KNOWLEDGE AND CAPABILITY TRANSFER THROUGH LARGE LANGUAGE MODELS' PARAMETERS FUSING

Anonymous authors
Paper under double-blind review

ABSTRACT

The post-training phase of large language models (LLMs) plays a pivotal role in refining models to follow instructions and align with human preferences. However, this phase is fraught with challenges, particularly in sourcing high-quality post-training data. This paper introduces a novel approach, termed Parameters Fusing, that simplifies the post-training process by amalgamating model parameters delta from existing instruct-tuned checkpoints with a new base model tailored to specific domain data obtained by continual pre-training. Utilizing open-weight models such as Meta's Llama, our method replicates the effects of the traditional post-training phase while significantly reducing both time and resource costs. This approach not only minimizes the challenges of post-training data acquisition but also provides a flexible and efficient framework for enhancing LLMs with domain-specific knowledge or capabilities.

1 INTRODUCTION

The creation of contemporary language foundation models involves a two-step process. The first step is the pre-training phase, where the model undergoes extensive training using simple tasks such as predicting the next token or generating captions. The second step is the post-training phase, where the model is fine-tuned to follow instructions, align with human preferences, and enhance specific skills (e.g., coding, tool-use, reasoning).

Post-training of large language models (LLMs) can be challenging due to their massive size and complexity: 1) the biggest difficulty is the need for large amounts of high quality instruction aligned training data, which can be time-consuming and expensive to obtain; 2) additionally, LLMs may not always generalize well to new tasks or domains, requiring sophisticated fine-tuning or domain adaptation; 3) another challenge is the risk of over-fitting (Chang et al., 2024; Kaplan et al., 2020; Tirumala et al., 2022), where the model becomes too specialized to the training data and fails to generalize to new inputs; 4) finally, there are also concerns about the safety implications of LLMs when they are not trained properly, which perpetuate biases and stereotypes present in the training data (1).

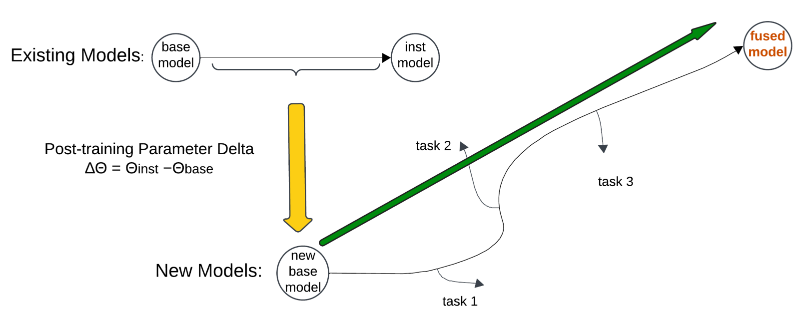


Figure 1: Parameter Delta Fusing: knowledge and capability transfer with no training cost

Drawing inspiration from seminal works 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), in this study, we explore the correlation between alterations in model parameters and model performance. Our findings suggest that enhancements in knowledge and capabilities can be achieved not only through post-training, but also simply adding the pre-existing models' parameters delta (derived by subtracting pre-trained checkpoints from the instruct-tuned checkpoints).

The contributions of this study are articulated as follows:

- We introduce the concept of the Model Parameters Space and demonstrate that parameter deltas can be added, subtracted, and transferred within homologous large language models that share identical architectural frameworks.
- We report the discovery of a concave relationship between variations in model parameters and their corresponding performances. We further explore and articulate the underlying reasons and hypotheses for this concaveness.
- We present a novel method termed Parameters Fusing, which effectively replicates the benefits of post-training without undergoing the entire post-training phase. We also illustrate the feasibility of enhancing the performance of a post-trained model by strategically adjusting parameter deltas derived from various existing checkpoints, leveraging the identified concaveness property.

2 Hypothesis and Theories

Building upon our foundational hypothesis, we undertake a step-by-step analysis to demonstrate the feasibility of transferring knowledge and capabilities through model parameter operations. In this section, we aim to provide a rigorous and systematic exploration of the underlying mechanisms that enable the transfer of knowledge and capabilities across models' parameters.

2.1 The Relationship Between The Change of Model Parameters and Amount of Training Received

Definition 2.1. Amount of Training Received is a conceptual metric, yet it can be quantified through various indicators such as the number of training steps completed, the volume of data (e.g., number of tokens) on which the model has been trained, or the computational resources expended (e.g., training FLOPs). We denote $\mathbb{T}$ as the amount of training received and $\mathbb{T}_i$ as the amount of training received at step i.

Conjecture 2.1. *The change of model parameters reflects the knowledge acquired through the amount of training received.*

Conventionally, the initialization of a model's parameters is conducted either randomly or based on a pre-trained checkpoint, with the model possessing no prior knowledge of the data it is tasked to learn. As the training progresses, the model increasingly processes data, enabling it to discern patterns and relationships within the dataset. This acquisition of knowledge is manifested in the modifications to the model's parameters, which become progressively refined and precise throughout the training duration. Consequently, the parameters are updated to encapsulate this newly acquired knowledge, proportionate to the extent of training received.

Conjecture 2.2. *The change of model parameters is a bijection relationship to the amount of training received.*

The bijection relationship is upheld due to each unique set of parameters changing corresponding to a distinct level of training, provided that there is no stochastic variability during the training process and all exogenous factors, like training hyper-parameters, remain constant. Consequently, given knowledge of the quantity of training data utilized and the constancy of all training hyper-parameters, it is possible to ascertain the specific set of model parameters that were updated to during the training phase.

2.2 Definitions of Model Parameters Space

Model parameters are pivotal in determining how a model processes input data and subsequently makes predictions or decisions. The Model Parameters Space can be defined as a representation of the underlying knowledge that the model has acquired through the training it has undergone.

Definition 2.2. Model Parameters Space refers to a high-dimensional space composed of the sets of parameters associated with homologous models, consistent within the same model architecture. We denote Θ as the model parameters. The homologous models encompasses consistent model configuration parameters, such as the number of layers, number of heads, and number of dimensions, as well as the tokenizer.

The model parameters space and text embedding space (Aggarwal & Zhai, 2012; Angelov, 2020; Church, 2017) can be considered analogous. Text embedding space is a high-dimensional domain where words, phrases, or entire documents are represented as dense vectors. These vectors encapsulate the semantic meanings and relationships among the input data, thereby facilitating the model's execution of various natural language processing tasks. Similarly, just as the embedding space conveys the meanings and relationships inherent in text-based data, the model parameters space embodies the model's comprehension of the tasks or the knowledge it has accrued through the training it has received. For a large language model, its model parameters encapsulates the knowledge it has assimilated, while the entirety of the parameters space represents the full spectrum of potential world knowledge that can be learned.

2.3 Additivity, Subtractivity and Transferability in Model parameters Space

Proposition 2.1. *The parameters of the language model are additive, subtractive, and transferable within its model parameters space, reflecting the corresponding amount of training involved in these operations.*

Our preceding hypothesis can be reformulated as $\Theta_j - \Theta_i \longleftrightarrow \mathbb{T}_j - \mathbb{T}_i$, where $\longleftrightarrow$ denotes both the reflection of the knowledge acquired and the bijection relationship. When $i = 0$, the model is initialized randomly, embodying a state of zero knowledge.

Additivity: Building upon the concept of model weights averaging(Izmailov et al., 2018; Nikishin et al., 2018; Rame et al., 2022; Su & Chen, 2015; Wortsman et al., 2022a), the parameters of a language model can be aggregated and averaged to forge a new set of parameters that encapsulate the collective knowledge from multiple training checkpoints. This relationship is mathematically represented as $\Theta_i + \Theta_j \longleftrightarrow \mathbb{T}_i + \mathbb{T}_j$.

Subtractability: In a similar vein, the parameters of a language model can be subtracted from one another to yield a new set of parameters that delineate the differential knowledge between two models. This is expressed as $\Theta_j - \Theta_i \longleftrightarrow \mathbb{T}_j - \mathbb{T}_i$.

Transferability: The knowledge encapsulated in a language model checkpoint can be transferred to another model within the same parameters space, enabling the recipient model to assimilate the knowledge and capabilities of the original model. This process is depicted as $\Theta_k + (\Theta_j - \Theta_i) \longleftrightarrow \mathbb{T}_k + (\mathbb{T}_j - \mathbb{T}_i)$. Transfer learning, a specific instance of transferability, can be formulated as $\Theta_k + (\Theta_j - \Theta_0) \longleftrightarrow \mathbb{T}_k + (\mathbb{T}_j - \mathbb{T}_0)$.

The subsequent definitions are derived based on the properties inherent to the model parameters space:

Definition 2.3. Model Parameters Delta is the subtraction of model parameters between two checkpoints within the same model parameters space. Mathematically, it is expressed as $\Delta\Theta_{ji} = \Theta_j - \Theta_i$. 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).

Definition 2.4. Parameters Fusing refers to the process of executing operations on parameters within the same model parameters space, with the objective of synthesizing an integrated model. These operations typically involve adding, subtracting and transferring the model parameters or model parameters delta. This 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).

Definition 2.5. Fused Model refers to a model characterized by the integration of parameters or model parameters delta from existing models, rather than being directly trained through data.

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

Notation	Description
Θ	model parameters
Θ_0, Θ_{base}, Θ_{pretrain}	pre-trained or base model parameters
Θ_{inst}, $\Theta_{\text{post-train}}$	post-trained or instruction-finetuned model parameters
Θ_{cpt}	continual pre-trained model parameters, it's also a base model
$\Delta\Theta$	parameter delta between a post-trained model and a pre-trained model
$\mathbb{T}$	amount of training received
α	scaling factor of parameter delta $\Delta\Theta$

2.4 Model Performance is Concave to the Change of Model Parameters

Conjecture 2.3. *The relationship between model performance and the amount of training received is characterized as a concave function.*

A concave function is one that is curved inward, where the rate of change of the function's slope is negative. Typically, model performance exhibits concavity relative to the amount of training received. This implies that as training progresses, model performance initially improves rapidly but eventually reaches a plateau and may even decline. This pattern occurs because the model quickly assimilates new information from the data at the outset but gradually approaches a saturation point, beyond which it may overfit if training persists excessively.

Proposition 2.2. *The performance of the language model is concave with respect to changes in the parameters within the model parameters space. Specifically, $f((1-\alpha)\Theta_i+\alpha\Theta_j) \geq (1-\alpha)f(\Theta_i)+\alpha f(\Theta_j)$ where $f(*)$ denotes the model performance.*

Given that model performance exhibits concavity with respect to the amount of training received, and considering our hypothesis that changes in model parameters reflect the knowledge acquired through training, it follows logically that language model performance is concave relative to changes in the parameters. This pattern will be substantiated through our experimental observations. Additionally, the following corollary can be logically inferred:

Conjecture 2.4. *The performance derived from averaging model parameters across several checkpoints will surpass the performance obtained by averaging the model outputs from those same checkpoints.*

Izmailov et al. (2018); Rame et al. (2022); Wortsman et al. (2022a) have also documented analogous outcomes in their research, highlighting enhancements in model accuracy, improved generalization capabilities, and increased stability and robustness, when implementing the model parameters averaging. This assertion supports the hypothesis that the fusion of model parameters typically results in equivalent, if not better, model performance. Additionally, we experiment to demonstrate the validity of this conjecture.

2.5 Knowledge and Capability Transfer Through Adding Existing Models' Parameters Delta

Drawing from Proposition 2.3, it is logically deduced that the manipulation of model parameters facilitates the effective transfer of knowledge and capabilities. This assertion represents the core claim and key contribution of this paper.

Proposition 2.3. *The language model can assimilate the knowledge and capabilities typical of the post-training stage by fusing the model parameters delta between an instruct-tuned checkpoint and a pre-trained checkpoint.*

Conventionally, the domain expansion a large language model involves initially continuing to train the pre-trained checkpoint through a next token prediction task (Ke et al., 2023) to acquire domain-specific knowledge (e.g., coding, mathematics, tool calling, or other domain-specific data). This is followed by a standard post-training process, which includes Supervised Fine-Tuning (SFT) (Brown, 2020; Dubey et al., 2024; Touvron et al., 2023) and either Direct Preference Optimization

(DPO)(Dubey et al., 2024; Rafailov et al., 2024; Touvron et al., 2023) or Proximal Policy Optimization (PPO) (Ouyang et al., 2022; Schulman et al., 2017). Our proposed approach bypasses the standard post-training process and instead fuses the available post-trained model parameters into pre-trained model. Mathematically, this is represented as

$$\Theta_{\text{cpt}} + \Delta\Theta = \Theta_{\text{cpt}} + (\Theta_{\text{inst}} - \Theta_{\text{base}}) \longleftrightarrow \mathbb{T}_{\text{cpt}} + (\mathbb{T}_{\text{inst}} - \mathbb{T}_{\text{base}}) \tag{1}$$

where Θ_{cpt} is from the newly continual-pre-trained base model, Θ_{inst} and Θ_{base} are from pre-existing models, $\mathbb{T}_{\text{inst}} - \mathbb{T}_{\text{base}}$ symbolizes the post-training process, and $f(\Theta_{\text{cpt}} + \Delta\Theta)$ yields legitimate outputs (i.e., the model performs effectively). We validate the efficacy of this approach will through experimental observations.

3 Experiments

3.1 Experiments Setup

We utilize open-weights checkpoints of Llama3 and Llama3.1 (Dubey et al., 2024), specifically employing the 8B and 70B model architectures. The 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).

3.2 Model Performance Concaveness Validation

3.2.1 Model Performance Concaveness on the Change of Model Parameters

In the initial experiment, we assess the performance of fused models with varying parameters to elucidate the concave relationship (referenced in Conjecture 2.4) between model performance and changes in parameters. The experiments are conducted using both Llama3.1-8b and Llama3.1-70b models. Each interim fused checkpoint is defined as $\Theta_i = \Theta_0 + \alpha\Delta\Theta$, where $\alpha \in \{0, 0.1, 0.2, \ldots, 2\}$. Θ_0 is the Llama3.1-base model and $\Delta\Theta$ is the parameter delta between Llama3.1-inst and Llama3.1-base model. When $\alpha = 0$, the checkpoint corresponds to the pre-trained model; at $\alpha = 1.0$, it aligns with the instruct model checkpoint.

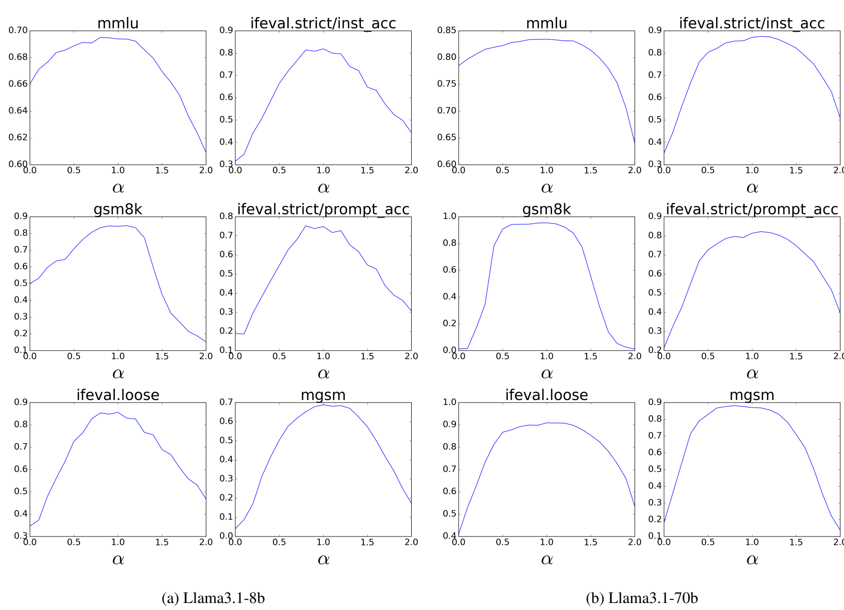
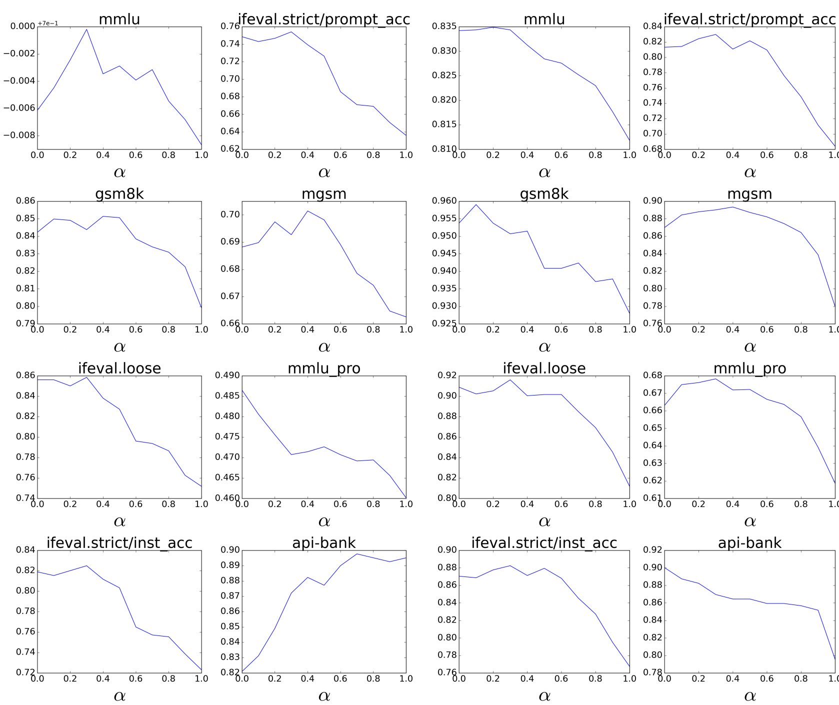


Figure 2: Concave shape of model performance to different scale of model parameters delta

Figure 2 presents a selected set of concave curves illustrating the relationship between model performance and changes in model parameters. All depicted curves demonstrate legitimate responses and exhibit concave relationships, corroborating our hypothesis. Notably, the apex of performance is consistently observed around $\alpha = 1$, suggesting that the publicly released Llama3.1 models are in their optimal states for each respective model architecture. Additional examples of concave curves can be found in Figures 5 and 6, all of which further confirm the concave relationships.

3.2.2 Model Performance Concaveness When Fusing Other Homologous Models in the Open Weight Community

In this section, we expand our investigation to include Llama-homologous models that are available to the community. Specifically, we incorporate the Llama3.1-8b-style model, Llama3.1-8B-Chinese-Chat (Wang et al., 2024), and the Llama3.1-70b-style model, Reflection-Llama-3.1-70B (Shumer, 2024), both sourced from HuggingFace. Our objective is to evaluate the performance implications of varying degrees of integration between the native Llama3.1 models and analogous community-developed models, thereby testing the concave relationship posited in Conjecture 2.4. The methodology for creating each fused model checkpoint involves calculating the weighted average of the model parameters from the Llama3.1-inst checkpoint and the corresponding community model checkpoint. The formula used is as follows: $\Theta_i = \Theta_0 + \alpha\Delta\Theta^* + (1-\alpha)\Delta\Theta$, where $\Delta\Theta^*$ is from the community model and $\Delta\Theta$ is from the vanilla Llama3.1 model, $\alpha \in \{0, 0.1, 0.2, \ldots, 1\}$. At $\alpha = 0$, the model strictly represents the official Llama3.1-inst model, whereas at $\alpha = 1$, it fully represents the community instruct model checkpoint.



(a) Llama3.1-8b different scale of fusion

(b) Llama3.1-70b different scale of fusion

Figure 3: Concave shape of model performance to different scale of model fusion between Llama3.1-inst checkpoints and Hugging Face open-weight checkpoints

Figure 3 presents a series of concave curves that illustrate the relationship between model performance and varying levels of parameter fusion. The analysis of these curves leads to two critical observations: 1) all curves demonstrate legitimate responses validate the presence of concave relationships, supporting the initial conjecture. 2) the concavity property indicates that some fused checkpoints outperform both the pure Llama3.1-inst model ($\alpha = 0$) and the standalone community model ($\alpha = 1$). This is noteworthy, especially considering that the performance of community models generally fall short of the Meta-released Llama-inst models on certain benchmarks.

3.3 Knowledge and Capability Acquisition via Post-Training Model Parameters Fusing

The experiment in the last sub-section demonstrated the legitimacy of responses generated from checkpoints with model parameter operations. The experiment in this subsection aims to show that knowledge gained through post-training can also be acquired by adding the $\Delta\Theta$ of pre-existing checkpoints to a new base model Θ_{base}.

3.3.1 Bypassing Post-Training Through Model Parameters Fusing on Continual-Pretrained Checkpoints

To contextualize our research within a practical framework, we designed an experiment that extends the pre-training phase by employing the next-token prediction task on a newly generated document (refer to Document A.5) that could not have been encountered by the Llama 3.1 pre-trained model during its initial pre-training phase. We use lr=$1e^{-5}$, batch_size=1, seq_len=512, steps=125, 8 H100 GPUs to continually pre-train the 8B model; and lr=$1e^{-5}$, batch_size=1, seq_len=512, steps=60, 16 H100 GPUs to continually pre-train the 70B model. Subsequently, we integrate the $\Delta\Theta$ from the prior post-training phase, into the newly updated continual-pre-trained model checkpoint Θ_{cpt} to produce an equivalent instruct-model ($\Theta = \Theta_{\text{cpt}} + \Delta\Theta$). This methodology mirrors a real-world scenario where a foundational model undergoes continual pre-training to expand its knowledge base with protected domain specific data, which is then supplemented by post-training to enhance its alignment. We then evaluate this "CPT-fused-inst" model's performance in terms of both domain knowledge acquisition and its ability to follow instructions (on standard benchmarks). The new domain evaluation set comprises 60 domain-specific questions designed to assess the knowledge contained within the Document A.5 undergoing continual pre-training. 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 accuracy of each response generated by the fused models and the vanilla Llama models.

Category	Benchmark	llama3.1-8b-inst	CPT-fused-8b-inst	llama3.1-70b-inst	CPT-fused-70b-inst
General	MMLU	0.6939	0.6926	0.8342	0.8256
	MMLU PRO	0.4865	0.4764	0.6631	0.6516
	IFEval	0.7154	0.7385	0.8769	0.9077
Code	HumanEval	0.6951	0.6037	0.8049	0.7927
	MBPP EvalPlus	0.7037	0.7143	0.8492	0.8255
Math	GSM8K	0.8423	0.8302	0.9538	0.9522
Reasoning	ARC Challenge	0.8369	0.8395	0.9468	0.9433
	GPQA	0.2969	0.2924	0.4665	0.4263
Tool use	BFCL	0.6790	0.5651	0.7570	0.7126
Multilingual	MGSM	0.6882	0.6156	0.8700	0.8695
New domain		0.0000	0.7667	0.0000	0.7667

Table 1: Performance comparison between Llama-CPT-fused models and original Llama-inst models

The parameters fusing can be deemed successful based on the following observations: 1) as shown in Table 1, the CPT-fused-inst 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 during the continual pre-training phase. 2) furthermore, as depicted in , the performance of the Llama-CPT-fused-inst models closely align with that of the original Llama-inst models. Notably, there is a slight decline in performance in certain benchmarks, primarily attributable to the transfer efficiency (to be discussed in Section 3.4), potential variability in the evaluation setup, or lack of annealing in our continual pre-training phrase. These findings underscore that the incorporation of $\Delta\Theta$ enables the bypassing of conventional post-training approaches while maintaining equivalent performance levels.

3.3.2 Further Studies on Llama3 and Llama3.1 Models Parameters Fusing

To further investigate the dynamics of parameter fusion within the homologous model parameter space, we consider two scenarios: (1) adding the $\Delta\Theta$ of Llama3 to Llama3.1-base model (denoted as $\Theta^{3.1}_{\text{base}} + \Delta\Theta^{3}$), and (2) adding the $\Delta\Theta$ of Llama3.1 to the base checkpoint of Llama3 (denoted as $\Theta^{3}_{\text{base}} + \Delta\Theta^{3.1}$). For the evaluation results, we categorized them two three groups: 1) General: common evaluation sets that both Llama3 and Llama3.1 perform well on; 2) Tool use: according to the Llama3 paper(Dubey et al., 2024), tool use knowledge is primarily learned during the post-training stage, an area where pre-trained models perform poorly.

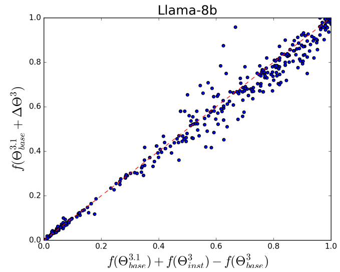
						Fused Models	
Category	**Benchmark**	$\Theta_{\text{Llama3-base}}$	$\Theta_{\text{Llama3.1-base}}$	$\Theta_{\text{Llama3-inst}}$	$\Theta_{\text{Llama3.1-inst}}$	$\Theta^{3.1}_{\text{base}}+\Delta\Theta^{3}$	$\Theta^{3}_{\text{base}}+\Delta\Theta^{3.1}$
8B							
General	MMLU	0.6161	0.6603	0.6848	0.6939	0.6863	0.6799
	MMLU PRO	0.3300	0.3637	0.4546	0.4865	0.4551	0.4394
	IFEval	0.4000	0.3231	0.6616	0.7154	0.7231	0.7154
Code	HumanEval	0.3049	0.2744	0.6220	0.6951	0.6463	0.6585
	MBPP EvalPlus	0.5900	0.5926	0.7063	0.7037	0.7328	0.7169
Math	GSM8K	0.5026	0.5004	0.8143	0.8423	0.7847	0.8218
	MATH	0.1044	0.1128	0.2760	0.4970	0.2888	0.4618
Reasoning	ARC Challenge	0.6258	0.6532	0.8206	0.8369	0.8275	0.8309
	GPQA	0.0513	0.0625	0.3281	0.2969	0.3058	0.2746
Tool use	BFCL	-	-	0.6010	0.6790	0.6087	0.6384
	API Bank	0.2532	0.2481	0.4885	0.8210	0.5192	0.8082
Multilingual	MGSM	0.0227	0.0399	0.6085	0.6882	0.6033	0.6400
70B							
General	MMLU	0.7878	0.7846	0.8198	0.8342	0.8167	0.8291
	MMLU PRO	0.5399	0.5126	0.6321	0.6631	0.6205	0.6546
	IFEval	0.6692	0.6615	0.8154	0.8769	0.8385	0.9231
Code	HumanEval	0.3963	0.3902	0.8049	0.8049	0.7866	0.8049
	MBPP EvalPlus	0.6693	0.7037	0.8280	0.8492	0.8069	0.8466
Math	GSM8K	0.0440	0.0129	0.9333	0.9538	0.9227	0.9530
	MATH	0.2752	0.1624	0.4996	0.6624	0.4984	0.6402
Reasoning	ARC Challenge	0.8773	0.8893	0.9425	0.9468	0.9425	0.9442
	GPQA	0.1429	0.2277	0.4062	0.4665	0.4219	0.4487
Tool use	BFCL	-	-	0.7683	0.7796	0.7773	0.7665
	API Bank	0.3785	0.1330	0.8517	0.9003	0.8286	0.9003
Multilingual	MGSM	0.0937	0.1830	0.8405	0.8700	0.8405	0.8624

Table 2: Knowledge and capability transfer through post-training parameters fusing on Llama3-8b and Llama3-70b models

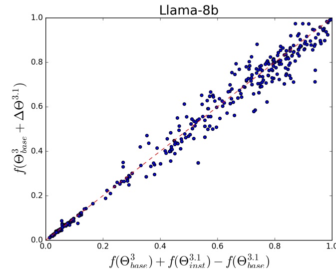
Tables 2 presents the performance of Llama-inst, Llama-pretrain, and fused models on various evaluation sets. Our observations are as follows: 1) all fused models ($\Theta_{\text{base}} + \Delta\Theta$) produce legitimate outputs; 2) for both 8b and 70b models, $\Theta^{3.1}_{\text{base}} + \Delta\Theta^{3}$ checkpoints exhibit performance similar to that of the $\Theta_{\text{Llama3.1-inst}}$ checkpoints, but with slightly lower scores. This indicates that the knowledge and capabilities gained from post-training of one model (Llama 3) can be transferred to a second model (Llama3.1) by adding the parameters delta of first model to the parameters of the second model and also aligns with the fact that Llama3.1 received more training during its post-training phase compared to Llama3; 3) for 70b model, the $\Theta^{3}_{\text{base}} + \Delta\Theta^{3.1}$ checkpoint demonstrates performance similar to, or slightly better than, the $\Theta_{\text{Llama3-inst}}$ checkpoints (although this pattern is not clear for the 8b model). This observation substantiates the transfer and alignment of knowledge and capabilities between the two models, indicating that Llama3.1 underwent a more extensive post-training phase compared to Llama3; 4) in the context of tool use, it is evident that the capability to utilize tools can be acquired by adding $\Delta\Theta$ to base models' parameters, as this tool calling capability is typically only gained during the post-training phase.

3.4 Quantitative Analysis on the Effectiveness and Efficiency of Transfer

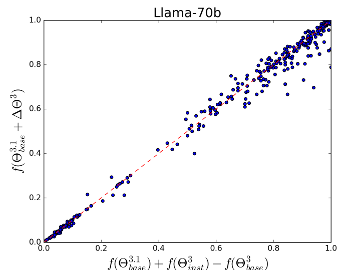
We further conduct a quantitative study to examine the performance of fused models, so as to estimate the effectiveness and efficiency of knowledge and capability transfer. We denoted the model performance of a checkpoint as $f(*)$ and represented the real performance of a fused checkpoint as $f(\Theta^{i}_{\text{base}} + \Delta\Theta^{j})$. 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 the fused model is represented mathematically as $f(\Theta^{i}_{\text{base}}) + f(\Theta^{j}_{\text{inst}}) - f(\Theta^{j}_{\text{base}})$.



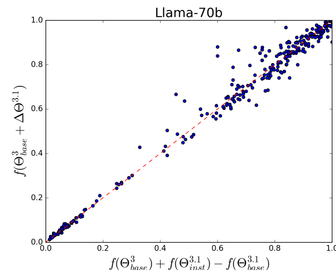
(a) $\Theta^{3.1}_{\text{base}}+\Delta\Theta^{3}$, Llama-8b
$\beta = 0.9786$, $R^2 = 0.9921$



(b) $\Theta^{3}_{\text{base}}+\Delta\Theta^{3.1}$, Llama-8b
$\beta = 0.9845$, $R^2 = 0.9915$



(c) $\Theta^{3.1}_{\text{base}}+\Delta\Theta^{3}$, Llama-70b
$\beta = 0.9805$, $R^2 = 0.9950$



(d) $\Theta^{3}_{\text{base}}+\Delta\Theta^{3.1}$, Llama-70b
$\beta = 0.9986$, $R^2 = 0.9943$

Figure 4: Relationship between the real performance of fused 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 fused models

The hypothetical model performance demonstrates a high degree of explanatory power, accounting for over 99.1% 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 fused 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 parameters delta from the first model to the parameters of the second model, with minimal loss or distortion. This is evidenced by the regression coefficient β being close to 100%, indicating a near-perfect linear relationship between the hypothetical and actual performances. Each subfigure in Figure 4 contains more than 500 data points, with each point representing an evaluation metric. The parameter β may also be interpreted as the coefficient of transfer efficiency.

4 Related Work

4.1 Large Language Model Post-training

Large language model post-training often consists of supervised finetuning (SFT)(Brown, 2020; Dubey et al., 2024; Touvron et al., 2023) and follow-up alignment such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Direct Preference Optimization (DPO) (Rafailov et al., 2024). GPT-series models (Brown, 2020; Ouyang et al., 2022) are using PPO according to their latest released tech report while Llama-series models are using DPO (Dubey et al., 2024; Touvron et al., 2023) according to their paper. Both PPO and DPO incorporate human preference data to LLM for model alignment.

Post-training can be a challenging process, as the size and quality of the dataset used can significantly impact the effectiveness and difficulty of post-training. Overfitting is a common phenomenon observed in large language model training, which can lead to reduced generalization performance and decreased robustness (Tirumala et al., 2022). Therefore, it is crucial to carefully select and preprocess the dataset (Dubey et al., 2024; Touvron et al., 2023), as well as employ appropriate regularization techniques to mitigate overfitting.

4.2 Parameter Space, Model Weights Averaging and Model Merging

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. Building upon these findings, our study expands the concept of model parameters space and hypothesizes that the variation in the model parameters are reflective of, and proportional to, the amount of training received.

Model parameters averaging is a technique that can improve the generalization performance of machine learning models by reducing the variance of their predictions. Izmailov et al. (2018) introduced Stochastic Weight Averaging (SWA) that finds much flatter solutions than Stochastic Gradient Descent (SGD) and achieves significant improvement in test accuracy. 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, transcending mere 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 fusing.

We expand the concept of additivity in model parameters to encompass subtractivity and transferability, enabling operations across various model checkpoints within the same parameters space. This enhancement allows for more flexible manipulation of model parameters. Furthermore, we elucidate the principles underlying model merging and parameter fusion by correlating them with the extent of training reflected in parameter changes. Additionally, we establish a concave relationship between model performance and parameter changes. Ultimately, we empirically demonstrate that traditional post-training processes can be effectively substituted with direct parameter fusion.

5 Conclusions

In this paper, we have discussed and shown that the change of model parameters carries the amount of training it has received, and therefore the language model can gain the knowledge and capability of the post-training stage from the model parameters delta between instruct-tuned checkpoint and pre-trained checkpoint. As a result, we present a novel approach for knowledge transfer through large language models' parameters fusing by incorporating parameters delta (derived from subtracting the pre-trained checkpoints from the instruct-tuned checkpoints from open-weight models such as Meta's Llama) into pre-trained checkpoints. This approach can effectively bypass the entire post-training process, while achieving the similar performance, therefore significantly reducing data annotation and training costs in terms of money, time, and compute.

We believe that our approach has the potential to advance the field of natural language processing by making it easier and more efficient to train and deploy large language models. We hope that our work will inspire further research in this area and lead to the development of even more advanced techniques for knowledge transfer through large language models' parameters fusion.

REFERENCES

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.

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.

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.

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.

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.

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.

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.

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.

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.

Matt Shumer. Reflection-llama-3.1-70b, 2024. URL `https://huggingface.co/mattshumer/Reflection-Llama-3.1-70B`.

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

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.

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. Llama3.1-8b-chinese-chat, 2024. URL `https://huggingface.co/shenzhi-wang/Llama3.1-8B-Chinese-Chat`.

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.

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 Impact Statement

The research presented in this paper has significant implications for the field of artificial intelligence, particularly in the development and optimization of large language models (LLMs). By introducing a novel methodology for parameter fusion, the study addresses the complex and resource-intensive phase of post-training LLMs. This innovative approach not only replicates the benefits traditionally achieved through post-training but also substantially reduces associated costs and time expenditures. Furthermore, it simplifies the process of expanding the model's domain knowledge and capabilities, such as tool-use, coding proficiency, and tonal qualities by allowing for the strategic combination of parameter deltas from various checkpoints, including those from open-weight models. This aspect is particularly impactful for the community, as it democratizes access to advanced model tuning techniques, enabling a broader base of researchers and developers to contribute to and benefit from state-of-the-art LLMs. The flexibility in model customization could lead to more tailored and efficient AI systems. The potential for this method to streamline the refinement of LLMs and reduce dependency on extensive computational resources could mark a transformative shift in how these models are developed, making advanced AI more accessible and sustainable. This could ultimately accelerate innovation and broaden the application of LLMs across different sectors, with the open-weight community playing a pivotal role in this evolution.

A.2 Comparison Between Fused Model and LoRA Models

We selected two LoRA models from the open-weight community on Hugging Face, specifically trained on Llama3-8b models: Llama3-lora-1 (SkyOrbis/SKY-Ko-Llama3-8B-lora) and Llama3-lora-2 (MadMarx37/llama3-8b-alpaca-lora-peft). These were compared to the Llama3-fused model, a parameter-delta-fused model based on Llama3-8b, with parameter deltas derived from the Llama3.1 model. Our findings indicate that the Llama3-fused model significantly outperforms the LoRA models, achieving superior results with zero training cost. This further demonstrates the resource efficiency and high performance of our method. The weak results of Llama3-lora-2 (especially in GSM8K and GPQA) are due to illegitimate responses and repeated tokens.

Metric Key	Llama3-lora-1	Llama3-lora-2	Llama3-fused
MMLU	0.6447	0.5826	0.6800
MML PRO	0.3725	0.2639	0.4394
IF EVAL	0.4462	0.2846	0.7154
GSM8K	0.6641	0.0000	0.8218
ARC Challenge	0.7957	0.6678	0.8309
GPQA	0.2790	0.0201	0.2746
MGSM	0.3349	0.1185	0.6400
GPU HOUR (A100)	25	-	0

Table 3: Comparison between Fused model and LoRA models: Fused model can achieve higher performance than LoRA models with zero training cost.

A.3 Concaveness on Other Performance Metrics

Figure 5 6 show more selected concave shapes of model performance to change of model parameters.

A.4 Fused Models With Continual-Pretrained Checkpoints Do Have Domain Knowledge and Instruction Following Capability

Query	llama3.1-8b-inst	llama3.1-8b-inst + document	llama3.1-8b-cpt-fused-inst
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-fused models and original Llama-inst models on 8b

Query	llama3.1-70b-inst	llama3.1-70b-inst + document	llama3.1-70b-CPT-fused-inst
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-fused models and original Llama-inst models on 70b

A.5 Pseudo biology document for large language model continual pre-training 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

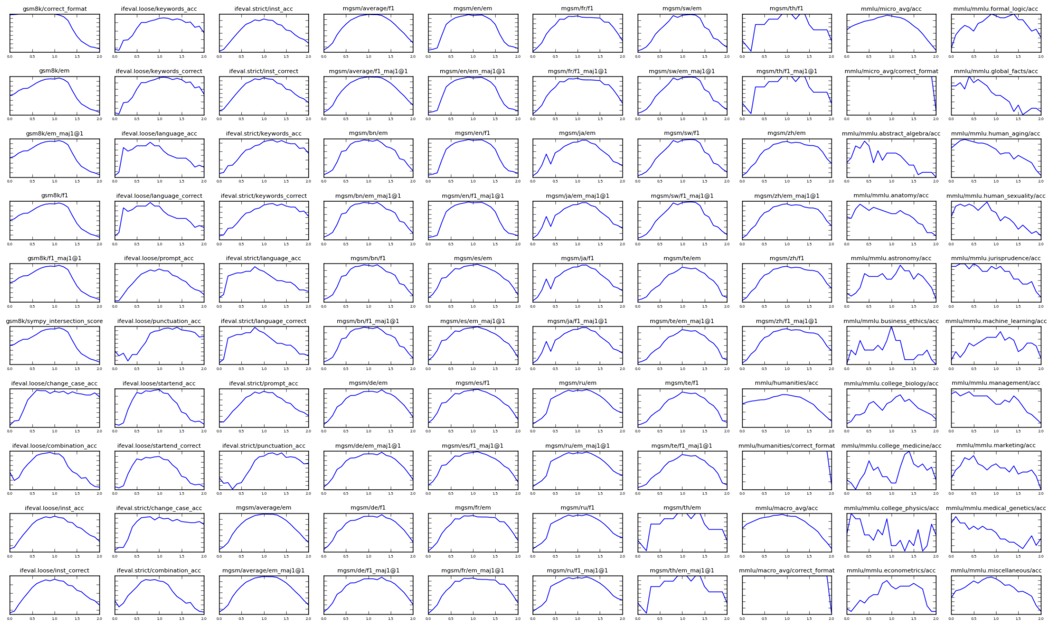


Figure 5: Concave shapes of model performance to scale of model parameters delta - Llama3.1-8b

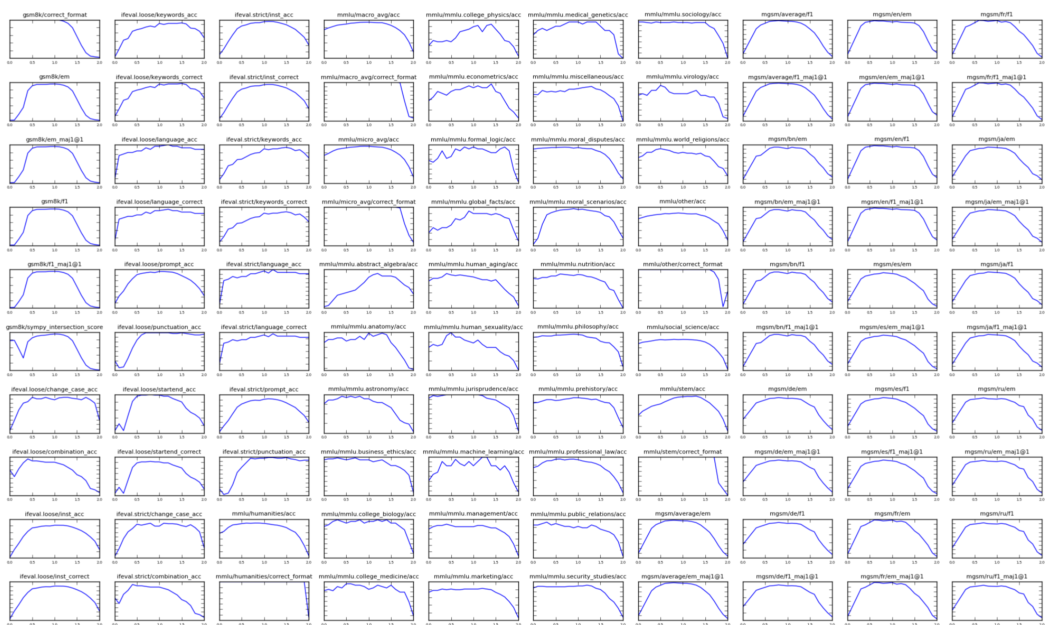


Figure 6: Concave shapes of model performance to scale of model parameters delta - Llama3.1-70b

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.”