# Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning

Zhaorui Yang[1], Tianyu Pang[2], Haozhe Feng[3], Han Wang[1]
Wei Chen[1], Minfeng Zhu[1], Qian Liu[2],

1 ZHEJIANG UNIVERSITY
2 sea AI Lab
3 Tencent 腾讯

ACL 2024 Bangkok, Thailand

## Background: The Challenge of Enhancing Existing Models

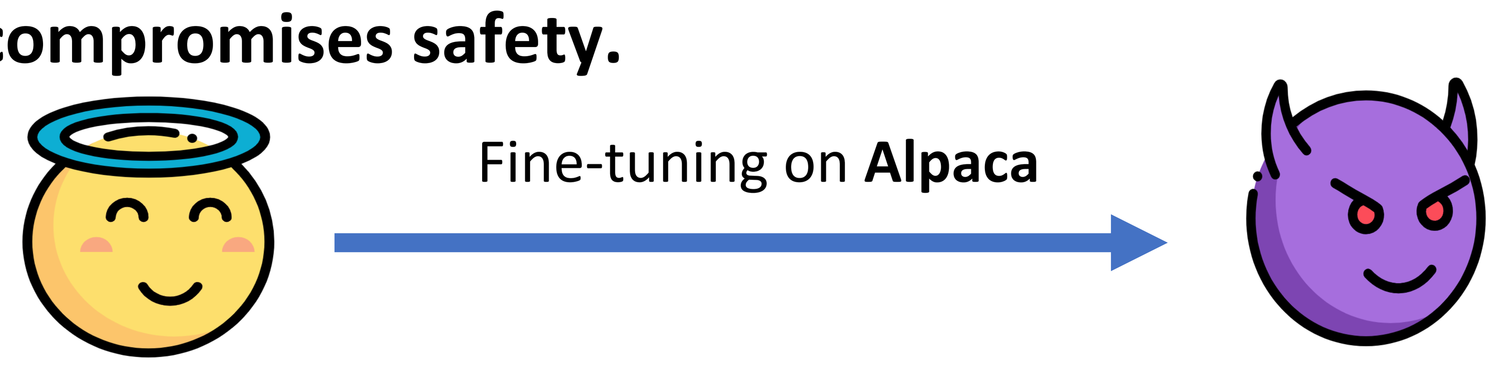
When Meta releases the powerful chat model **Llama-3-Instruct** without deets on its fine-tuning, and you need to enhance its capabilities further on some tasks, sounds easy, right ?

**Wrong! Reality is way tougher.**

**How come?**

**(1) Performance issue:** It has already leveraged *10M* human examples and never released. Enhancing performance with vanilla fine-tuning is **non-trivial.**

**(2) Catastrophic forgetting:** Vanilla Fine-tuning **compromises safety.**

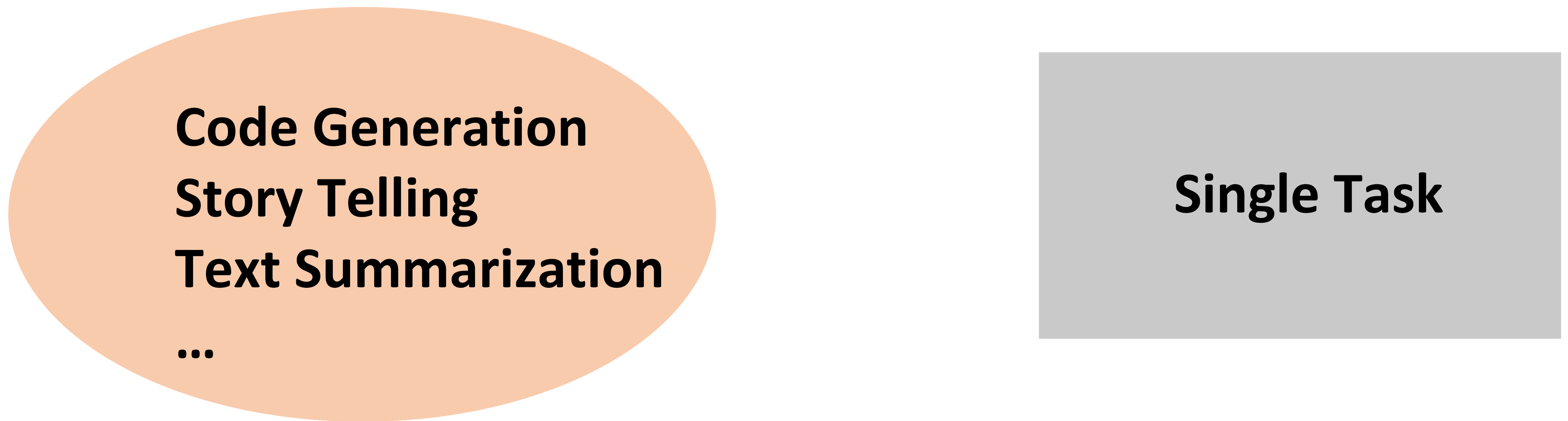



**Llama-3-Instruct** is already an aligned model

Fine-tuning aligned models compromises **safety**, even when you do not intend to

## The Root Cause of Challenge

The primary cause of the fine-tuning challenge lies in the **distribution gap** between the task data and the original LLM.

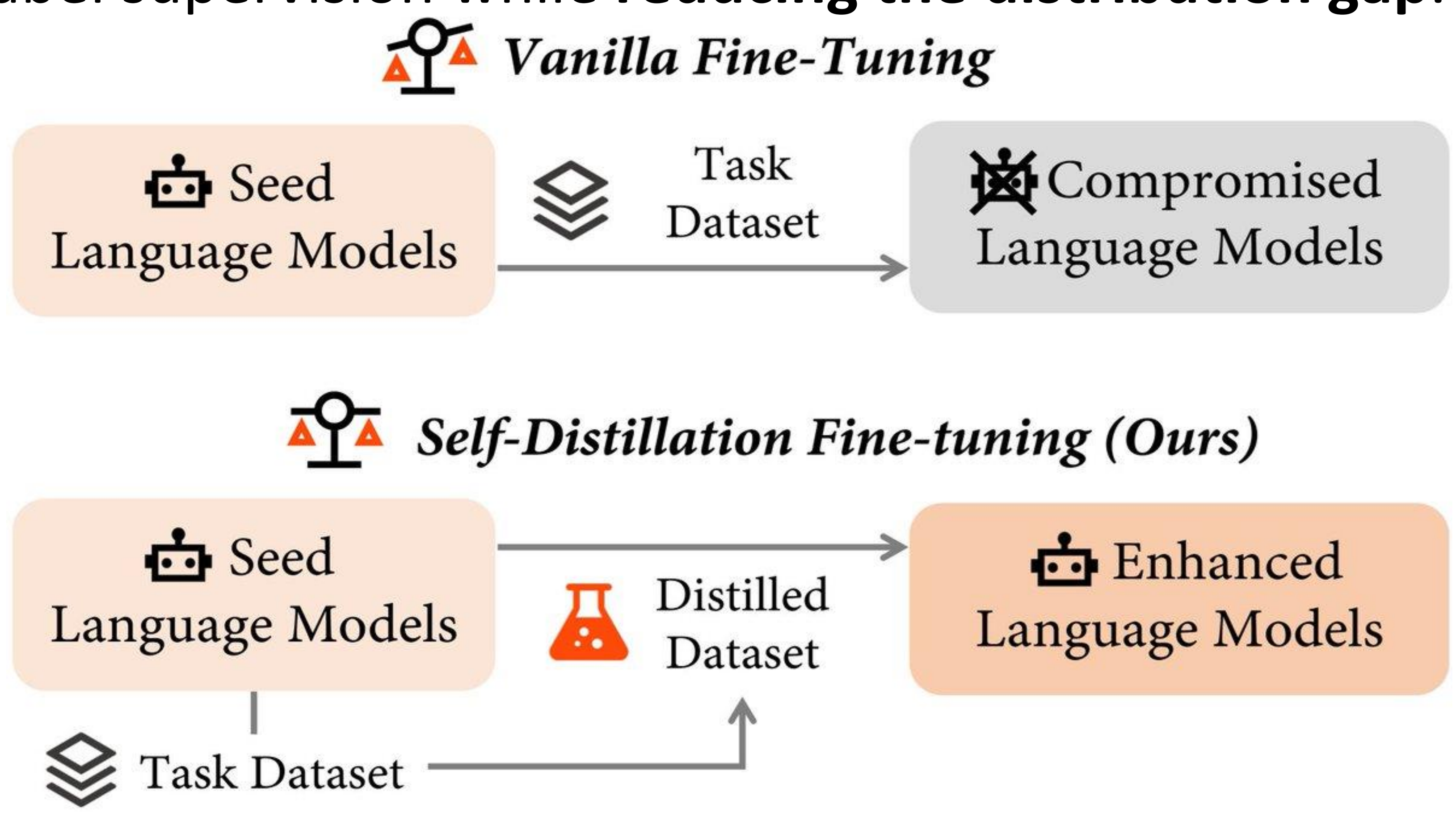



**Llama-3-Instruct** capabilities: diverse, aligned with human values

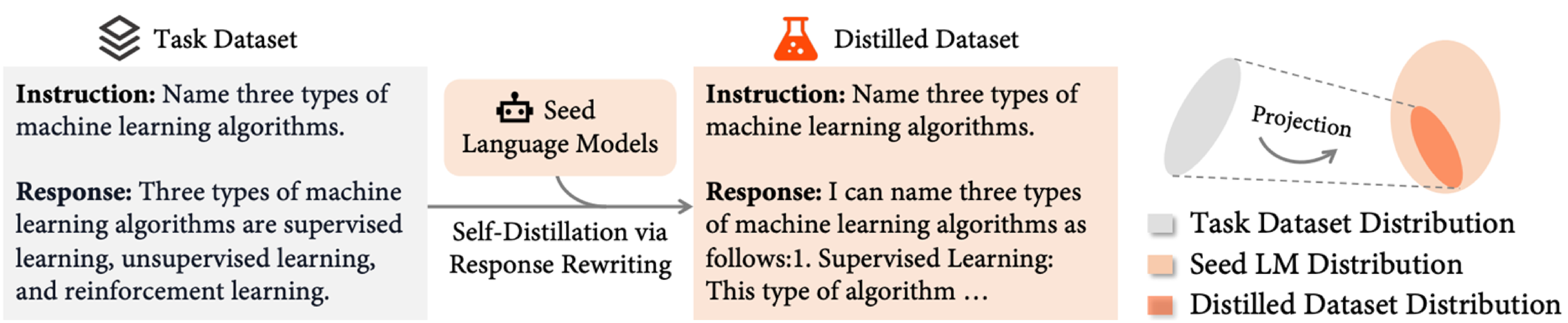
**Task** data: narrowed distribution and focused on certain tasks or domains

## Introducing Self-Distillation Fine-Tuning (SDFT)

SDFT aligns task data with the LLMs' distribution, preserving label supervision while **reducing the distribution gap**.




It achieves this by having the LLM rewrite target labels, integrating new tasks with the model's existing knowledge.




## Method: Self-Distillation Fine-tuning

1. Start with a **chat model** (i.e., seed language model)
2. Curate a **task dataset** targeting areas where the model underperforms
3. Use the model to rewrite responses in the dataset, creating a **distilled dataset**
4. **Fine-tune** on the distilled dataset, balancing new skills and original capabilities

## Experiments: SDFT vs. Vanilla Fine-tuning

While both vanilla fine-tuning and SDFT can improve **target task performance**, SDFT excels in preserving the model's **broad capabilities**.

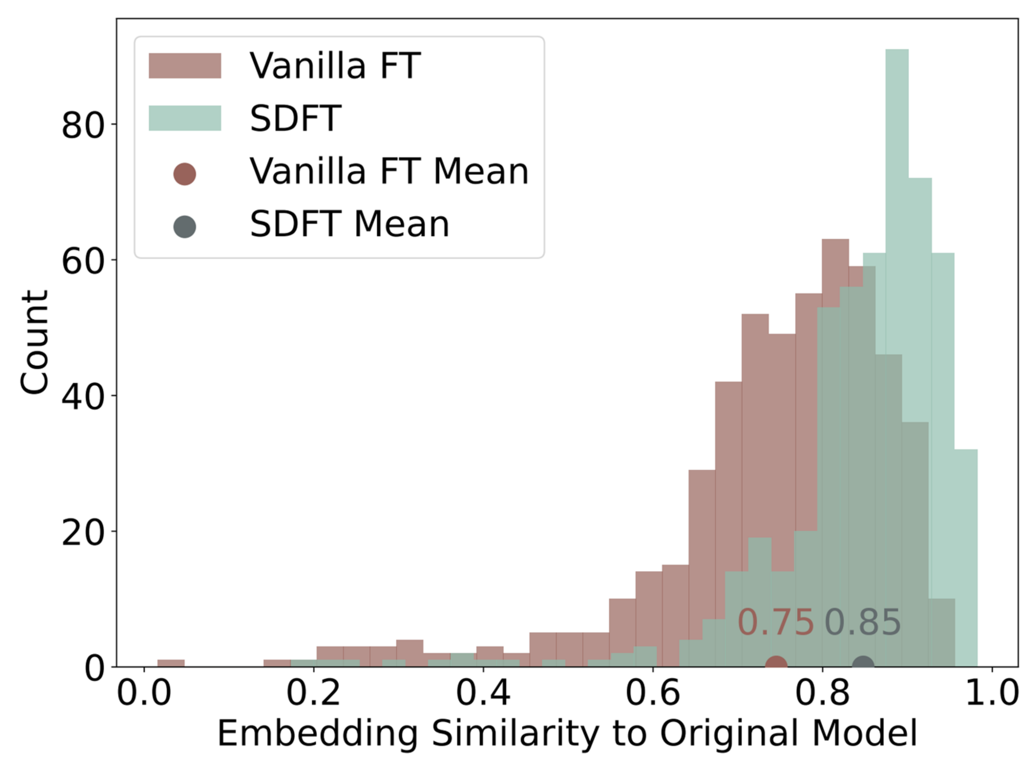
| Method | Dataset | OpenFunctions | GSM8K | HumanEval | Average |
|---|---|---|---|---|---|
| Seed LM | — | 19.6 | 29.4 | 13.4 | 20.8 |
| Vanilla FT | OpenFunctions | 34.8 | 21.5 | 9.8 | 22.0 |
| | GSM8K | 17.9 | 31.9 | 12.2 | 20.7 |
| | MagiCoder | 3.6 | 23.2 | 18.9 | 15.2 |
| SDFT (Ours) | OpenFunctions | 36.6 ↑ 1.8 | 29.1 ↑ 7.6 | 15.2 ↑ 5.4 | 27.0 ↑ 5.0 |
| | GSM8K | 17.9 ↑ 0.0 | 34.4 ↑ 2.5 | 14.6 ↑ 2.4 | 22.3 ↑ 1.6 |
| | MagiCoder | 8.0 ↑ 5.4 | 24.9 ↑ 1.7 | 18.3 ↓ 0.6 | 17.1 ↑ 1.9 |

Vanilla fine-tuning leads to notable degradation in safety and general helpfulness, while SDFT maintains strong alignment after fine-tuning.

| Dataset for FT | Raw Safe Rate | Jailbreak Safe Rate | AlpacaEval Win Rate |
|---|---|---|---|
| Seed LM | 99.81 | 88.85 | 66.04 |
| OpenFunctions | 98.27 → 99.23 (↑ 0.96) | 87.31 → 94.42 (↑ 7.11) | 35.49 → 67.66 (↑ 32.17) |
| GSM8K | 82.12 → 87.12 (↑ 5.00) | 54.81 → 65.58 (↑ 10.77) | 23.38 → 66.73 (↑ 43.35) |
| MagiCoder | 96.73 → 97.88 (↑ 1.15) | 83.65 → 88.65 (↑ 5.00) | 76.52 → 76.09 (↓ 0.43) |

## Analysis: Distribution Gap

We assess shifts in model representation by measuring **embedding similarity** between the original model and the fine-tuned one.




SDFT mitigates the distribution shift, thus alleviating forgetting.

## Take Away

**Finding:** distribution shift leads to catastrophic forgetting in vanilla fine-tuning

**Method:** self-distillation => bridge distribution gap => mitigate forgetting

**Experiments:** improve the target task performance and keep the original capabilities

**Paper**: Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning **Presenter**: Zhaorui Yang **Email**: zhaorui.yang@zju.edu.cn