

Shadow-FT: Tuning Instruct via Base

Taiqiang Wu^{1*} Runming Yang^{2,3*}

Jiayi Li² Pengfei Hu³ Ngai Wong¹ Yujia Yang^{2†}

¹The University of Hong Kong ²Tsinghua University ³Tencent
takiwu@connect.hku.hk yang.yujia@sz.tsinghua.edu.cn

Abstract

Large language models (LLMs) consistently benefit from further fine-tuning on various tasks. However, we observe that directly tuning the INSTRUCT (i.e., instruction tuned) models often leads to marginal improvements and even performance degeneration. Notably, paired BASE models, the foundation for these INSTRUCT variants, contain highly similar weight values (i.e., less than 2% on average for Llama 3.1 8B). Therefore, we propose a novel **Shadow-FT** framework to tune the INSTRUCT models by leveraging corresponding BASE models. The key insight is to fine-tune the BASE model, and then *directly* graft the learned weight updates to the INSTRUCT model. Our proposed Shadow-FT introduces no additional parameters, is easy to implement, and significantly improves performance. We conduct extensive experiments on tuning mainstream LLMs, such as Qwen 3 and Llama 3 series, and evaluate them across 19 benchmarks covering coding, reasoning, and mathematical tasks. Experimental results demonstrate that Shadow-FT consistently outperforms conventional full-parameter and parameter-efficient tuning approaches. Further analyses indicate that Shadow-FT can be applied to multimodal large language models (MLLMs) and combined with direct preference optimization (DPO). Codes and weights are available at Github.

1 Introduction

Large Language Models (LLMs), such as Qwen [2], Llama [1], and Gemma [3], have demonstrated remarkable performance across diverse disciplines [4, 7]. Such a strong capability is always attributed to the pre-training on massive data with billions of parameters [9, 10]. When applied in real-world scenarios, there are several challenges. The users want the LLMs to follow their instructions helpfully and honestly [11], which is not covered during the pre-training [4, 8]. Meanwhile, the downstream tasks always involve specific domain knowledge requiring adaptation [12, 13].

To tackle these issues, one predominant approach is further tuning LLMs on desired tasks, including full parameter fine-tuning and parameter-efficient fine-tuning [15, 14]. Typically, for each model size, two paired variants are provided: the pretrained base model (denoted as BASE) and its instruction-tuned version (denoted as INSTRUCT). The BASE model exhibits relatively poor instruction-following ability, while the INSTRUCT model performs better. However, we observe that tuning the INSTRUCT models brings marginal improvements and even a performance degeneration. Therefore, how to tune the INSTRUCT model effectively gains increasing importance.

In this paper, we first analyze the weights of paired BASE and INSTRUCT models considering the relative absolute difference σ . Fortunately, we find that the weights of BASE and INSTRUCT are highly similar. As shown in Figure 1, the gap σ is quite low, such as an average σ of 0.016 for the Llama-3.1-8B model. Intuitively, the contained instruction following ability of INSTRUCT model

*Equal contributions. Work was done when Runming was interning at Tencent.

†Corresponding author.

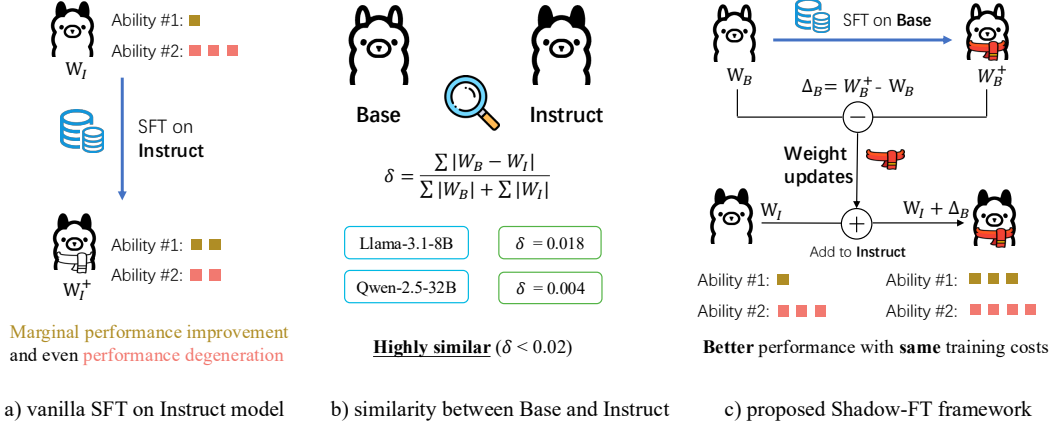


Figure 1: The similarity between BASE and INSTRUCT models (part a), vanilla SFT on INSTRUCT (part b), and the Shadow-FT framework (part c). Based on the observations that the two models are highly similar (gap σ less than 0.02), we propose to tune the paired BASE model and then graft the weight updates onto INSTRUCT model.

might disturb the learning of new knowledge, while BASE can avoid it. Motivated by these, we thus propose a novel **Shadow-FT** framework to employ the BASE model as 'shadow' of INSTRUCT. The key is to tune the BASE for better weight updates and directly graft these updates to INSTRUCT, as they share the same structures.

To evaluate the performance, we conduct extensive experiments tuning mainstream LLMs such as Qwen 3 [2] and Llama 3 [1]. For the tuning data, we employ the BAAI-Infinity-Instruct Dataset³ and extract 2000 samples named as BAAI-2k following [17, 18]. Without the loss of generality, we apply Shadow-FT on full parameter and low-rank settings, and then report the performance on 19 datasets. Experimental results indicate that Shadow-FT consistently outperforms the baselines under various settings, demonstrating the effectiveness and robustness. Further analyses show that Shadow-FT can be applied to MLLMs and combined with DPO for alignment. Our contributions can be concluded as follows:

- We find that paired BASE and INSTRUCT are highly similar considering weight values, and thus propose a novel Shadow-FT framework. The key is to tune the BASE for better weight updates and directly graft these updates to INSTRUCT.
- We conduct extensive experiments tuning various mainstream LLMs and report the performance on 19 benchmarks across math, code, and reasoning. Experimental results demonstrate the effectiveness and robustness of Shadow-FT.
- This work highlights the potential of leveraging BASE models to enhance their INSTRUCT counterparts, and we hope it inspires further research and broader applications in the future.

2 Preliminaries and Motivation

2.1 Background

Basic tuning methods. Supervised Fine-tuning (SFT) is a fundamental approach to updating the knowledge of LLMs. Vanilla SFT methods update all the parameters by gradient descent following $W^+ \leftarrow W + \Delta W$, where $W \in \mathbb{R}^{d_1 \times d_2}$ is an arbitrary weight and W^+ is the updated variant. To reduce the update costs, LoRA [14] introduces a low-rank branch to learn the weight updates following $W^+ \leftarrow W + AB$, where $A \in \mathbb{R}^{d_1 \times r}$, $B \in \mathbb{R}^{r \times d_2}$ and $r \ll \min\{d_1, d_2\}$. The original weight W is frozen during training, and only the low-rank branch is updated.

BASE and INSTRUCT. Current LLMs typically follow a two-stage training pipeline, including pre-training and post-training. During pre-training, LLMs are trained on massive training data on

³<https://huggingface.co/datasets/BAAI/Infinity-Instruct>

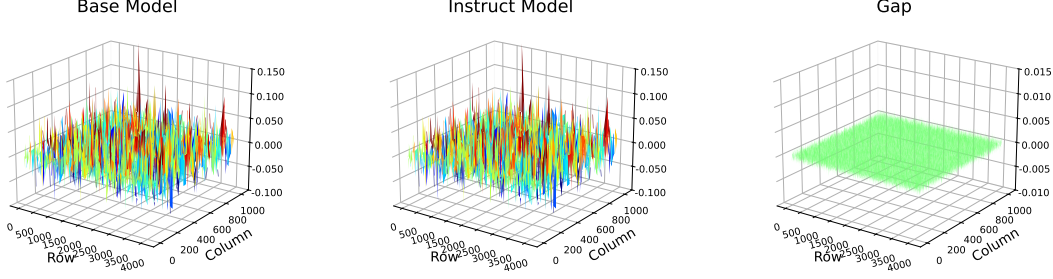


Figure 2: Weight distributions for Llama-3.1-8B. We visualize the same linear layer (layer.0.k_proj) for BASE model (left), INSTRUCT model (middle), and their gap (right). Though zoomed in 10x in the z-axis, the gap is negligible and the average σ value is 0.016.

next token prediction tasks [31], and the weights would be released as BASE version. The INSTRUCT variant, post-trained upon the BASE model, is further aligned with human preference and tuned for reasoning tasks [20]. Therefore, INSTRUCT model performs better than BASE model regarding instruction-following ability.

2.2 Directly Tuning INSTRUCT

However, tuning the INSTRUCT models often leads to marginal improvements and even performance degeneration. Table 1 shows the tuned performance of the INSTRUCT models using the BAAI-2k. We report the average scores of popular benchmarks. Compared to the vanilla INSTRUCT, the tuned version shows marginal improvement, and even degeneration in more cases. Specifically, as shown in Table 1, tuning Qwen-3-4B on BAAI-2k dataset via conventional LoRA would lead to a drop of 2.6 in Math-7 (from 73.8 to 71.2), 6.8 in Code-3 (from 66.4 to 59.6), and 2.6 in Reasoning-9 (from 63.7 to 61.1). Therefore, how to effectively tune INSTRUCT remains a challenge.

2.3 Similar Weights: BASE & INSTRUCT

Fortunately, we observe that the weights of BASE and INSTRUCT are highly similar. To calculate the similarity, we first define the relative gap ratio σ as follows:

$$\sigma = \frac{\sum |W_B - W_I|}{\sum |W_B| + \sum |W_I|}, \quad (1)$$

where \sum is the element-wise sum and $|\cdot|$ means the absolute operations. The σ would be 1 if one is much larger than the other, and be 0 if the two matrices are exactly the same. The smaller the σ , the more similar the two matrices are. Figure 2 shows the weights of the same layer from BASE and INSTRUCT, and also their differences with $\sigma = 0.016$. We can find that the gaps are quite small and negligible after zooming in 10x in the z-axis. Please refer to Appendix A for more σ regarding various LLMs. In summary, these paired BASE and INSTRUCT models are highly similar with $\sigma < 0.02$.

3 Methodology

3.1 Shadow-FT

To tackle the issue that directly tuning INSTRUCT fails, we propose a novel framework, Shadow-FT, to tune the INSTRUCT on BASE. Motivated by the observation that BASE and INSTRUCT models are highly similar, we argue that the weight updates of BASE can be directly added to INSTRUCT. Since they share the same structures, no extra operations are required. Specifically, in Shadow-FT, we first tune the BASE model:

$$W_B^+ \leftarrow \text{Tune}(W_B), \quad (2)$$

where Tune is the fine-tuning method, such as full-parameter fine-tuning and LoRA. After that, we would like to get the weights updates as the learned knowledge using, and directly graft these updates to the INSTRUCT model as:

$$W_I^+ = W_I + (W_B^+ - W_B) = W_I + (\text{Tune}(W_B) - W_B). \quad (3)$$

Traditional tuning on INSTRUCT can be formulated as:

$$W_I^+ = W_I + (W_I^+ - W_I) = W_I + (\text{Tune}(W_I) - W_I). \quad (4)$$

We can find that Shadow-FT introduces no extra training costs. The only difference is the basic weights to learn the weight updates for INSTRUCT model. Vanilla FT methods rely on the INSTRUCT model while Shadow-FT on the BASE model. Since the BASE version is pre-trained only, we believe that the weight updates would be more suitable for modeling the knowledge with less priority, compared to updates of the INSTRUCT version.

3.2 Relation with Task Vectors

Task Vectors aim to represent the ability on tasks as vectors, and are widely used for arithmetic operations on these tasks regarding the same base model [34]. Chat Vector [33] extends such an idea to LLMs, which models weight differences between INSTRUCT and BASE models as vectors and then adds the vectors to continually pretrained BASE models. Specifically, Chat Vector continually pre-trains Llama2 [35] on the Traditional Chinese corpus, and then adds on the chat vectors. Compared to Chat Vector [33], the differences are as follows: 1) *task*: Chat Vector focuses on continual pertaining while Shadow-FT can be applied to board tuning methods, including full-parameter fine-tuning, LoRA, and DPO. 2) *motivation*: Chat Vector aims to extend the language ability. Shadow-FT aims to tackle the degeneration issue based on the weight similarity.

4 Experiments

4.1 Experimental Setup

Training. For the tuning data, we build BAAI-2k by extracting 2000 samples from BAAI-Infinity-Instruct Dataset ⁴ following [17, 18]. We select the samples with high rewards to ensure the data quality and uniform sampling among all categories for data diversity. Without loss of generality, we tune various LLMs, including Qwen 3 series [39] and Llama 3 series [1]. Also, we report the results on Gemma-3 series [3], Yi series [69], and Falcon series [70] in Section 5.5. We employ LLaMA-Factory [36] for the code base and apply two tuning strategies: full-parameter and LoRA. All experiments are conducted on 8 A100 GPUs. Please refer to Appendix B.1 for detailed hyperparameters.

Evaluation. To evaluate the tuned LLMs on downstream benchmarks, we employ the OpenCompass framework [37] and lmdeploy as the acceleration framework [38]. During inference, we set the cutoff length as 4096 and the batch size as 512. Considering the benchmarks, we select three representative abilities, i.e., mathematical, coding, and commonsense reasoning ability, and report the average scores marked as Math-7, Code-3, and Reasoning-9. Specifically, Math-7 denotes the results of AIME24 [5], GSM8K (0-shot and 8-shot) [6], MATH [63], MATH-500, Minerva_Math [65], SVAMP [59]. Code-3 for HumanEval [52], HumanEval+ [53], LiveCodeBench [58]. Reasoning-9 for ARC-challenge [55], BBH (0-shot and few-shot), DROP [51], GPQA Diamond [60], MMLU [56], MMLU Pro [47], Winogrande [57], TheoremQA [54]. To avoid the impact of different prompts, we mainly evaluate under a zero-shot setting. Please refer to Appendix B.2 for more details.

4.2 Main Results

Table 1 shows the results of tuning various mainstream LLMs on BAAI-2k using full-parameter fine-tuning and LoRA. We set the rank as 128 in LoRA. Some findings can be summarized as follows:

- **Conventional tuning methods lead to marginal improvements and even performance degeneration.** Considering the average score, we can find that conventional tuning methods bring marginal improvements, such as 74.8 vs. 74.5 on Qwen-2.5-32B and 47.4 vs. 47.5 on Llama-3.2-3B. Moreover, they would lead to performance degeneration, such as 68.0 vs. 65.7 on Qwen-3-4B and 70.6 vs. 69.2 on Qwen-3-8B. The observations are consistent across full-parameter tuning and LoRA.
- **While conventional tuning fails, Shadow-FT performs well in adaptation at the same cost.** Across all model sizes and tasks, Shadow-FT consistently outperforms tuning baselines

⁴<https://huggingface.co/datasets/BAAI/Infinity-Instruct/tree/main/Gen>

Table 1: Performance comparison of different methods tuning popular LLMs. **Math-7** denotes the average score of 7 mathematical benchmarks including AIME24, **Code-3** for 3 code benchmarks including LiveCodeBench, and **Reasoning-9** for 9 commonsense reasoning benchmarks including MMLU Pro. For **Math-7** and **Code-3**, we report the mean value of three runs. We employ the Instruct version and report the final average scores. Please refer to Appendix B.3 for detailed scores.

Model	Method	Math-7		Code-3		Reasoning-9		Avg.
		Full	LoRA	Full	LoRA	Full	LoRA	
Qwen-3-4B	Vanilla	73.8		66.4		63.7		68.0
	FT	72.9	71.2	66.4	59.6	62.9	61.1	65.7
	Shadow-FT	73.7	75.9	67.4	69.7	64.9	65.0	69.4
Qwen-3-8B	Vanilla	74.5		72.7		64.7		70.6
	FT	74.0	71.3	71.2	69.6	64.6	64.3	69.2
	Shadow-FT	75.9	74.8	73.1	71.9	65.6	67.8	71.5
Qwen-3-14B	Vanilla	75.8		76.8		71.2		74.6
	FT	75.2	73.3	76.2	74.4	70.6	70.4	73.4
	Shadow-FT	78.9	78.6	77.0	77.8	71.4	71.5	75.9
Qwen-2.5-32B	Vanilla	74.1		75.9		73.4		74.5
	FT	75.7	74.3	75.8	75.9	73.6	73.8	74.8
	Shadow-FT	74.9	75.7	76.1	76.2	73.5	73.8	75.0
Llama-3.2-1B	Vanilla	23.8		26.5		34.2		28.2
	FT	24.5	25.3	26.1	26.6	32.8	33.3	28.1
	Shadow-FT	25.2	27.2	28.2	27.9	32.7	32.3	29.0
Llama-3.2-3B	Vanilla	53.6		39.3		49.3		47.4
	FT	52.7	51.9	40.2	41.4	49.4	49.1	47.5
	Shadow-FT	54.9	56.2	40.3	42.8	49.5	48.9	48.8
Llama-3.1-8B	Vanilla	56.8		50.9		56.6		54.8
	FT	56.8	57.8	53.4	51.8	58.5	57.5	56.0
	Shadow-FT	58.7	59.4	51.8	50.9	57.6	58.7	56.2

and vanilla INSTRUCT model. For example, on Qwen-3-4B, Shadow-FT archives an average score of 69.4, which is 3.7 higher than the 65.7 of conventional tuning methods and 1.4 higher than the vanilla INSTRUCT model. The conclusion is consistent on larger models such as Qwen-3-14B. Moreover, Shadow-FT does not introduce any extra training overheads. These consistent gains demonstrate that our proposed Shadow-FT can effectively learn the knowledge contained in training data.

- **Shadow-FT works well under both full-parameter setting and LoRA.** For instance, when tuning Qwen-3-4B under full-parameter setting, Shadow-FT achieves 73.7/67.4/64.9 on Math-7/Code-3/Reasoning-9 compared to 72.9/66.4/62.9 of conventional tuning methods. When applying a low-rank setting, Shadow-FT achieves 75.9/69.7/65.0, which is 4.7/10.1/3.9 higher than conventional LoRA. These indicate that Shadow-FT is effective with different tuning strategies, showing its robustness.
- **LoRA can outperform full-parameter.** When tuning using our BAAI-2k dataset, we find that Shadow-FT (LoRA) can outperform Shadow-FT (full), such as 69.7 vs. 67.4 on Code-3 when tuning Qwen-3-4B. Interestingly, we find that Shadow-FT (LoRA) typically performs better than Shadow-FT (full) on Math-7. However, considering the conventional tuning

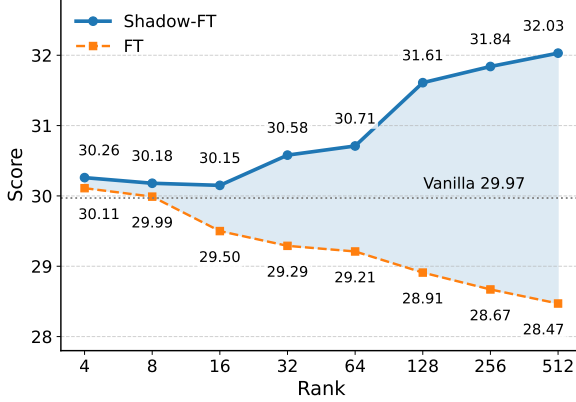


Figure 3: The average of Math-7, Code-3, and Reasoning-9 for different ranks when tuning Llama-3.2-1B using LoRA. We report the best performance searching learning rates in $\{5e-5, 1e-4, 2e-4, 5e-4\}$.

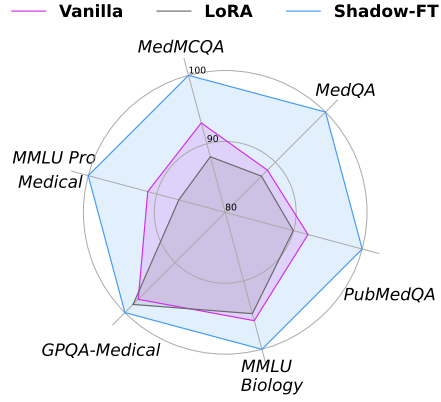


Figure 4: Performance of various methods when tuning Llama-3.2-1B on the Medical-o1-reasoning-SFT dataset. Detailed scores at Table 6.

methods, FT (full) would perform better [40], such as 75.9 vs. 74.8 on Qwen-3-8B. We leave it to future work for further investigation.

5 Extensive Analysis

5.1 Ranks in LoRA

We fine-tune the Llama-3.2-1B using LoRA with different ranks (from 4 to 512), and report the average scores after searching learning rates in $\{5e-5, 1e-4, 2e-4, 5e-4\}$. As shown in Figure 3, our proposed Shadow-FT (LoRA) can always outperform conventional LoRA with different ranks, demonstrating the robustness. With a larger rank, the conventional LoRA would perform worse, indicating more severe degeneration when tuning the INSTRUCT model [43]. In contrast to that, Shadow-FT (LoRA) can consistently benefit from more parameters (with larger ranks) and achieves better performance. For the results on Llama-3.1-8B, please refer to Appendix B.5.

5.2 Tuning on Domain Data

Tuning methods are typically employed to adapt LLMs for a specific domain, such as medical. Therefore, we perform tuning experiments on specific domain data, including Medical-o1-reasoning-SFT [41] in the medical domain, Code-Z1 [42] in the code domain, and LIMO [44] & OpenR1-Math [49] in the math domain. Following LIMO [44], we uniformly down sample the Medical-o1-reasoning-SFT to 1,000, and Code-Z1/OpenR1-Math to 2,000. On these domain tasks, we employ the LoRA with rank 128 and optimize with a learning rate of $2e-4$.

Figure 4 reports the results of tuning Llama-3.2-1B on Medical-o1-reasoning-SFT. We report the results on MMLU Pro-Medical [47], MedMCQA [46], PubMedQA [45], MMLU-Biology [61], and GPQA-Medical following [41], while normalizing the maximum score to 1 for better visualization. We can find that conventional LoRA would lead to performance degeneration, while Shadow-FT (LoRA) improves the performance, which is consistent with the conclusion on BAAI-2k. Please refer to Appendix B.4 for detailed scores.

Table 2 shows the detailed results of Math-7 and Code-3 tuning Qwen-3-8B and Llama-3.1-8B on Code-Z1, LIMO, and OpenR1-Math. The observations are consistent, i.e., conventional LoRA would lead to degeneration, while the proposed shadow-FT (LoRA) can effectively adapt LLMs on specific domain knowledge. For instance, Shadow-FT (LoRA) achieves a Math-7 score of 77.4 on Qwen-3-8B, which is 6.2 higher than 71.2 of LoRA, and 2.9 higher than the vanilla INSTRUCT model. Moreover, we also find that tuning LLM via Shadow-FT on code data can improve the math capability [42], and vice versa. In particular, when tuned via shadow-FT on Code-z1, the Qwen-3-8B

Table 2: The detailed mathematical and code performance tuning Qwen-3-8B and Llama-3.1-8B on Code-Z1, LIMO, and OpenR1-Math. *Va.* denotes the vanilla INSTRUCT baseline, *LoRA* for conventional LoRA, and *Shadow* for proposed Shadow-FT (LoRA).

Benchmark	Qwen-3-8B							Llama-3.1-8B						
	Code-Z1			LIMO		OpenR1-Math		Code-Z1			LIMO		OpenR1-Math	
	<i>Va.</i>	<i>LoRA</i>	<i>Shadow</i>	<i>LoRA</i>	<i>Shadow</i>	<i>LoRA</i>	<i>Shadow</i>	<i>Va.</i>	<i>LoRA</i>	<i>Shadow</i>	<i>LoRA</i>	<i>Shadow</i>	<i>LoRA</i>	<i>Shadow</i>
AIME24	20.0	13.3	36.7	23.3	26.7	16.7	26.7	6.7	3.3	20.0	6.7	3.3	3.3	6.7
GSM8K(8shot)	87.4	84.1	88.3	85.2	88.7	83.1	86.8	84.2	84.1	85.8	80.5	83.8	82.3	84.8
GSM8K(0shot)	93.0	91.9	93.6	91.7	92.4	92.7	92.9	84.2	85.4	85.7	82.5	86.1	86.1	85.9
MATH	70.9	69.4	69.1	70.0	67.6	70.6	66.5	48.0	48.8	51.3	44.3	45.8	39.8	47.7
MATH-500	83.2	79.8	88.0	77.0	80.4	80.2	85.0	48.4	50.8	55.4	44.4	43.8	41.8	48.8
Minerva_Math	73.0	69.7	72.9	69.9	73.1	70.8	73.2	40.6	39.6	45.5	37.1	41.2	44.0	44.2
SVAMP	91.4	90.3	93.3	90.9	92.9	90.3	93.0	83.1	86.5	86.9	83.7	85.9	85.1	87.1
Math-7	74.5	71.2 ↓	77.4 ↑	72.6 ↓	75.1 ↑	72.1 ↓	75.7 ↑	56.8	57.1 ↑	61.5 ↑	54.2 ↓	55.7 ↓	54.6 ↓	58.0 ↑
HumanEval	84.2	82.3	87.8	84.2	86.0	78.1	83.5	71.3	64.6	70.1	68.9	70.7	72.6	72.0
HumanEval+	79.9	76.8	78.1	79.3	81.1	75.6	81.1	63.4	48.2	64.6	62.2	64.0	61.6	62.8
LiveCodeBench	51.5	43.2	54.7	48.7	53.1	47.6	54.6	19.8	11.8	20.5	18.6	20.7	15.6	19.9
Code-3	72.7	67.4 ↓	73.5 ↑	70.7 ↓	73.4 ↑	67.1 ↓	73.1 ↑	50.9	41.5 ↓	51.7 ↑	49.9 ↓	51.8 ↑	49.9 ↓	51.6 ↑

Table 3: Benchmark results when tuning Llama-3.1-8B using DPO and Shadow-DPO. We report Math-7, Code-3, and Reasoning-9 scores with different ranks (8 and 128).

Method	Rank	Math-7	Code-3	Reasoning-9	Avg.
Vanilla	–	56.81	50.88	56.61	54.77
DPO	8	57.00	50.26	57.14	54.80
Shadow-DPO	8	57.00	50.61	57.27	54.96
DPO	128	56.24	50.28	57.33	54.62
Shadow-DPO	128	57.48	50.94	57.76	55.39

can achieve a score of 36.7 on the tough AIME-24 benchmark, showing superior adaptation and generalization ability.

5.3 Combined with DPO

Direct Preference Optimization (DPO), which directly optimizes a language model to adhere to human preferences without explicit reward modeling or reinforcement learning, shows promising performance when applying RL to LLMs [25]. Therefore, we try to combine Shadow-FT with DPO, i.e., applying DPO on BASE and then grafting the weight to INSTRUCT, termed as Shadow-DPO. Specifically, we achieve Shadow-DPO using LoRA on 1,000 paired samples from the Math-Step [62] dataset and set the rank to 8 and 128. As shown in Table 3, shadow-DPO outperforms DPO under two settings, such as 55.39 vs. 54.62 of vanilla DPO. It shows that the strategy employing the BASE as proxy of INSTRUCT also works for DPO. Meanwhile, a larger rank leads to better results for shadow-DPO, which is consistent with results tuning on BAAI-2k shown in Figure 3.

5.4 Performance on MLLM

For generality, we further conduct experiments tuning Multimodal Large Language Models (MLLMs). For the dataset, we select 10,000 samples from ChartMoE [50], which takes a chart and a natural language question as input to predict the answer. For MLLM, we select Gemma-3 [3] 12B/27B and Llama-3.2-Vision [67] 11B/90B. During training, we employ LoRA and set rank as 128. The learning rate is $2e-4$. We evaluate the tuned model via lmms-eval framework [68]. As shown in Table 4, both conventional LoRA and Shadow-FT (LoRA) effectively adapt MLLMs on ChartQA [66] tasks. Meanwhile, our proposed Shadow-FT outperforms LoRA especially on larger models, such

Table 4: Performance of Gemma-3 and Llama-3.2-Vision on Multi-modal ChartQA tasks.

Model	Size	Method	Aug. Split	Human Split	Overall
Gemma-3	12B	Vanilla	45.20	29.52	37.36
		LoRA	61.84	45.12	53.48
		Shadow-FT	63.36	46.48	54.92
	27B	Vanilla	56.56	27.28	41.92
		LoRA	70.88	49.68	60.28
		Shadow-FT	72.24	55.36	63.80
Llama-3.2-Vision	11B	Vanilla	27.04	17.20	22.12
		LoRA	93.04	55.84	74.44
		Shadow-FT	92.48	55.76	74.12
	90B	Vanilla	42.56	19.28	30.92
		LoRA	93.68	66.16	79.92
		Shadow-FT	93.44	67.76	80.60

as 63.80 on Gemma-3-27B compared to 60.28 of vanilla LoRA and 80.6 on Llama-3.2-Vision-90B compared to 79.92.

5.5 Model Zoo: More LLMs

We further apply Shadow-FT to more LLMs, including Gemma-3 series [3], Yi series [69], and Falcon series [70]. The hyperparameters are the same as tuning Qwen 3 and Llama 3. Table 5 shows the results of Math-7, Code-3, and Reasoning-9. We can find that proposed Shadow-FT consistently outperforms conventional tuning methods. For instance, Shadow-FT gets an average of 52.55 when tuning Gemma-3-4B, which is 1.1 higher than vanilla INSTRUCT model and 7.91 higher than conventional tuning methods. All the tuned models will be made public in the future.

6 Related Work

6.1 Tuning For LLMs

Large language models (LLMs) gain superior ability from pre-training on tremendous data [19], followed by tuning on various downstream tasks [20, 18]. These methods can be categorized into: 1) full-parameters method, which updates all the parameters, and 2) parameter-efficient fine-tuning (PEFT) method, lowering the tuning costs via parameter selection [21] or low-rank branches [14, 22]. More recently, Reinforcement Learning from Human Feedback (RLHF) methods show promising performance in aligning models to human preferences and improving the reasoning ability [25, 2, 23, 24]. These methods focus on improving the training strategy and involve the target model only. In this paper, we propose Shadow-FT to tune INSTRUCT model on BASE model. Also, our proposed Shadow-FT can be combined with these baselines to enhance the performance.

6.2 Model Guidance in Tuning

Introducing extra knowledge from other models has been proven as a promising way to enhance tuning performance, such as knowledge distillation [26, 27] and proxy-tuning [8]. Knowledge distillation methods aim to transfer the knowledge from a larger teacher model to a compact student model, via aligning the outputs [27, 28] or employing the teacher’s outputs as training data [29, 30]. Proxy-tuning first tunes a smaller LLM and then applies the logit differences to a larger model [8]. These methods transfer knowledge at the feature level or data level, while our proposed Shadow-FT directly grafts the weight updates. We also notice a very recent concurrent work [32] to transfer the fine-tuning ability. Differently, our proposed Shadow-FT focuses on tuning INSTRUCT via BASE model based on the *observation* that the weights are highly similar. Moreover, we conduct experiments on more LLMs across more benchmarks, and further extend the idea to MLLMs and DPO.

Table 5: Performance comparison of different methods tuning more LLMs. We employ the Instruct version and report the final average scores.

Model	Method	Math-7		Code-3		Reasoning-9		Avg.
		Full	LoRA	Full	LoRA	Full	LoRA	
Falcon Family								
Falcon3-3B	Vanilla	53.33		38.09		47.28		46.23
	FT	55.83	58.70	39.57	41.23	48.43	49.50	48.88
	Shadow-FT	56.74	60.31	41.02	43.69	48.16	48.25	49.70
Falcon3-10B	Vanilla	57.23		60.03		53.85		57.04
	FT	59.33	68.74	60.95	61.54	54.17	55.72	60.08
	Shadow-FT	58.27	70.40	61.35	62.20	53.19	52.83	59.71
Gemma Family								
Gemma-3-4B	Vanilla	54.02		48.33		52.01		51.45
	FT	35.34	49.12	48.15	43.03	43.83	48.37	44.64
	Shadow-FT	56.68	56.30	48.87	48.93	52.88	51.62	52.55
Gemma-3-12B	Vanilla	60.82		58.06		61.54		60.14
	FT	56.56	62.84	58.17	59.21	61.63	61.99	60.07
	Shadow-FT	61.05	64.59	58.17	60.86	61.59	62.66	61.49
Yi Family								
Yi-6B	Vanilla	17.34		8.40		38.63		21.46
	FT	18.93	18.39	10.64	11.89	40.84	40.46	23.53
	Shadow-FT	17.73	17.21	13.35	14.30	38.70	38.25	23.26
Yi-Coder-9B	Vanilla	28.01		61.85		40.73		43.53
	FT	26.05	26.11	52.70	53.95	39.55	37.22	39.26
	Shadow-FT	28.41	29.09	62.07	64.72	40.27	39.88	44.07

7 Conclusion and Limitations

In this work, we propose Shadow-FT, a novel framework to fine-tune INSTRUCT models by leveraging their corresponding BASE models. Inspired by the observation that the weights of BASE and INSTRUCT are highly similar, we propose Shadow-FT to tune INSTRUCT vis BASE, aiming to tune INSTRUCT better. Extensive experiments across multiple LLM series, including Qwen, Llama, Gemma, and Falcon, demonstrate that Shadow-FT consistently outperforms conventional full-parameter and parameter-efficient fine-tuning methods. Notably, Shadow-FT introduces no additional training cost or parameters, yet it achieves superior performance across diverse benchmarks covering math, coding, and reasoning tasks. We further show that Shadow-FT generalizes well to multimodal large language models (MLLMs) and can be seamlessly combined with alignment techniques such as DPO, offering a simple yet effective solution for improving instruction-following models.

In Shadow-FT, we first tune the BASE model and then graft the weight updates to the INSTRUCT model. However, there are some LLMs for which the paired BASE models are not available, such as Qwen-3-32B. For these LLMs, we can not apply Shadow-FT. Therefore, finding a proper 'shadow' for these models is an interesting topic for future work.

References

- [1] AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [3] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- [4] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.
- [5] MAA. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME 2024*, February 2024. URL <https://maa.org/math-competitions/american-invitational-mathematics-examination-aime>.
- [6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [7] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*, 2024.
- [8] Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A Smith. Tuning language models by proxy. *arXiv preprint arXiv:2401.08565*, 2024.
- [9] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiusi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- [10] Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *arXiv preprint arXiv:2407.13623*, 2024.
- [11] Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, et al. A survey on the honesty of large language models. *arXiv preprint arXiv:2409.18786*, 2024.
- [12] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.
- [13] Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, et al. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association*, 31(9): 1865–1874, 2024.
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [15] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.
- [16] Yubo Wang, Xiang Yue, and Wenhui Chen. Critique fine-tuning: Learning to critique is more effective than learning to imitate. *arXiv preprint arXiv:2501.17703*, 2025.

- [17] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.
- [18] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- [19] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.
- [20] 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.
- [21] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.
- [22] Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. *arXiv preprint arXiv:2406.11909*, 2024.
- [23] 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.
- [24] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- [25] 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*, 36:53728–53741, 2023.
- [26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [27] Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*, 2024.
- [28] Runming Yang, Taiqiang Wu, and Yujiu Yang. Loca: Logit calibration for knowledge distillation. *arXiv preprint arXiv:2409.04778*, 2024.
- [29] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*, 2024.
- [30] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.
- [31] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [32] Pin-Jie Lin, Rishab Balasubramanian, Fengyuan Liu, Nikhil Kandpal, and Tu Vu. Efficient model development through fine-tuning transfer. *arXiv preprint arXiv:2503.20110*, 2025.
- [33] Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung-yi Lee. Chat vector: A simple approach to equip llms with instruction following and model alignment in new languages. *arXiv preprint arXiv:2310.04799*, 2023.

- [34] 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.
- [35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [36] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL <http://arxiv.org/abs/2403.13372>.
- [37] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023.
- [38] LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. <https://github.com/InternLM/lmdeploy>, 2023.
- [39] Qwen Team. Qwen3, April 2025. URL <https://qwenlm.github.io/blog/qwen3/>.
- [40] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.
- [41] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.
- [42] Zhaojian Yu, Yinghao Wu, Yilun Zhao, Arman Cohan, and Xiao-Ping Zhang. Z1: Efficient test-time scaling with code. *arXiv preprint arXiv:2504.00810*, 2025.
- [43] Runming Yang, Taiqiang Wu, Jiahao Wang, Pengfei Hu, Yik-Chung Wu, Ngai Wong, and Yujiu Yang. Llm-neo: Parameter efficient knowledge distillation for large language models. *arXiv preprint arXiv:2411.06839*, 2024.
- [44] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.
- [45] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- [46] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.
- [47] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhramil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [48] 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.
- [49] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL <https://github.com/huggingface/open-r1>.
- [50] Zhengzhuo Xu, Bowen Qu, Yiyan Qi, Sinan Du, Chengjin Xu, Chun Yuan, and Jian Guo. Chartmoe: Mixture of diversely aligned expert connector for chart understanding. *ArXiv*, abs/2409.03277, 2024.

- [51] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*, 2019.
- [52] 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.
- [53] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=1qv610Cu7>.
- [54] Wenhui Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023.
- [55] 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:1803.05457v1*, 2018.
- [56] 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)*, 2021.
- [57] Winogrande: An adversarial winograd schema challenge at scale. 2019.
- [58] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *CoRR*, abs/2403.07974, 2024. URL <https://doi.org/10.48550/arXiv.2403.07974>.
- [59] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL <https://aclanthology.org/2021.naacl-main.168>.
- [60] 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. In *First Conference on Language Modeling*, 2024.
- [61] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [62] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.
- [63] 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*, 2021.

- [64] 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.
- [65] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022.
- [66] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- [67] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [68] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. URL <https://arxiv.org/abs/2407.12772>.
- [69] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- [70] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

A Similarity on More LLMs

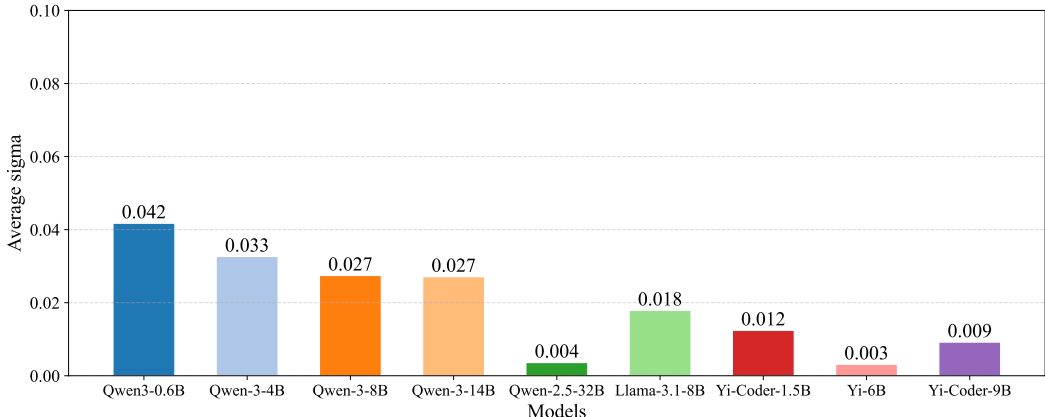


Figure 5: Average σ values of more LLMs.

As shown in Figure 5, we can find that all $\sigma < 0.05$, indicating high similarity between BASE and INSTRUCT. Also, the larger the LLMs, the smaller the gaps.

B Experimental Details

B.1 Hyper-parameters

For the experiments, we set the hyperparameters after grid search. The batch size is 2, with the gradient_accumulation_steps as 16. During experiments, the cutoff the inputs to 4096 and train for 1 epoch.

B.2 Benchmarks

The details about the benchmarks are detailed in Table 7. Since the n-shot setting are unstable, we prefer to report the 0-shot results. For the popular GSM8K and BBH, we also report 8-shot and 3-shot results.

B.3 Detailed Data of Table 1

The detailed scores are listed in Table 8, Table 9, and Table 10.

B.4 Detailed table for Medical Benchmarks

Table 6 reports the detailed results of tuning Llama-3.2-1B on the Medical-o1-reasoning-SFT dataset.

Table 6: Performance of LLAMA-3.2-1B-INSTRUCT on medical QA benchmarks.

Benchmark	Vanilla	FT	Shadow-FT
GPQA-Medical	23.85	24.10	24.50
MMLU Pro-Medical	25.20	23.95	27.60
MedMCQA	30.15	28.55	32.40
MedQA	25.95	25.60	29.35
PubMedQA	55.85	54.55	60.65
MMLU-Biology	49.68	49.15	51.85

Table 7: Details on instruction-model evaluations. CoT denotes the chain-of-thought setting.

Evaluation	Metric	Type	n-shot	CoT
<i>Math-7</i>				
AIME24	pass@1	sampling	0-shot	
GSM8K(0-shot)	Accuracy	sampling	0-shot	✓
GSM8K(8-shot)	Accuracy	sampling	8-shot	✓
MATH	Accuracy	sampling	0-shot	✓
MATH-500	Accuracy	sampling	0-shot	
Minerva Math	Accuracy	sampling	4-shot	
SVAMP	Accuracy	sampling	0-shot	
<i>Code-3</i>				
HumanEval	pass@1	sampling	0-shot	
HumanEval+	pass@1	sampling	0-shot	
LiveCodeBench		average		
- generation	pass@1	sampling	0-shot	✓
- test	pass@1	sampling	0-shot	✓
- prediction	pass@1	sampling	0-shot	✓
<i>Reasoning-9</i>				
ARC-Challenge	Accuracy	sampling	0-shot	✓
BBH(0-shot)	Accuracy	sampling	0-shot	
BBH(3-shot)	Accuracy	sampling	3-shot	
Drop	Accuracy	sampling	0-shot	
GPQA Diamond	Accuracy	sampling	0-shot	✓
MMLU	Accuracy	sampling	0-shot	
MMLU Pro	Accuracy	sampling	0-shot	
Winogrande	Accuracy	sampling	0-shot	
TheoremQA	Accuracy	sampling	0-shot	

B.5 Ranks in LoRA on Llama-3.1-8B

As shown in Figure 6, the conclusions regarding Llama-3.1-8B are consistent with Section 5.1.

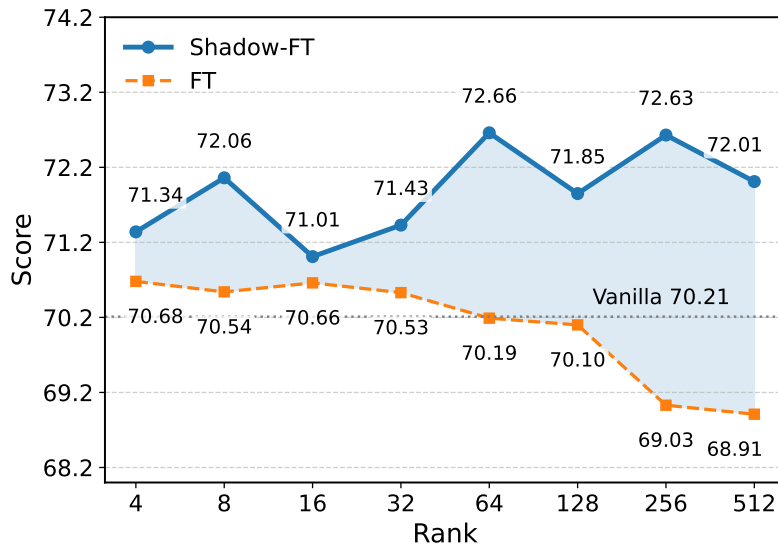


Figure 6: Performance tuning Llama-3.1-8B with different ranks.

Table 8: Detailed results on the math benchmarks for Table 1. Three times independent training and three times evaluation average are reported.

Model	Method	AIME24	GSM8K (8-shot)	GSM8K (0-shot)	MATH	MATH 500	Minerva Math	SVAMP	Math-7
Llama-3.2-1B	Vanilla	1.1	46.9	1.8	15.8	15.1	20.3	65.3	23.8
	FT (<i>full</i>)	1.1	46.8	0.8	18.6	19.1	19.0	66.5	24.5
	Shadow-FT (<i>full</i>)	0.0	47.2	1.0	18.9	18.3	23.1	67.9	25.2
	FT (<i>LoRA</i>)	1.1	45.2	2.6	21.8	20.3	18.7	67.5	25.3
	Shadow-FT (<i>LoRA</i>)	0.0	47.8	4.6	22.1	24.5	25.1	66.4	27.2
Llama-3.2-3B	Vanilla	4.4	76.6	79.5	45.8	47.9	36.1	84.7	53.6
	FT (<i>full</i>)	8.9	77.0	77.0	45.2	44.2	32.4	84.1	52.7
	Shadow-FT (<i>full</i>)	8.9	77.8	80.4	47.2	47.9	37.6	84.8	54.9
	FT (<i>LoRA</i>)	4.4	76.5	73.5	46.4	47.7	31.7	83.1	51.9
	Shadow-FT (<i>LoRA</i>)	11.1	78.1	77.6	49.8	52.0	39.3	85.4	56.2
Llama-3.1-8B	Vanilla	6.7	83.6	85.0	49.1	49.2	40.8	83.2	56.8
	FT (<i>full</i>)	1.1	84.0	85.4	51.0	51.5	39.0	85.8	56.8
	Shadow-FT (<i>full</i>)	6.7	85.0	84.0	52.2	53.2	43.8	86.3	58.7
	FT (<i>LoRA</i>)	6.7	83.8	83.8	50.2	52.5	41.3	86.5	57.8
	Shadow-FT (<i>LoRA</i>)	6.7	85.0	84.5	52.0	53.0	48.3	86.0	59.4
Qwen-3-4B	Vanilla	18.9	87.8	92.2	70.3	82.3	73.4	91.5	73.8
	FT (<i>full</i>)	14.4	88.1	91.6	70.1	82.4	72.1	91.2	72.9
	Shadow-FT (<i>full</i>)	16.7	87.4	92.3	70.0	84.3	73.3	91.7	73.7
	FT (<i>LoRA</i>)	18.9	84.2	91.6	68.1	77.3	67.4	90.7	71.2
	Shadow-FT (<i>LoRA</i>)	28.9	88.3	92.5	70.4	84.5	73.8	92.8	75.9
Qwen-3-8B	Vanilla	22.2	87.3	93.4	70.8	83.1	73.2	91.6	74.5
	FT (<i>full</i>)	22.2	86.2	93.1	70.6	80.7	72.7	92.1	74.0
	Shadow-FT (<i>full</i>)	32.2	87.5	93.3	70.6	82.9	73.2	91.4	75.9
	FT (<i>LoRA</i>)	17.8	83.6	92.1	68.9	77.3	68.4	90.7	71.3
	Shadow-FT (<i>LoRA</i>)	22.2	88.5	92.9	70.5	84.1	73.6	91.8	74.8
Qwen-3-14B	Vanilla	20.0	90.0	95.3	72.1	85.2	75.7	92.6	75.8
	FT (<i>full</i>)	17.8	88.9	94.9	72.2	85.5	75.5	91.3	75.1
	Shadow-FT (<i>full</i>)	40.0	90.7	95.2	71.7	86.3	76.0	92.7	78.9
	FT (<i>LoRA</i>)	14.4	87.3	94.5	71.7	81.3	72.8	91.0	73.3
	Shadow-FT (<i>LoRA</i>)	36.7	90.7	95.9	71.3	86.7	76.1	93.2	78.7
Qwen-2.5-32B	Vanilla	16.7	84.3	95.5	78.0	83.1	71.7	89.3	74.1
	FT (<i>full</i>)	21.1	86.6	95.4	74.8	82.9	76.8	92.1	75.7
	Shadow-FT (<i>full</i>)	13.3	85.0	95.5	76.8	84.1	78.0	91.3	74.9
	FT (<i>LoRA</i>)	14.4	85.7	95.3	73.6	83.8	75.0	92.1	74.3
	Shadow-FT (<i>LoRA</i>)	18.9	86.3	95.6	76.0	84.3	77.3	91.3	75.7

Table 9: Detailed data on the code benchmarks for Table 1. Three times independent training and three times evaluation averages are reported.

Model	Method	HumanEval	HumanEval ⁺	LiveCodeBench				Code-3
				Exec	Gen	Out	Avg	
Llama 3.2-1B	Vanilla	40.9	35.0	4.0	7.0	0.2	3.7	26.5
	FT (<i>full</i>)	38.2	34.2	9.3	6.9	1.2	5.8	26.1
	Shadow-FT (<i>full</i>)	42.9	36.8	5.9	7.3	1.1	4.8	28.2
	FT (<i>LoRA</i>)	39.2	34.4	11.0	6.5	0.8	6.1	26.6
	Shadow-FT (<i>LoRA</i>)	41.9	35.4	10.5	6.9	2.2	6.5	27.9
Llama 3.2-3B	Vanilla	60.0	52.2	0.0	16.8	0.7	5.8	39.3
	FT (<i>full</i>)	57.9	53.7	4.5	16.3	6.2	9.0	40.2
	Shadow-FT (<i>full</i>)	60.6	54.1	0.0	16.8	2.1	6.3	40.3
	FT (<i>LoRA</i>)	59.1	50.8	9.7	17.1	15.5	14.1	41.4
	Shadow-FT (<i>LoRA</i>)	61.2	55.3	14.4	16.0	5.8	12.1	42.9
Llama 3.1-8B	Vanilla	69.7	62.8	17.3	19.8	23.2	20.1	50.9
	FT (<i>full</i>)	70.7	67.3	16.9	22.3	27.5	22.2	53.4
	Shadow-FT (<i>full</i>)	70.1	63.6	16.6	20.8	27.8	21.7	51.8
	FT (<i>LoRA</i>)	70.7	63.4	16.3	21.0	26.8	21.4	51.8
	Shadow-FT (<i>LoRA</i>)	71.1	50.4	14.9	21.3	27.3	21.2	50.9
Qwen-3-4B	Vanilla	77.9	71.3	41.8	48.8	59.7	50.1	66.4
	FT (<i>full</i>)	80.9	70.9	43.1	46.1	53.0	47.4	66.4
	Shadow-FT (<i>full</i>)	80.3	71.1	42.5	49.7	60.1	50.8	67.4
	FT (<i>LoRA</i>)	76.4	69.1	13.1	41.1	45.6	33.3	59.6
	Shadow-FT (<i>LoRA</i>)	81.3	76.8	43.2	49.1	60.6	51.0	69.7
Qwen-3-8B	Vanilla	85.8	79.9	42.3	51.3	63.4	52.3	72.7
	FT (<i>full</i>)	82.7	79.3	42.9	51.8	59.7	51.5	71.2
	Shadow-FT (<i>full</i>)	86.8	79.3	41.9	52.3	65.2	53.1	73.1
	FT (<i>LoRA</i>)	84.2	78.5	42.0	45.7	50.9	46.2	69.6
	Shadow-FT (<i>LoRA</i>)	84.6	77.6	41.9	52.4	66.1	53.5	71.9
Qwen-3-14B	Vanilla	86.8	83.1	51.9	55.8	74.2	60.6	76.8
	FT (<i>full</i>)	87.6	83.5	50.9	54.3	67.3	57.5	76.2
	Shadow-FT (<i>full</i>)	87.4	82.9	52.1	55.6	74.4	60.7	77.0
	FT (<i>LoRA</i>)	85.6	82.3	51.2	51.3	62.5	55.0	74.4
	Shadow-FT (<i>LoRA</i>)	87.8	84.4	50.7	56.8	76.4	61.3	77.8
Qwen-2.5-32B	Vanilla	86.4	82.1	58.3	54.6	64.6	59.1	75.9
	FT (<i>full</i>)	85.6	81.1	60.3	55.8	66.4	60.8	75.8
	Shadow-FT (<i>full</i>)	86.6	81.5	60.5	55.7	64.0	60.1	76.1
	FT (<i>LoRA</i>)	85.4	81.7	60.9	55.0	64.8	60.7	75.9
	Shadow-FT (<i>LoRA</i>)	87.4	80.5	61.8	55.0	64.9	60.6	76.2

Table 10: Detailed results on the general Reasoning benchmarks for Table 1.

Method	MMLU	MMLU Pro	WinoG	DROP	ARC Challenge	BBH (0-shot)	BBH (3-shot)	GPQA Diamond	TheoremQA	Reasoning-9
<i>Llama-3.2-1B</i>										
Vanilla	46.8	21.4	51.9	42.7	56.6	24.4	26.1	27.8	9.9	34.2
FT (<i>full</i>)	46.9	21.8	50.4	39.0	56.6	20.4	24.8	24.8	10.8	32.8
Shadow-FT (<i>full</i>)	47.1	22.7	51.1	41.6	52.9	22.2	20.8	26.3	9.6	32.7
FT (<i>LoRA</i>)	46.7	22.1	51.2	40.8	56.6	20.9	26.3	23.2	11.8	33.3
Shadow-FT (<i>LoRA</i>)	46.6	23.2	51.4	43.9	52.2	17.2	20.4	25.3	10.6	32.3
<i>Llama-3.2-3B</i>										
Vanilla	62.4	39.7	53.9	71.8	78.6	41.8	49.2	29.3	17.4	49.3
FT (<i>full</i>)	62.0	39.2	54.5	71.7	79.0	41.8	51.8	25.8	18.5	49.4
Shadow-FT (<i>full</i>)	62.4	40.4	54.3	72.1	79.0	41.7	50.0	28.3	17.6	49.5
FT (<i>LoRA</i>)	61.9	39.9	51.1	71.7	82.7	41.4	49.4	25.3	18.4	49.1
Shadow-FT (<i>LoRA</i>)	62.1	41.6	54.6	72.0	79.0	38.6	49.6	26.8	16.1	48.9
<i>Llama-3.1-8B</i>										
Vanilla	69.5	48.5	59.4	81.4	85.4	44.6	67.6	25.8	27.3	56.6
FT (<i>full</i>)	69.7	49.2	60.9	80.0	87.1	46.8	71.1	30.8	30.6	58.5
Shadow-FT (<i>full</i>)	69.6	49.3	60.2	81.7	85.8	46.8	67.0	28.3	29.5	57.6
FT (<i>LoRA</i>)	69.3	48.9	60.0	79.5	86.4	48.8	68.0	30.3	26.8	57.5
Shadow-FT (<i>LoRA</i>)	69.4	50.8	60.2	80.1	85.4	51.6	68.8	32.8	29.1	58.7
<i>Qwen-3-4B</i>										
Vanilla	70.7	57.1	57.7	77.3	91.5	57.7	78.7	37.4	44.6	63.6
FT (<i>full</i>)	70.7	54.2	56.8	75.9	91.2	57.3	77.2	38.9	44.4	63.0
Shadow-FT (<i>full</i>)	71.4	57.0	57.4	77.7	92.2	58.4	78.4	45.0	46.5	64.9
FT (<i>LoRA</i>)	71.9	51.2	59.0	69.1	91.5	54.7	73.5	39.4	39.9	61.1
Shadow-FT (<i>LoRA</i>)	71.8	58.2	58.8	79.1	91.9	59.6	77.0	46.0	42.4	65.0
<i>Qwen-3-8B</i>										
Vanilla	76.5	55.8	55.7	85.2	91.9	59.8	80.0	46.5	31.0	64.7
FT (<i>full</i>)	76.3	53.0	54.8	84.8	91.2	60.1	80.1	44.4	36.5	64.6
Shadow-FT (<i>full</i>)	76.6	56.0	54.8	85.8	92.2	59.2	79.8	53.5	32.1	65.6
FT (<i>LoRA</i>)	76.1	57.2	55.9	80.6	92.5	59.0	75.5	41.4	40.9	64.3
Shadow-FT (<i>LoRA</i>)	78.6	61.5	55.0	85.8	92.5	59.3	79.6	56.6	41.1	67.8
<i>Qwen-3-14B</i>										
Vanilla	79.4	64.2	68.5	86.3	94.6	61.4	84.2	47.0	54.6	71.1
FT (<i>full</i>)	79.7	61.3	67.8	85.5	94.9	61.2	84.1	47.5	53.0	70.6
Shadow-FT (<i>full</i>)	79.6	64.9	68.7	86.9	94.6	60.3	83.9	46.5	57.6	71.4
FT (<i>LoRA</i>)	79.6	60.7	68.5	84.0	95.3	63.3	83.0	47.0	51.9	70.4
Shadow-FT (<i>LoRA</i>)	79.8	66.1	69.1	88.1	93.6	58.2	83.6	48.0	56.8	71.5
<i>Qwen-2.5-32B</i>										
Vanilla	83.4	68.8	82.2	88.1	95.3	63.6	84.6	39.9	54.3	73.4
FT (<i>full</i>)	83.4	68.3	81.9	88.7	94.6	63.0	83.8	42.4	56.5	73.6
Shadow-FT (<i>full</i>)	83.2	69.1	82.6	88.4	95.6	64.3	82.9	39.4	55.8	73.5
FT (<i>LoRA</i>)	83.6	68.9	82.2	88.8	94.9	62.8	83.7	44.4	54.8	73.8
Shadow-FT (<i>LoRA</i>)	83.0	68.8	82.7	88.6	94.6	64.3	83.4	43.9	54.8	73.8