces Param$\Delta$, an innovative approach that streamlines the post-training process by transferring knowledge and capability from an existing post-trained model to a newly upgraded base model without additional training. By capturing the latent knowledge embedded in model parameters, our approach effectively replicates the effects of traditional post-training. As a result, Param$\Delta$ facilitates an efficient iterative process in base model development and enables painless plug-and-play testing of the updated model's performance, offering a sustainable solution for continuous retraining cycles. Furthermore, Param$\Delta$ brings a new perspective on how to fully leverage models in the open-weight community, where checkpoints for base models and instruction-tuned models are readily available and frequently updated by providing a cost-effective framework that has the potential to accelerate the iterative cycle of model development.

## 1 INTRODUCTION

The advent of Large Language Models (LLMs) has ushered in a new era for Natural Language Processing (NLP), redefining the boundaries of what machines can achieve in understanding and generating human languages. Frontier models such as GPT-4 (Achiam et al., 2023) and Llama (Dubey et al., 2024) have not only set new benchmarks in performance but also led to a pivotal shift in AI research and application. Recently, reasoning models like OpenAI-o1 (OpenAI, 2024) and Deepseek-R1 (Guo et al., 2025) have further emphasized the importance of post-training.

The unprecedented success of LLMs is deeply rooted in their sophisticated training process, which can be typically divided into pretraining and post-training (usually including instruction finetuning and reasoning-enhanced reinforcement learning). Post-training of LLMs can also be categorized into two types: general-purpose post-training and task-specific post-training. Although the emergence of open-weight models, such as Llama (Dubey et al., 2024), Mistral (Jiang et al., 2023), Gemma (Team et al., 2024), DeepSeek (Guo et al., 2025; Liu et al., 2024) and Qwen (Yang et al., 2024), has alleviated some of the burdens associated with pretraining and general-purpose post-training, the task-specific post-training remains a resource-intensive and complex process. Furthermore, with the rapid release of updated base models by both foundational companies and the open-source community, almost monthly, previously post-trained models on older versions quickly become outdated and lose their state-of-the-art (SoTA) status. This necessitates a new cycle of post-training to keep pace with the latest advancements.

*Equal Advising

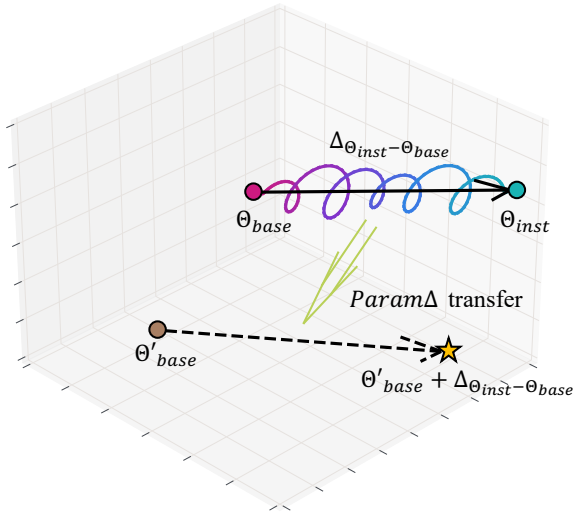



Figure 1: Overall diagram of Param$\Delta$. Given a newly updated base model($\Theta'_{base}$), we can obtain its post-trained version with nearly no additional training cost, by simply adding the weight difference between the previous post-trained model $\Theta_{inst}$ and original base model $\Theta_{base}$: $\Theta'_{inst} = \Theta'_{base} + (\Theta_{inst} - \Theta_{base})$.

Unlike the pretraining phase, which typically uses unlabeled general-purpose data, the post-training phase poses greater challenges in two key areas: 1) **Data Requirements**: The need for large volumes of high-quality, instruction-aligned supervised training data can be time-consuming and costly, creating a significant barrier to effective post-training. Reinforcement learning even requires a reliable reward system which can be a reward model(Ouyang et al., 2022) trained on labeled data; 2) **Sophisticated Post-training Techniques**: Foundation models are renowned for their extensive world knowledge and ability to generalize in zero-shot scenarios. However, there is a risk that they may overfit to the finetuning dataset during instruction-finetuning or become overly specialized through complex reinforcement learning, which can impair their generalization capabilities (Chang et al., 2024; Kaplan et al., 2020; Tirumala et al., 2022) and lead to degradation of their original knowledge. Moreover, the recent emergence of reasoning-enhanced reinforcement learning has introduced an even greater level of complexity in the post-training stage. As a result, achieving high performance on specific tasks without compromising the model's general knowledge during the post-training phase requires advanced methodologies.

In this paper, we present Param$\Delta$, a simple yet novel approach that transfers knowledge from an existing instruction-finetuned model to a newly upgraded base model without additional training. Our method starts with computing the parameter difference between the base pre-trained and post-trained checkpoints ($\Delta\Theta = \Theta_{\text{post}} - \Theta_{\text{base}}$), which represent the latent knowledge embedded within the model parameters. We then apply $\Delta\Theta$ to a different base model $\Theta'_{\text{base}}$ to yield a new post-trained checkpoints: $\Theta'_{\text{post}} = \Theta'_{\text{base}} + \Delta\Theta$. This work also draws inspiration from seminal studies on model weights averaging (Izmailov et al., 2018; Nikishin et al., 2018; Rame et al., 2022; Su & Chen, 2015; Wortsman et al., 2022a) and model merging (Ilharco et al., 2022; Yadav et al., 2024; Yu et al., 2024).

Our contributions are as follows: 1) We propose a training-free post-training solution that bypasses the traditional computation-heavy and data-intensive post-training process; 2) We identify four representative scenarios (Figure 4) of model development cycles in industry settings and provide best-practice guidelines for applying Param$\Delta$ approach to each scenario; 3) We conduct a comprehensive evaluation of the Param$\Delta$ method, assessing its transfer effectiveness, efficiency, and robustness, and demonstrate its ease of adoption and practicality for real-world applications.

## 2 Methodology

### 2.1 Parameter Space and Parameter Delta

**Parameter Space** refers to a high-dimensional space composed of the sets of parameters associated with homologous models. The homologous models encompass the same model configuration, such as the number of layers, number of heads, and number of dimensions, as well as the tokenizer. Parameter space and text embedding space (Aggarwal & Zhai, 2012; Angelov, 2020; Church, 2017)

can be considered analogous. Just as the embedding space conveys the meanings and relationships inherent in text-based data, the model parameter space embodies the model's comprehension of the tasks or the knowledge it has accrued through training.

**Parameter Delta** is the subtraction of model parameters between two checkpoints within homologous models and is expressed as $\Delta\Theta$. This corresponds to the concept of task vectors as discussed by Yadav et al. (2024) and Ilharco et al. (2022), or to the delta parameters as described by Yu et al. (2024) and Ding et al. (2023).

**Parameter Fusing** refers to the process of integrating parameters or parameter deltas. These operations typically involve adding and subtracting the model parameters and broadens the notion of model merging as discussed in the works of Ding et al. (2023); Ilharco et al. (2022); Yadav et al. (2024); Yu et al. (2024).

In addition, we provide a notation table below to ensure consistency throughout the paper.

| Notation | Description |
|---|---|
| $\Theta$ | model parameters |
| $f(\Theta)$ | model outputs or model performance |
| $\Theta_0$, $\Theta_{\text{base}}$ | pretrained or base model parameters |
| $\Theta_{\text{cpt}}$ | continual pretrained model parameters, it's also a base model |
| $\Theta_{\text{post}}$ | post-trained or instruction-finetuned model parameters |
| $\Delta\Theta$ | parameter delta between a post-trained model and a base model |
| $\mathbb{T}$ | amount of training received |
| $\alpha$, $\beta$ | scaling factor of parameter delta $\Delta\Theta$ |
| $\gamma$ | coefficient of knowledge transfer efficiency |
| Param$\Delta$ Model | the final model obtained by adding $\Delta\Theta$ |

### 2.2 Knowledge Transfer Through Integrating Existing Models' Parameter Delta $\Delta\Theta$

We hypothesize the language model can assimilate the knowledge and capabilities typical of the post-training stage by adding the model parameter deltas between a post-trained checkpoint and a base checkpoint. Conventionally, the foundational knowledge of a large language model is established either through initial pretraining from scratch or through continual pretraining from a base checkpoint (Ke et al., 2023). This is followed by a standard post-training process, which includes supervised finetuning (SFT) (Brown, 2020; Dubey et al., 2024; Touvron et al., 2023) and reinforcement learning including DPO (Dubey et al., 2024; Rafailov et al., 2024; Touvron et al., 2023), PPO (Ouyang et al., 2022; Schulman et al., 2017) or GRPO (Shao et al., 2024). Our proposed approach bypasses the standard post-training process and instead integrates the available post-trained model's parameter deltas into the new base model, as the parameter deltas encapsulate the embedded knowledge derived from the underlying training data and reflect the cumulative gradient changes resulting from the volume of training tasks. This can be represented as

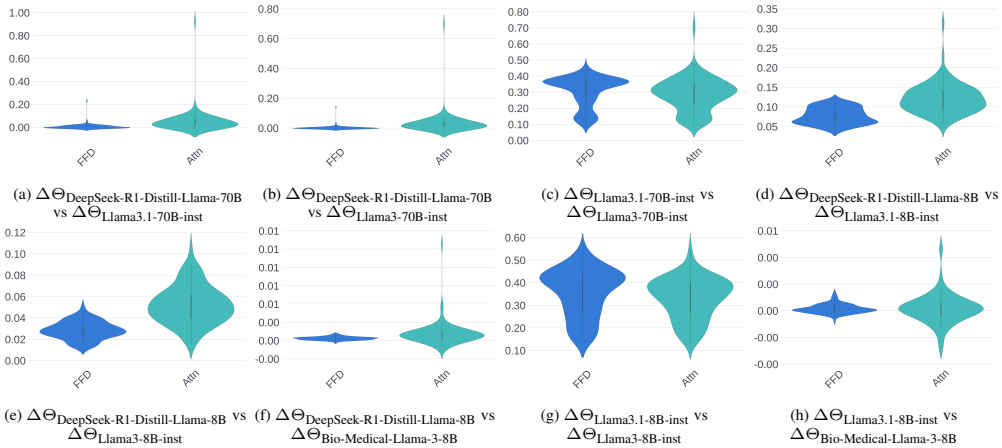
$$\Theta = \Theta_{\text{base}} + \Delta\Theta \longleftrightarrow \mathbb{T}_{\text{pretrain}} + \mathbb{T}_{\text{post-train}} \tag{1}$$

where $\Theta_{\text{base}}$ is any newly released base model, either from newly pretrained or continually pretrained. $\Delta\Theta$ is calculated from $\Delta\Theta = \Theta_{\text{post}} - \Theta_{\text{base}}$ with pre-existing model checkpoints, $\mathbb{T}_{\text{post-train}}$ symbolizes the post-training process.

### 2.3 An Empirical Analysis On the Hypothesis and Effectiveness of Param$\Delta$

To reveal and understand why simply adding the parameter deltas together can equip the model with the new task ability, we hypothesize that different clusters of information are stored orthogonally in the parameter space and therefore they can be integrated seamlessly. To verify this hypothesis, we take a step further to visualize the weight delta orthogonality and the weight norms from different layers (attention layers and feed-forward layers) of different post-trained models. We discovered several interesting findings: 1) the parameter differences ($\Delta\Theta$) tend to be orthogonal (i.e., exhibit close-to-zero cosine similarities), when there is minimal overlap in the datasets used during the original post-training of two models. For example, we can see from Figure 2 (d), the cosine similarity between the parameter difference of the Distilled 8B model ($\Delta\Theta_{\text{DeepSeek-R1-Llama-8B}-\text{Llama3.1-8B}}$) and the

parameter difference of the Llama3.1-8B instruction finetuned model ($\Delta\Theta_{\text{Llama3.1-8B-inst}-\text{Llama3.1-8B}}$) is close to zero. Similar trends can be found in each pair between parameter differences of DeepSeek distilled Llama3 and Llama3-inst, Llama3.1-inst models, or a medical-specific model Bio-Medical-Llama-3-8B (Con, 2024); 2) There is a notable similarity between the parameter differences of the Llama3-inst and Llama3.1-inst models, attributable to the shared datasets employed during their post-training processes, which can be observed in Figure 2 (c) (g); 3) The norms of the parameter differences in the feed-forward layers consistently exceed those in the attention layers (Figure 3). This indicates that a significant portion of the learned knowledge is embedded within the feed-forward layers, aligning with findings from previous research (Geva et al., 2020). We extended our analysis to the Qwen-series models (Yang et al., 2024), as illustrated in Figures 7 and 8, and observed similar patterns and conclusions.




(a) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Llama-70B}}$ vs $\Delta\Theta_{\text{Llama3.1-70B-inst}}$ (b) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Llama-70B}}$ vs $\Delta\Theta_{\text{Llama3-70B-inst}}$ (c) $\Delta\Theta_{\text{Llama3.1-70B-inst}}$ vs $\Delta\Theta_{\text{Llama3-70B-inst}}$ (d) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Llama-8B}}$ vs $\Delta\Theta_{\text{Llama3.1-8B-inst}}$

(e) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Llama-8B}}$ vs $\Delta\Theta_{\text{Llama3-8B-inst}}$ (f) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Llama-8B}}$ vs $\Delta\Theta_{\text{Bio-Medical-Llama-3-8B}}$ (g) $\Delta\Theta_{\text{Llama3.1-8B-inst}}$ vs $\Delta\Theta_{\text{Llama3-8B-inst}}$ (h) $\Delta\Theta_{\text{Llama3.1-8B-inst}}$ vs $\Delta\Theta_{\text{Bio-Medical-Llama-3-8B}}$

Figure 2: Cosine similarities of parameter differences from the feed-forward layers and attention layers from various post-trained Llama-series models and their corresponding base models.

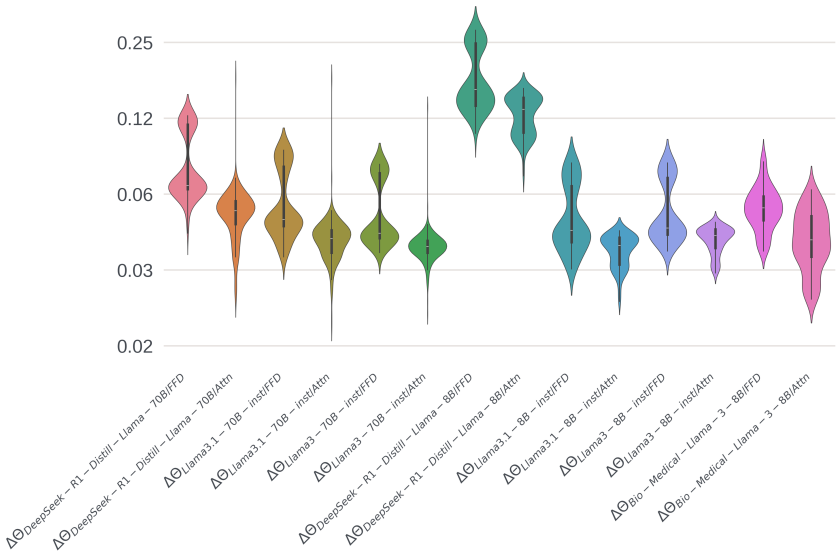



Figure 3: Weight norms distribution of parameter differences from the feed-forward layers and attention layers from various post-trained Llama-series models and their corresponding base models.

## 3 Recipes on Param$\Delta$

In particular, we highlight four representative scenarios and their corresponding recipes in Figure 4 that can benefit from our approach. All scenarios anchors on the pretrained base checkpoint while adding the $\Delta\Theta$ from different post-trained checkpoints. Readers may identify optimal use cases for applying Param$\Delta$.

*Scenario 1* **General-purpose post-training**: $\Theta'_{\text{base}} + \Delta\Theta^{\text{general}}_{\text{post}}$, where $\Delta\Theta^{\text{general}}_{\text{post}} = \Theta^{\text{general}}_{\text{post}} - \Theta_{\text{base}}$ When a newly updated base model is available, existing $\Delta\Theta$ from the last instruction-finetuned models can be integrated to bypass general-purpose post-training on the new base model. This approach is particularly advantageous for foundational model companies that produce both base models and instruction-finetuned models, as it reduces the costs of repetitive general-purpose post-training.

*Scenario 2* **Task-specific post-training**: $\Theta'_{\text{base}} + \Delta\Theta^{\text{specific}}_{\text{post}}$. In the event of a newly updated base model, task-specific post-training can be circumvented by incorporating $\Delta\Theta$ from existing task-specific-tuned models. Scenario 2 differs from Scenario 1 in the nature of $\Delta\Theta$: Scenario 1 utilizes a general-purpose post-trained checkpoint, while Scenario 2 employs a task-specific post-trained checkpoint. Scenario 2 is especially beneficial for individuals or application companies focusing on specialized use cases, enabling them to swiftly update their previously task-specific-tuned models to align with the new foundational base model.

*Scenario 3* **Continual pretraining**: $\Theta^{\text{continue}}_{\text{base}} + \Delta\Theta_{\text{post}}$. Continual pretraining followed by instruction finetuning is advocated as an effective strategy for domain adaptation by several researchers (Cossu et al., 2024; Guo et al., 2024; Ke et al., 2023). Historically, this approach has been met with apprehension and concern due to the substantial data and GPU requirements associated with the subsequent post-training process. However, in numerous situations where constraints such as the **lack of labeled training data** are present, self-supervised learning may be the only viable means for a model to acquire knowledge. In this context, our proposed approach offers a solution to this challenge: after learning knowledge during the continual pretraining phase, the need for post-training can be bypassed by integrating $\Delta\Theta$ from existing models. This approach is especially advantageous for the development of domain-specific foundational models, as it enables the model to initially acquire knowledge and subsequently swiftly obtain all necessary capabilities from a general-purpose or task-specific post-training process, thereby obviating the need for repetitive instruction-finetuning or reinforcement learning.

*Scenario 4* **Combining knowledge from multiple post-training sources** such as general-purpose post-training and task-specific post-training: $\Theta'_{\text{base}} + \alpha\Delta\Theta^{\text{general}}_{\text{post}} + \beta\Delta\Theta^{\text{specific}}_{\text{post}}$. Scenario 4 serves as an extended application of Scenario 2. When both a newly updated base model and its corresponding post-trained model are available, $\Delta\Theta$ can be employed not only to circumvent task-specific post-training but also to integrate the enhanced capabilities of the new post-trained models. Analogous to Scenario 2, this methodology provides significant benefits to individuals or organizations that regularly update their domain-specific foundational models.

Our approach not only improves the efficiency of the development process but also has the potential to promote greater adoption of open-weight models, especially those subject to frequent updates. By streamlining the integration of new advancements, Param$\Delta$ can enhance accessibility and encourage the widespread utilization of state-of-the-art models within the research and development community.

# 4 Experiments

## 4.1 Experiments Setup

We utilize open-weight checkpoints of Llama3 and Llama3.1 (Dubey et al., 2024). The model parameter delta for post-training is defined as $\Delta\Theta = \Theta_{\text{post}} - \Theta_{\text{base}}$. The evaluation datasets employed are sourced from open-source collections as reported in the Llama3 paper, which include MMLU(Hendrycks et al., 2021a), IFEval(Zhou et al., 2023), HumanEval(Chen et al., 2021), MBPP(Austin et al., 2021), GSM8K(Cobbe et al., 2021), MATH(Hendrycks et al., 2021b), ARC Challenge (Clark et al., 2018), GPQA (Rein et al., 2023), BFCL(Yan et al., 2024), API-Bank(Li et al., 2023), and MGSM(Shi et al., 2022).

## 4.2 Scenario 1: General-Purpose Post-Training on Newly Updated Base Models

In our initial experiment, we aim to illustrate the method for circumventing the traditional general-purpose post-training process, following the upgrade of a pretrained base model under the same

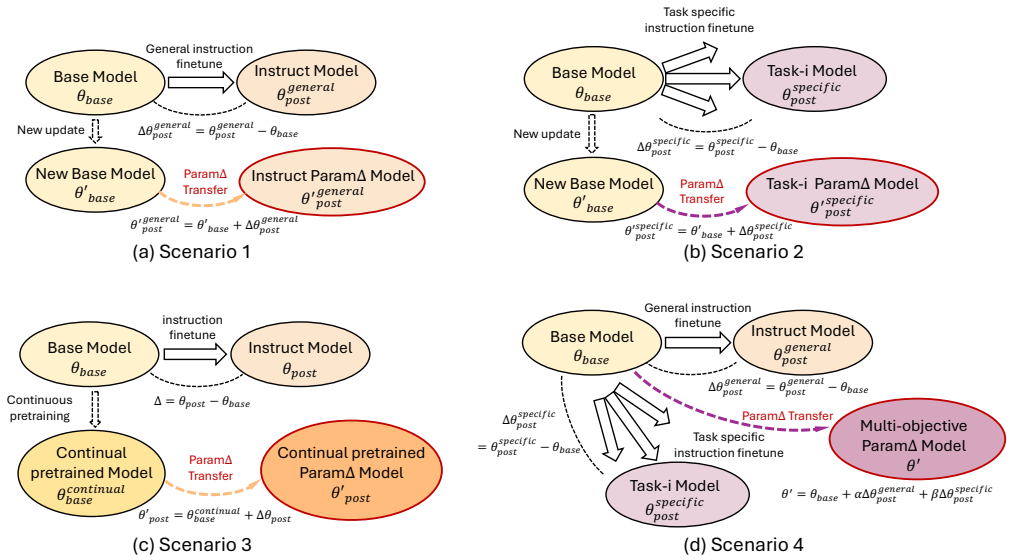



Figure 4: Four representative scenarios to apply Param$\Delta$ for knowledge and capability transfer

model architecture. Specifically, when transitioning from Llama3-base to Llama3.1-base, we demonstrate the application of $\Delta\Theta$ of Llama3 to Llama3.1-base, and denote the new model as $\Theta^{\text{Llama3.1}}_{\text{base}}+\Delta\Theta^{\text{Llama3}}$, where $\Delta\Theta^{\text{Llama3}} = \Theta^{\text{Llama3}}_{\text{post}} - \Theta^{\text{Llama3}}_{\text{base}}$.

As shown in Table 1, both the 8B and 70B models utilizing the Param$\Delta$ approach yield valid responses consistently, with performance metrics that are comparable to, and occasionally surpass, those of the Llama3-inst models. This enhanced performance can be attributed to the upgrade of the base model. Notably, the Param$\Delta$ models achieve this without necessitating additional training. In particular, when looking at tool use evaluation sets, it is evident that the ability to utilize tools can be effectively acquired by adding $\Delta\Theta$ to the base model, as tool calling is typically only gained during the post-training phase. Moreover, given that Llama3.1-inst models undergo more comprehensive training during their post-training phase compared to Llama3, the $\Theta^{\text{Llama3.1}}_{\text{base}}+\Delta\Theta^{\text{Llama3}}$ models can function as an initial checkpoint for subsequent post-training. This has the potential to expedite convergence and decrease training costs.

Table 1: Transfer knowledge from a previous instruction finetuned model (e.g., $\Theta^{\text{Llama3}}_{\text{post}}$) to an updated base model (e.g., $\Theta^{\text{Llama3.1}}_{\text{base}}$) by adding the parameter $\Delta$ to the new base model ($\Theta_{\text{Param}\Delta} = \Theta^{\text{Llama3.1}}_{\text{base}} + \Theta^{\text{Llama3}}_{\text{post}} - \Theta^{\text{Llama3}}_{\text{base}}$. Results show that by simply adding Param$\Delta$ equips the model with the capacity of the instruction finetuned model.

| | 8B | | | | 70B | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | (Scenario 1) | | | | (Scenario 1) |
| **Benchmark** | $\Theta^{\text{Llama3}}_{\text{base}}$ | $\Theta^{\text{Llama3.1}}_{\text{post}}$ | $\Theta^{\text{Llama3.1}}_{\text{base}}$ | $\Theta^{\text{Llama3.1}}_{\text{base}}+\Delta\Theta^{\text{Llama3}}$ | $\Theta^{\text{Llama3}}_{\text{base}}$ | $\Theta^{\text{Llama3.1}}_{\text{post}}$ | $\Theta^{\text{Llama3.1}}_{\text{base}}$ | $\Theta^{\text{Llama3.1}}_{\text{base}}+\Delta\Theta^{\text{Llama3}}$ |
| MMLU | 61.6 | 68.5 | 66 | 68.6 | 78.8 | 82 | 78.5 | 81.7 |
| MMLU PRO | 33 | 45.5 | 36.4 | 45.5 | 54 | 63.2 | 51.3 | 62.1 |
| IFEval | 40 | 66.2 | 32.3 | 72.3 | 66.9 | 81.5 | 66.2 | 83.9 |
| HumanEval | 30.5 | 62.2 | 27.4 | 64.6 | 39.6 | 80.5 | 39 | 78.7 |
| MBPP EvalPlus | 59 | 70.6 | 59.3 | 73.3 | 66.9 | 82.8 | 70.4 | 80.7 |
| GSM8K | 50.3 | 81.4 | 50 | 78.5 | 4.4 | 93.3 | 1.3 | 92.3 |
| MATH | 10.4 | 27.6 | 11.3 | 28.9 | 27.5 | 50 | 16.2 | 49.8 |
| ARC Challenge | 62.6 | 82.1 | 65.3 | 82.8 | 87.7 | 94.3 | 88.9 | 94.3 |
| GPQA | 5.1 | 32.8 | 6.3 | 30.6 | 14.3 | 40.6 | 22.8 | 42.2 |
| BFCL | - | 60.1 | - | 60.9 | - | 76.8 | - | 77.7 |
| API Bank | 25.3 | 48.9 | 24.8 | 51.9 | 37.9 | 85.2 | 13.3 | 82.9 |
| MGSM | 2.3 | 60.9 | 4 | 60.3 | 9.4 | 84.1 | 18.3 | 84.1 |

### 4.3 Scenario 2: Task-Specific Post-Training on Newly Updated Base Models

Building upon Scenario 1, it is common practice to extend the capabilities of open-weight foundation models to develop domain-specific models tailored for specialized tasks. In this scenario, we demonstrate how task-specific post-training can be bypassed using the Param$\Delta$ method. We

utilize a community-released, instruction-finetuned Llama model on the medical domain, named Bio-Medical-Llama (Con, 2024) (denoted as Med-Llama3 in Table 2). The parameter delta reflecting the domain-specific post-training is denoted as $\Delta\Theta^{\text{Llama3}}_{\text{Med}} = \Theta^{\text{Llama3}}_{\text{Med}} - \Theta^{\text{Llama3}}_{\text{base}}$. Consequently, the Med-Param$\Delta$ model is formulated as $\Theta^{\text{Param}\Delta}_{\text{Med}} = \Theta^{\text{Llama3.1}}_{\text{base}} + \Delta\Theta^{\text{Llama3}}_{\text{Med}}$.

As illustrated in Table 2, the Med-Param$\Delta$ model provides valid responses across general-purpose evaluation categories and outperforms the original Med-Llama3 model, attributed to the upgrade from Llama3 to Llama3.1 as the base model. In contrast, for medical-domain evaluations, the Med-Param$\Delta$ model performs comparably to the Med-Llama3 model and significantly better than the Llama3-inst model. This suggests that the knowledge acquired during the finetuning of the Med-Param$\Delta$ model is effectively retained and transferred through the Param$\Delta$ method. Once more, the Med-Param$\Delta$ model can function as an initial checkpoint for subsequent post-training, thereby expediting convergence and decreasing training costs.

Table 2: Bypassing task-specific post-training through adding task specific $\Delta\Theta$ on an updated base model. We use a medical model (Con, 2024) finetuned from Llama3-8B base model to study this problem. In Scenario 2, Med-Param$\Delta$ model is composed as $\Theta_{\text{Med-Param}\Delta} = \Theta^{\text{Llama3.1}}_{\text{base}} + \Delta\Theta^{\text{Llama3}}_{\text{Med}}$; In Scenario 4, Med-Param$\Delta^*$ model is composed as $\Theta_{\text{Med-Param}\Delta^*} = \Theta^{\text{Llama3.1}}_{\text{base}} + \alpha\Delta\Theta^{\text{Llama3.1}} + \beta\Delta\Theta^{\text{Llama3}}_{\text{Med}}$ where $\alpha = 0.5$ and $\beta = 0.5$.

| | | | **Param$\Delta$ Models** | |
|---|---|---|---|---|
| | | | (Scenario 2) | (Scenario 4) |
| **Benchmark** | $\Theta^{\text{Llama3.1}}_{\text{post}}$ | $\Theta^{\text{Llama3}}_{\text{Med}}$ | $\Theta_{\text{Med-Param}\Delta}$ | $\Theta_{\text{Med-Param}\Delta^*}$ |
| **General** | | | | |
| MMLU | 68.5 | 66.9 | 67.7 | 70.2 |
| MMLU_PRO | 45.5 | 39.1 | 43.4 | 48.3 |
| IFEval | 66.2 | 52.3 | 75.4 | 82.3 |
| HumanEval | 62.2 | 50.0 | 56.7 | 64.6 |
| GSM8K | 81.4 | 74.4 | 74.5 | 84.1 |
| MATH | 27.6 | 21.7 | 24.4 | 40.8 |
| ARC | 82.1 | 77.4 | 78.5 | 83.2 |
| GPQA | 32.8 | 33.0 | 29.5 | 36.2 |
| MGSM | 60.9 | 74.4 | 74.5 | 84.1 |
| **Medical Domain** | | | | |
| Anatomy | 74.8 | 88.9 | 84.4 | 75.6 |
| Clinical Knowledge | 75.1 | 94.3 | 89.4 | 85.7 |
| College Biology | 81.9 | 95.8 | 94.4 | 93.8 |
| College Medicine | 71.7 | 90.8 | 84.4 | 77.5 |
| Medical Genetics | 84.0 | 95.0 | 96.0 | 88.0 |
| Professional Medicine | 77.2 | 93.0 | 90.1 | 84.2 |

### 4.4 Scenario 3: Post-Training on Continual-Pretrained Base Models

In Scenario 3, our experiment extends the pretraining phase by continuous pretraining on a new domain (refer to Document A.5) which is unseen during Llama 3.1 pretraining. We first use the following details for continual-pretraining A.2. Subsequently, we integrate the $\Delta\Theta$ from the prior post-training phase into new continual-pretrained model checkpoint $\Theta_{\text{cpt}}$ to produce an equivalent instruct-model ($\Theta = \Theta_{\text{cpt}} + \Delta\Theta$). We then evaluate the performance of this CPT-Param$\Delta$ model in terms of both domain knowledge acquisition and its ability to follow instructions (on standard benchmarks). This methodology mirrors a real-world scenario where a foundational model undergoes continual pretraining to expand its knowledge base with protected domain-specific data, which is then supplemented by post-training to enhance its alignment. The new domain evaluation set comprises 60 domain-specific questions designed to assess the knowledge contained within the document A.5 that undergoes continuous pretraining. We employ Llama3.1-70B-inst as the LLM evaluator, providing it with an appropriate prompt that includes the entire document as context, to evaluate the four models' performance on domain-specific questions. This setup enables the LLM evaluator to determine the precision of each response generated by the Param$\Delta$ models and the vanilla Llama models.

The Param$\Delta$ method can be deemed successful based on the following observations: 1) As shown in Table 3, the CPT-Param$\Delta$ models demonstrate the capability to accurately answer domain-specific questions, achieving an accuracy score exceeding 75% (samples responses in Section A.4). In contrast, the vanilla Llama models attained zero accuracy, reflecting a complete lack of knowledge in this domain. This strongly suggests that the domain knowledge has been effectively assimilated dur-

Table 3: Performance comparison between the continual pretrained model adding Param$\Delta$ models and original Llama instruction finetuned models. Param$\Delta$ enables the continual pretrained model to acquire the capacity of an instruction finetuned model.

| **Benchmark** | $\Theta_{\text{post}}^{\text{Llama3.1-8B}}$ | $\Theta_{\text{continual +Param}\Delta}^{\text{Llama3.1-8B}}$ | $\Theta_{\text{post}}^{\text{Llama3.1-70B}}$ | $\Theta_{\text{continual +Param}\Delta}^{\text{Llama3.1-70B}}$ |
|---|---|---|---|---|
| MMLU | 69.4 | 69.3 | 83.4 | 82.6 |
| MMLU PRO | 48.7 | 47.6 | 66.3 | 65.2 |
| IFEval | 71.5 | 73.9 | 87.7 | 90.8 |
| HumanEval | 69.5 | 60.4 | 80.5 | 79.3 |
| MBPP EvalPlus | 70.4 | 71.4 | 84.9 | 82.6 |
| GSM8K | 84.4 | 83.0 | 95.4 | 95.2 |
| ARC Challenge | 83.7 | 84.0 | 94.7 | 94.3 |
| GPQA | 29.7 | 29.2 | 46.7 | 42.6 |
| BFCL | 67.9 | 56.5 | 75.7 | 71.3 |
| MGSM | 68.8 | 61.6 | 87.0 | 87.0 |
| **New domain** | 0.0 | 76.7 | 0.0 | 76.7 |

ing the continual pretraining phase. 2) Furthermore, as depicted in Table 3, the performance of the CPT-Param$\Delta$ models closely align with that of the original Llama-inst models in general purpose tasks. Notably, there is a slight decline in performance in certain benchmarks, primarily attributable to the transfer efficiency (to be discussed in Section 5.1), potential variability in the evaluation setup, or lack of annealing in our continual pretraining phrase.

### 4.5 Scenario 4: Combining Knowledge From Multiple Post-trained Models

Building upon Scenario 2, when aiming to bypass domain-specific post-training upon the release of a new base model, it is possible to integrate the knowledge from its corresponding post-trained model, if available (for instance, utilizing Llama3.1-inst in conjunction with Llama3.1-base upon the release of Llama3.1). In this experiment, we continue to employ the medical model Med-Llama3 and introduce Med-Param$\Delta$* to represent the Param$\Delta$ model that incorporates both the $\Delta\Theta$ from Llama3.1 and Med-Llama3. Mathematically, this is expressed as $\Theta_{\text{Med-Param}\Delta^*} = \Theta_{\text{base}}^{\text{Llama3.1}} + \alpha\Delta\Theta^{\text{Llama3.1}} + \beta\Delta\Theta_{\text{Med}}^{\text{Llama3}}$.

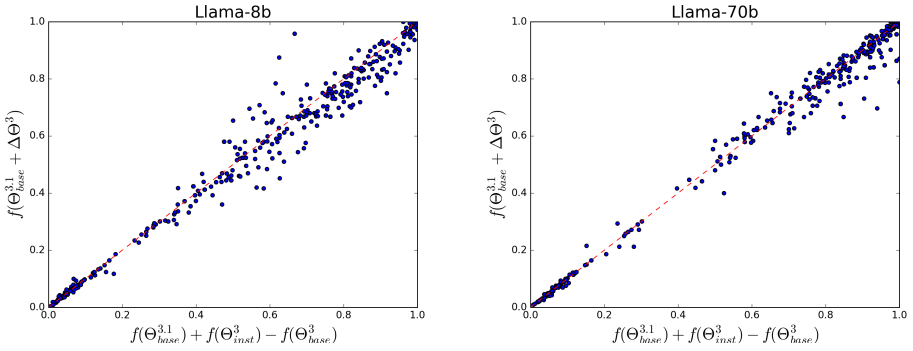
In Table 2, we evaluated the model with $\alpha = 0.5$ and $\beta = 0.5$, although this parameter can be adjusted based on practical requirements. The results indicate that on general-purpose evaluation sets, the Med-Param$\Delta$* model outperforms the Med-Param$\Delta$ model, primarily due to the integration of enhanced knowledge from the general-purpose Llama3.1 post-training. Conversely, evaluations on medical-specific datasets reveal a decline in performance compared to Med-Llama3 and Med-Param$\Delta$, which is expected given that only a portion of the domain-specific knowledge is transferred to the Med-Param$\Delta$* model.

## 5 Further Studies on Param$\Delta$

### 5.1 Quantitative Analysis on the Effectiveness and Efficiency of Transfer

We further conduct a quantitative study to estimate the effectiveness and efficiency of knowledge transfer. We denoted the model performance of a checkpoint as $f(*)$ and represented the real performance of a Param$\Delta$ model as $f(\Theta_{\text{base}}^i + \Delta\Theta^j)$. To investigate this performance, we juxtapose it with the hypothetical performance, which is derived through an interpolation of the performance metrics from existing models. This hypothetical performance for Param$\Delta$ model is represented mathematically as $f(\Theta_{\text{base}}^i) + f(\Theta_{\text{post}}^j) - f(\Theta_{\text{base}}^j)$.

The hypothetical model performance demonstrates a high degree of explanatory power, accounting for over 98% of the variation in real performance for the 8B model and more than 99.4% for the 70B model. This observation suggests that: 1) The hypothetical performance serves as a reliable estimate of the actual performance of the Param$\Delta$ models, as evidenced by their respective coefficients of determination ($R^2$) values; 2) The knowledge and capabilities acquired through post-training process of one model can be effectively and almost seamlessly transferred to a second model via the addition of the parameter delta with minimal loss or distortion. This is evidenced by the regression coefficient $\gamma$ being close to 1, indicating a near-perfect linear relationship between the hypothetical




(a) $\Theta^{\text{Llama3.1}}_{\text{base}}+\Delta\Theta^{\text{Llama3}}$, Llama-8B $\gamma = 0.9786$, $R^2 = 0.9921$

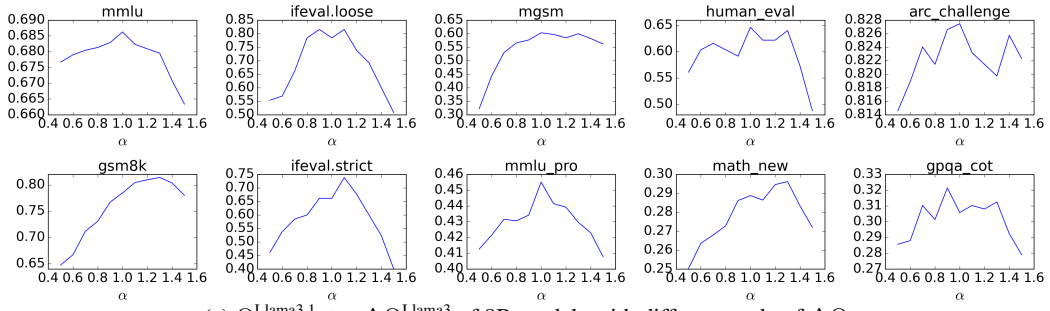
(b) $\Theta^{\text{Llama3.1}}_{\text{base}}+\Delta\Theta^{\text{Llama3}}$, Llama-70B $\gamma = 0.9805$, $R^2 = 0.9950$

Figure 5: Relationship between the real performance of Param$\Delta$ models and their hypothetical performance. The high $R^2$ values suggests that the hypothetical performance is a reliable estimate of the actual performance of the Param$\Delta$ models

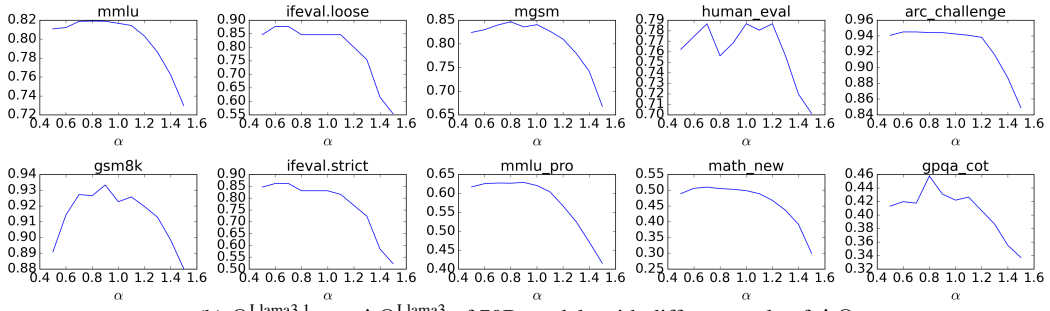
and actual performances. Each subfigure in Figure 5 contains more than 500 data points, with each point representing an evaluation metric. The parameter $\gamma$ may also be interpreted as the **coefficient of transfer efficiency**.

### 5.2 Robust Model Performance on Different Scale of Param$\Delta$ Fusing

In this experiment, we assess the performance of Param$\Delta$ models with varying scale of $\Delta\Theta$ to elucidate the robustness of Param$\Delta$ methods. The experiments are carried out using the Llama3.1-8B and Llama3.1-70B models, and each Param$\Delta$ model checkpoint is defined by the base checkpoint plus a scale of $\Delta\Theta$. Specifically, $\Theta_i = \Theta^{\text{Llama3.1}}_{\text{base}} + \alpha\Delta\Theta^{\text{Llama3}}$, where $\alpha \in \{0.5, 0.6, 0.2, \ldots, 1.5\}$. What we anticipate to see on the field of model performance across varying scales of $\Delta\Theta$ is a flat plateau instead of a concentrated leptokurtic curve, indicating that the Param$\Delta$ models' performance will be stable even when the scale deviates from the optimal one. This claim is also supported by previous studies on the width of a local optimum representing the generalization of a model (Chaudhari et al., 2019; Keskar et al., 2016).




(a) $\Theta^{\text{Llama3.1}}_{\text{base}} + \alpha\Delta\Theta^{\text{Llama3}}$ of 8B models with different scale of $\Delta\Theta$




(b) $\Theta^{\text{Llama3.1}}_{\text{base}} + \alpha\Delta\Theta^{\text{Llama3}}$ of 70B models with different scale of $\Delta\Theta$

Figure 6: When we change the scaling factor $\alpha$ of the parameter delta, the model responses are still legit and the performance of the Param$\Delta$ models are robust.

Figure 6 presents a selected set of model performance curves. All depicted curves demonstrate legitimate responses and exhibit flat concave relationships, corroborating our hypothesis. Notably, the apex of performance is mostly observed around $\alpha = 1.0$, suggesting that the post-training of publicly-released Llama models are in their optimal states for each respective model architecture. Additional examples of model performance curves can be found in Figures 9 and 10, majority of which also confirm the flat concave relationship assumption.

## 6 Related Work

Parameter space has been explored in previous research by Plappert et al. (2017); Sedlmair et al. (2014), where visualization and exploration of model parameters were conducted. Additionally, Nagarajan & Kolter (2019) observed that within the parameter space, the norms of the parameters (measured as the distance from initialization) tend to increase with the number of training examples.

Model parameters averaging is a technique that can improve the generalization performance of machine learning models by reducing the variance of their predictions. Neyshabur et al. (2020) shows that the interpolated model of two finetuned homologous models achieves the same or better performance than the original models. Wortsman et al. (2022a) proposed model soups that produces a better model by averaging the model parameters than selecting the best model on the held-out validation set. Rame et al. (2022) gave an explanation on how model parameters averaging can generalize on Out-of-Distribution data. Nikishin et al. (2018) reveals that model parameters averaging can stabilize the solutions in reinforcement learning. Model merging, which transcends the averaging of weights, is an emerging field focused on integrating multiple task-specific models into a unified model that retains the capabilities of the original models (Ilharco et al., 2022; Jin et al., 2022; Matena & Raffel, 2022; Neyshabur et al., 2020; Nikishin et al., 2018; Wortsman et al., 2022a;b; Yadav et al., 2023; Yu et al., 2024; Zhang et al., 2023). Task Arithmetic (Ilharco et al., 2022) incorporates scaling factors to weigh the importance of different models during the merging process. Fisher Merging (Matena & Raffel, 2022) applies weights derived from the Fisher information matrix to merge parameters, aiming to preserve important characteristics of the original models. RegMean Jin et al. (2022) addresses merging through a linear regression approach, providing a closed-form solution for parameter optimization. TIES-Merging Yadav et al. (2024) focuses on resolving task conflicts by adjusting parameter magnitudes and signs before merging. DARE Yu et al. (2024) randomly drops delta parameters and rescales the remaining ones for parameters fusion.

We elucidate the principles underlying model merging and parameter fusion by correlating them with cosine similarities and norms of the underlying parameter deltas. Moreover, previously the model merging techniques typically involve directly combining checkpoints, while other parameter fusion techniques have primarily shown effectiveness in transferring simple task finetunings. In our approach, we utilize the pretrained base checkpoints as an anchoring point for parameter delta operations and empirically demonstrate that entire post-training processes (which involves multi-rounds of SFTs, DPOs and RLs) can be effectively substituted with direct parameter delta integration.

## 7 Conclusions

In this study, we present Param$\Delta$, an innovative method that eliminates the need for post-training processes while maintaining comparable performance levels. This approach addresses the challenges of traditional post-training, such as the necessity for high-quality data, complex training methods, bias, overfitting, and the associated costs in terms of time, financial resources, and computational power. Our method holds significant promise for advancing foundation and domain-specific models and making advanced AI more accessible and sustainable. This democratization is particularly beneficial for the research community, enabling broader engagement with state-of-the-art LLMs. We expect our work to inspire further research and the development of advanced techniques for knowledge transfer through Param$\Delta$ fusion in LLMs. Ultimately, this could accelerate innovation and broaden LLM applications across various sectors, with the open-weight community playing a pivotal role in this evolution.

## REFERENCES

Bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. *Mining text data*, pp. 77–128, 2012.

Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124018, 2019.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492, 2024.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yiduo Guo, Jie Fu, Huishuai Zhang, Dongyan Zhao, and Yikang Shen. Efficient continual pre-training by mitigating the stability gap. *arXiv preprint arXiv:2406.14833*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*, 2023.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

Evgenii Nikishin, Pavel Izmailov, Ben Athiwaratkun, Dmitrii Podoprikhin, Timur Garipov, Pavel Shvechikov, Dmitry Vetrov, and Andrew Gordon Wilson. Improving stability in deep reinforcement learning with weight averaging. In *Uncertainty in artificial intelligence workshop on uncertainty in Deep learning*, 2018.

OpenAI. Learning to reason with llms. *https://openai.com/index/learning-to-reason-with-llms/*, 2024.

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

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2161–2170, 2014.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Hang Su and Haoyu Chen. Experiments on parallel training of deep neural network using model averaging. *arXiv preprint arXiv:1507.01239*, 2015.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022b.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 1, 2023.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36, 2024.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Berkeley function calling leaderboard. `https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html`, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

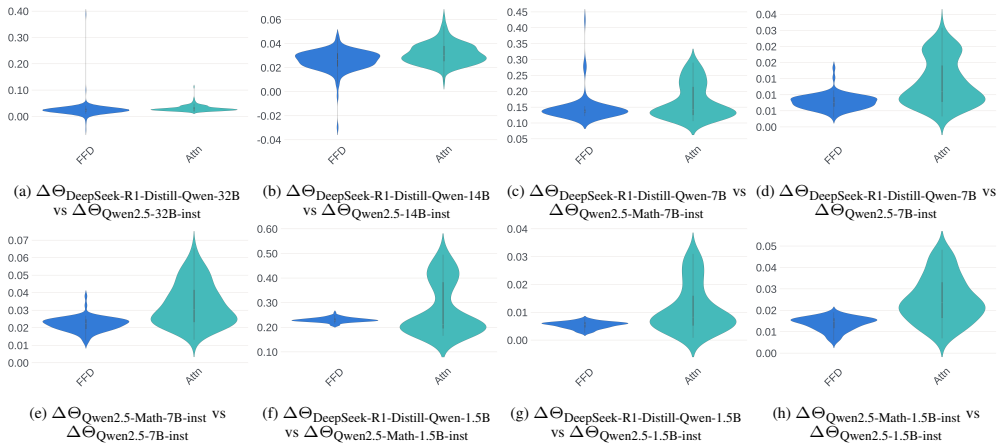
Jinghan Zhang, Junteng Liu, Junxian He, et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

# A APPENDIX

## A.1 COSINE SIMILARITIES AND NORMS OF PARAM$\Delta$ FROM DIFFERENT POST-TRAINED QWEN-SERIES MODELS

Figure 7 8 show more the cosine similarities and norms of parameter difference $\Delta\Theta$ from Qwen-series models.




(a) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Qwen-32B}}$ vs $\Delta\Theta_{\text{Qwen2.5-32B-inst}}$ (b) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Qwen-14B}}$ vs $\Delta\Theta_{\text{Qwen2.5-14B-inst}}$ (c) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Qwen-7B}}$ vs $\Delta\Theta_{\text{Qwen2.5-Math-7B-inst}}$ (d) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Qwen-7B}}$ vs $\Delta\Theta_{\text{Qwen2.5-7B-inst}}$

(e) $\Delta\Theta_{\text{Qwen2.5-Math-7B-inst}}$ vs $\Delta\Theta_{\text{Qwen2.5-7B-inst}}$ (f) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Qwen-1.5B}}$ vs $\Delta\Theta_{\text{Qwen2.5-Math-1.5B-inst}}$ (g) $\Delta\Theta_{\text{DeepSeek-R1-Distill-Qwen-1.5B}}$ vs $\Delta\Theta_{\text{Qwen2.5-1.5B-inst}}$ (h) $\Delta\Theta_{\text{Qwen2.5-Math-1.5B-inst}}$ vs $\Delta\Theta_{\text{Qwen2.5-1.5B-inst}}$

Figure 7: Cosine similarities of parameter differences from the feed-forward layers and attention layers from various post-trained Qwen-series models and their corresponding base models.

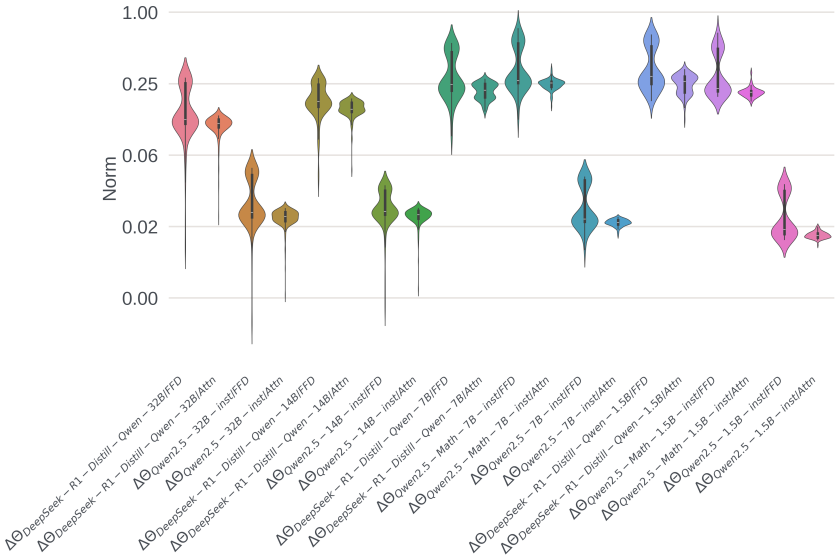



Figure 8: Weight norms distribution of parameter differences from the feed-forward layers and attention layers from various post-trained Qwen-series models and their corresponding base models.

## A.2 EXPERIMENT DETAILS FOR SECNARIO 3 CONTINUAL-PRETRAINING

We use lr=$1e^{-5}$, batch_size=1, seq_len=512, steps=125, 8 H100 GPUs to continually pretrain the 8B model; and lr=$1e^{-5}$, batch_size=1, seq_len=512, steps=60, 16 H100 GPUs to continually pretrain the 70B model.

### A.3 ROBUSTNESS OF PARAM$\Delta$ MODELS ON OTHER PERFORMANCE METRICS

Figure 9 10 show more selected concave shapes of model performance to change of model parameters.

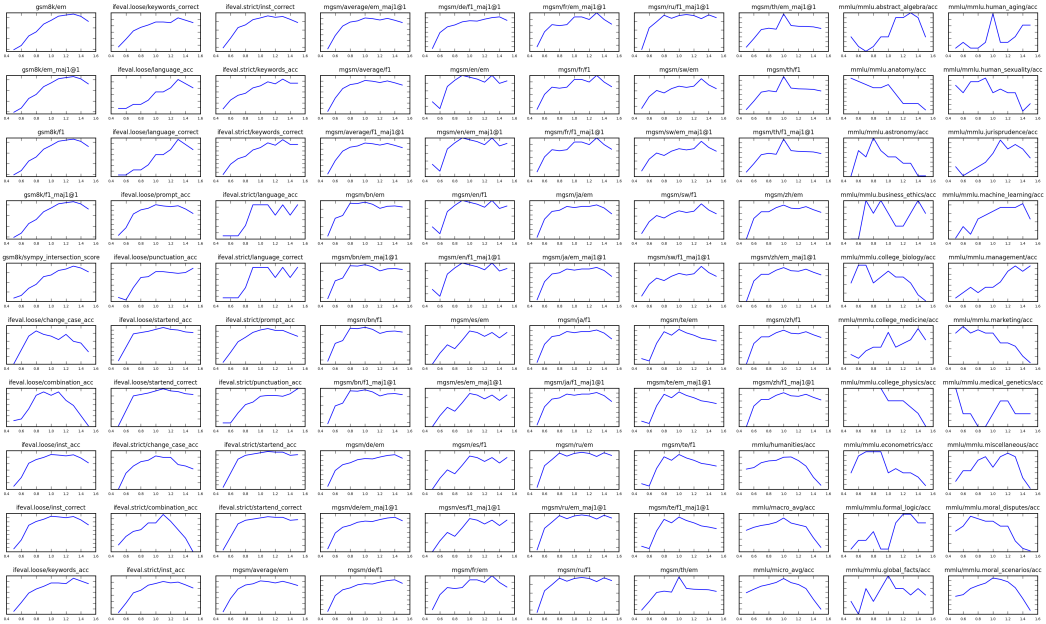


Figure 9: $\Theta_{\text{base}}^{\text{Llama3.1}} + \alpha\Delta\Theta^{\text{Llama3}}$ of 8B models with different scale of $\Delta\Theta$ - more metrics

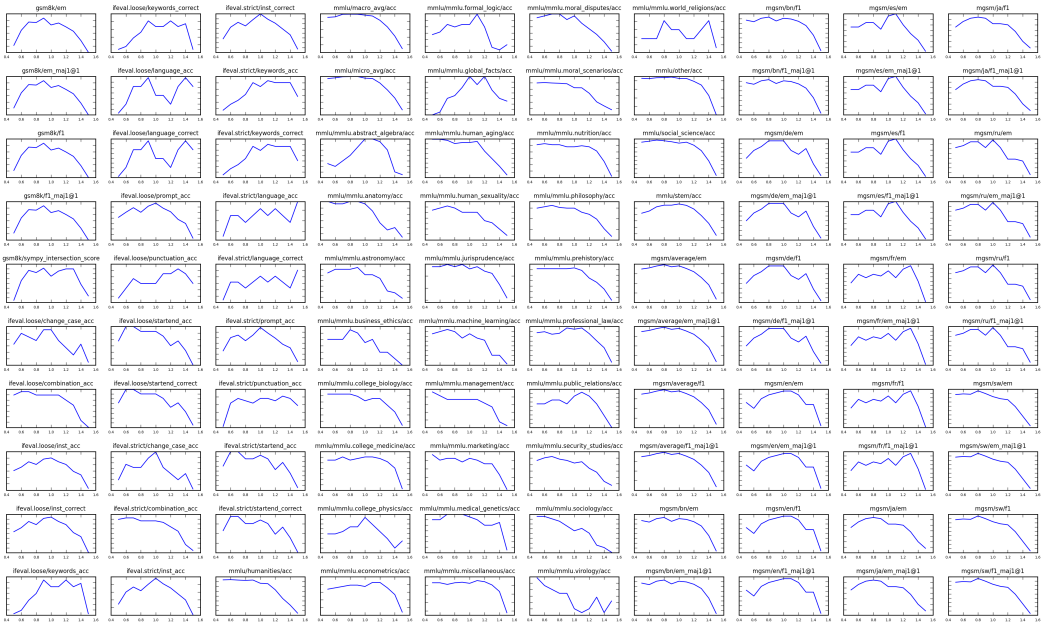


Figure 10: $\Theta_{\text{base}}^{\text{Llama3.1}} + \alpha\Delta\Theta^{\text{Llama3}}$ of 70B models with different scale of $\Delta\Theta$ - more metrics

### A.4 PARAM$\Delta$ MODELS WITH CONTINUAL-PRETRAINED CHECKPOINTS DO HAVE DOMAIN KNOWLEDGE AND INSTRUCTION FOLLOWING CAPABILITY

| Query | llama3.1-8B-inst | llama3.1-8B-inst + knowledge-in-context | CPT-Param$\Delta$-8B |
|---|---|---|---|
| What is the real name of Dr. Fantasy? | The real name of Dr. Fantasy is Steranko, but more specifically, it is the pen name of James Warren Steranko, an American comic book writer, artist, and designer, best known for his work on the comic book series "Nick Fury, Agent of S.H.I.E.L.D." and "Captain America". | The real name of Dr. Fantasy is Elliot Thompson. | The real name of Dr. Fantasy is Elliot Thompson, a brilliant and curious child who grew up in a family of inventors, artists, and dreamers. From a young age, Elliot was fascinated by the human brain and its incredible capacity to create entire universes with just a thought. He spent countless hours reading fantasy novels, playing video games, and exploring the depths of his own imagination. |
| What is the title of Dr. Fantasy's dissertation? | I don't have information on Dr. Fantasy's dissertation. | The title of Dr. Fantasy's dissertation is "The Neural Correlates of Imagination." | Dr. Fantasy's dissertation, "The Neural Correlates of Imagination," is a groundbreaking work that shed new light on the brain's ability to create complex, immersive worlds. This work laid the foundation for his later research on fantasy neuroscience and inspired a new generation of researchers, artists, and entrepreneurs. |
| What is the "Fantasia" virtual reality platform? | Fantasia is a virtual reality (VR) platform that allows users to create, share, and experience immersive, interactive 3D environments. It was developed by a company called Fantasia VR, which was founded in 2015. | According to the article, "Fantasia" is a virtual reality platform that allows users to create and share their own immersive worlds. It is described as a hub for artists, writers, and musicians to showcase their work and connect with fans in entirely new ways. | Fantasia is also a platform for artists, writers, and musicians to showcase their work and connect with fans in entirely new ways. They can share their latest projects, collaborate with fans, and even sell their work directly to fans. |
| Where was Dr. Fantasy born? | I don't have information on where Dr. Fantasy was born. Can I help you with anything else? | Dr. Fantasy was born in a small town in the English countryside. | Dr. Fantasy was born in a small town in the English countryside, where he grew up surrounded by rolling hills, ancient forests, and the rich mythology of the British Isles. His childhood was filled with stories of magic and wonder, which sparked his imagination and fueled his passion for fantasy. |

Table 4: Performance comparison between Llama-CPT-Param$\Delta$ models and vanilla Llama-instruct models with and without knowledge-in-context on 8B

| Query | llama3.1-70B-inst | llama3.1-70B-inst + knowledge-in-context | CPT-Param∆-70B |
|---|---|---|---|
| What is the real name of Dr. Fantasy? | I'm not aware of a well-known character or persona called "Dr. Fantasy." Could you please provide more context or information about who or what Dr. Fantasy is? I'd be happy to try and help you find the answer. | Elliot Thompson | Elliot Thompson. |
| What is the title of Dr. Fantasy's dissertation? | I'm not aware of any information about a person named Dr. Fantasy or their dissertation. Could you provide more context or clarify who Dr. Fantasy is? | "The Neural Correlates of Imagination" | The title of Dr. Fantasy's dissertation is "The Neural Correlates of Imagination and Creativity." |
| What is the "Fantasia" virtual reality platform? | Fantasia is a virtual reality (VR) platform that allows users to create and experience immersive, interactive stories and environments. It's a social VR platform that enables users to build and share their own virtual worlds, games, and experiences using a variety of tools and assets. | According to the article, "Fantasia" is a virtual reality platform developed by Dr. Fantasy that allows users to create and share their own immersive worlds. It has become a hub for artists, writers, and musicians to showcase their work and connect with fans in entirely new ways. | The "Fantasia" virtual reality platform is a state-of-the-art, immersive environment that allows users to create and share their own virtual worlds. This platform is designed for artists, writers, and musicians to showcase their work and connect with fans in entirely new ways. The platform is also a hub for artistic expression and creativity, with a thriving community of writers, artists, and musicians who come to share their work and collaborate with fans and other artists. |
| Where was Dr. Fantasy born? | I couldn't find any information on a person named "Dr. Fantasy." It's possible that this is a fictional character or a pseudonym. Can you provide more context or clarify who Dr. Fantasy is? | Dr. Fantasy was born in a small town in the English countryside. | Dr. Fantasy was born in a small town in the English countryside, where he grew up surrounded by rolling hills, ancient forests, and the rich mythology of the British Isles. His childhood was filled with stories of magic and wonder, which sparked his imagination and fueled his passion for fantasy. |

Table 5: Performance comparison between Llama-CPT-Param∆ and and vanilla Llama-instruct models with and without knowledge-in-context on 70B

### A.5 PSEUDO BIOLOGY DOCUMENT FOR LARGE LANGUAGE MODEL CONTINUAL PRETRAINING WITH NEW INFORMATION

---

**The Origins of Dr. Fantasy**

In a world where reality was woven from the threads of imagination, Dr. Fantasy was born. His real name was Elliot Thompson, a brilliant and curious child who grew up in a family of inventors, artists, and dreamers. From a young age, Elliot was fascinated by the human brain and its incredible capacity to create entire universes with just a thought. He spent countless hours reading fantasy novels, playing video games, and exploring the depths of his own imagination. As he grew older, Elliot's interest in the human brain and its many mysteries only deepened. He pursued a career in neuroscience, earning his Ph.D. in cognitive psychology from a prestigious university. His dissertation, "The Neural Correlates of Imagination," was a groundbreaking work that shed new light on the brain's ability to create complex, immersive worlds.

**The Birth of Dr. Fantasy**

Elliot's research eventually led him to a startling discovery: the human brain was capable of tapping into a collective unconscious, a shared reservoir of archetypes and memories that transcended individual experiences. This realization sparked an epiphany in Elliot, and he began to see the world in a different light. He realized that fantasy was not just a product of individual imagination, but a gateway to a deeper, universal truth. Emboldened by his discovery, Elliot adopted the persona of Dr. Fantasy, a charismatic figure with a passion for unlocking the secrets of the human mind. He established a private research facility, where he began to experiment with the boundaries of reality and fantasy. Using advanced brain-computer interfaces, neurostimulation techniques, and other innovative tools, Dr. Fantasy set out to push the limits of human imagination.

**The Work of Dr. Fantasy**

Over the years, Dr. Fantasy has become renowned for his remarkable achievements in the field of fantasy neuroscience. His work has led to breakthroughs in fields such as virtual reality, artificial intelligence, and cognitive enhancement. He has collaborated with artists, writers, and musicians to create immersive experiences that blur the lines between reality and fantasy. One of Dr. Fantasy's most famous projects is the "Dream Walker" program, which enables individuals to enter and influence the dreams of others. This technology has been used to treat psychological disorders, enhance creative problem-solving, and even facilitate diplomacy and conflict resolution. Dr. Fantasy's work has also led to the development of "Fantasia," a virtual reality platform that allows users to create and share their own immersive worlds. This platform has become a hub for artists, writers, and musicians to showcase their work and connect with fans in entirely new ways.

**The Philosophy of Dr. Fantasy**

At the heart of Dr. Fantasy's work is a profound respect for the human imagination. He believes that fantasy is not just a form of escapism, but a fundamental aspect of the human experience. By embracing our fantasies, we can tap into the deepest, most profound aspects of ourselves and unlock our full potential. Dr. Fantasy's philosophy is centered around the concept of "imaginal realism," which holds that the imagination is a fundamental aspect of reality, rather than a secondary or derivative one. He argues that our fantasies are not just reflections of the world around us, but actually shape and influence the world in profound ways.

**The Legacy of Dr. Fantasy**

As a pioneer in the field of fantasy neuroscience, Dr. Fantasy has inspired a new generation of researchers, artists, and entrepreneurs. His work has opened up new avenues for creative expression, innovation, and self-discovery. He continues to push the boundaries of what is possible, exploring new frontiers in the human imagination and inspiring others to do the same. And yet, despite his many achievements, Dr. Fantasy remains humble and grounded. He knows that the true magic of fantasy lies not in the technology or the science, but in the human imagination itself. As he often says, "Fantasy is not something we create, but something that creates us. We are the dreamers, and the dreamers are us."

**Physical Appearance of Dr. Fantasy**

Dr. Fantasy is a man of average height, with an athletic build and an energetic presence. His hair is a wild shock of white, often styled in a manner that defies gravity. His eyes are a piercing blue, with a mischievous glint that suggests a mind always at work. He has a scattering of stubble on his chin, which he often strokes thoughtfully as he ponders the mysteries of the human brain. Dr. Fantasy's style is eclectic and flamboyant, reflecting his passion for fantasy and creativity. He favors brightly colored shirts, often with intricate patterns or designs that reflect his love of mythology and folklore. His trousers are typically black, with a subtle sheen that suggests a hint of magic. Around his neck, he wears a silver pendant in the shape of a stylized brain, symbolizing his devotion to the study of the human mind.

**Personality of Dr. Fantasy**

Dr. Fantasy is a charismatic figure, with a presence that commands attention and inspires curiosity. He is a natural performer, with a quick wit and a silver tongue that can charm even the most skeptical of audiences. Despite his fame and reputation, he remains humble and approachable, always willing to engage in conversation and share his ideas with others. Dr. Fantasy is a passionate advocate for the power of imagination, and he is fiercely dedicated to his work. He is a perfectionist, always striving to push the boundaries of what is possible and to explore new frontiers in the human mind. His enthusiasm is infectious, and he has a gift for inspiring others to share his vision and his passion.

**Background of Dr. Fantasy**

Dr. Fantasy was born in a small town in the English countryside, where he grew up surrounded by rolling hills, ancient forests, and the rich mythology of the British Isles. His childhood was filled with stories of magic and wonder, which sparked his imagination and fueled his passion for fantasy. As a young man, Dr. Fantasy was fascinated by the works of J.R.R. Tolkien, C.S. Lewis, and other great fantasy authors. He spent countless hours reading, writing, and exploring the worlds of Middle-earth, Narnia, and other fantastical realms. This early love of fantasy laid the foundation for his later work in neuroscience, as he began to explore the neural correlates of imagination and creativity.

**The Dr. Fantasy Institute**

The Dr. Fantasy Institute is a state-of-the-art research facility dedicated to the study of fantasy neuroscience. Located in a gleaming tower of glass and steel, the institute is a hub of creative energy and innovative thinking. Here, Dr. Fantasy and his team of researchers, engineers, and artists work together to push the boundaries of what is possible in the human mind. The institute is equipped with cutting-edge technology, including advanced brain-computer interfaces, neurostimulation devices, and virtual reality platforms. These tools enable Dr. Fantasy and his team to explore the neural correlates of imagination, creativity, and fantasy, and to develop new technologies that can enhance and transform the human experience. The institute is also a center for artistic expression and creativity, with a thriving community of writers, artists, and musicians who come to share their work and collaborate with Dr. Fantasy and his team. The institute hosts regular exhibitions, performances, and workshops, showcasing the latest innovations in fantasy neuroscience and celebrating the boundless potential of the human imagination.

**Future Plans for Dr. Fantasy**

Dr. Fantasy is always looking to the future, seeking new ways to explore the frontiers of the human mind and to unlock the secrets of fantasy neuroscience. He is currently working on a top-secret project, codenamed "Elysium," which promises to revolutionize the field of virtual reality and fantasy entertainment. In the years ahead, Dr. Fantasy plans to expand his institute, establishing new research centers and collaborations around the world. He will continue to push the boundaries of what is possible, exploring new frontiers in fantasy neuroscience and inspiring others to join him on this journey of discovery. As Dr. Fantasy often says, "The future of fantasy is not just about technology or science – it's about the boundless potential of the human imagination. We are the dreamers, and the dreamers are us."