TeCho Reading Group
Oct 24, 2025

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

The Delta Learning Hypothesis:
Preference Tuning on Weak Data can Yield Strong Gains

COLM 2025

Authors: Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, Pang Wei Koh

University of Washington, Allen Institute for AI, Carnegie Mellon University

Presented by Sriram Sai Ganesh

Sections

1. Introduction
2. Background work
3. The Delta Learning Hypothesis
4. Results: Controlled Experiments
5. Results: LM Post-Training
6. Analysis

Introduction

Intuition: *strong data builds strong models.*
Task: ***preference tuning*** – alignment using preference annotations on paired data.

Predominant approach:

(humans, 100B+ parameter LMs)

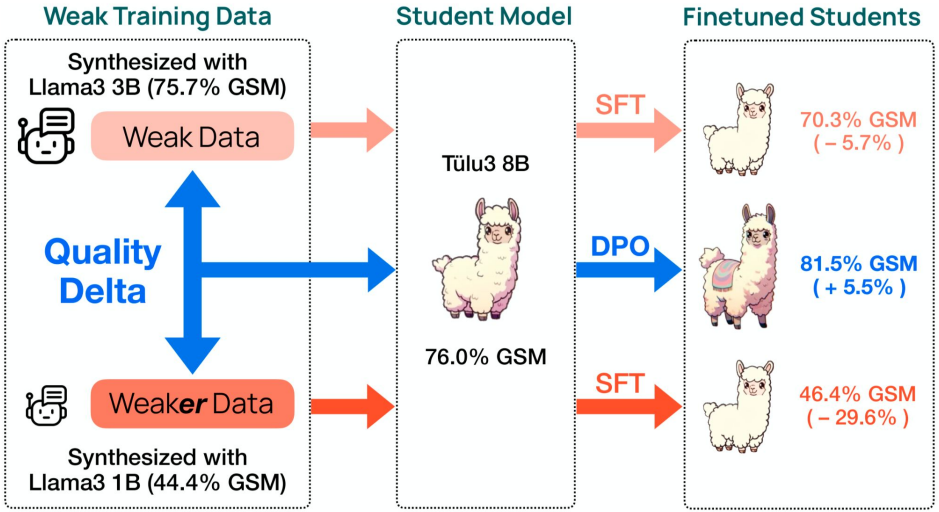
1. **Collect preference annotations from strong supervisors.**
2. **Optimize LMs by maximizing some objective based on 1.**

Limitation: **<u>High collection cost of strong data.</u>**

- May require expert annotations.
- May exceed current human expertise.

This paper: **Can we instead train on the Δ between two weaker models?**

Introduction

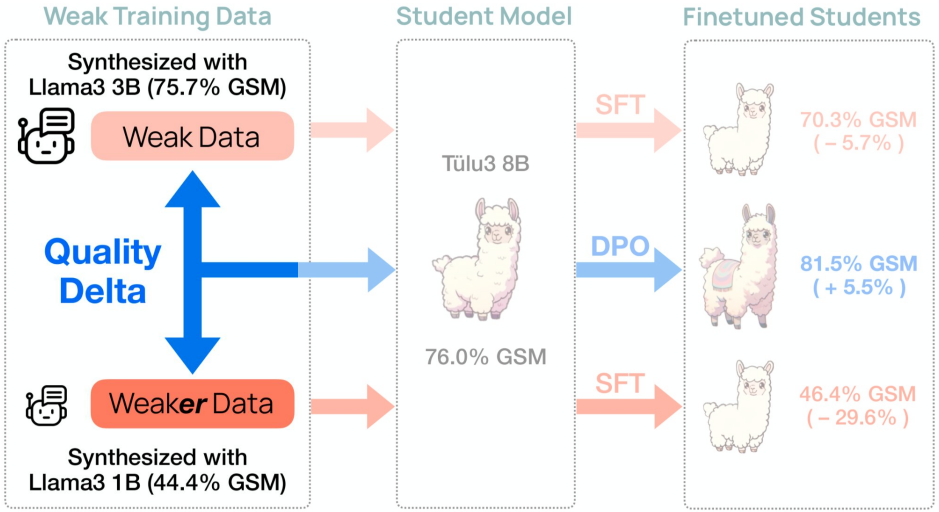


Limitation: **<u>High collection cost of strong data.</u>**

- May require expert annotations.
- May exceed current human expertise.

This paper: **Can we instead train on the Δ between two weaker models?**

Introduction

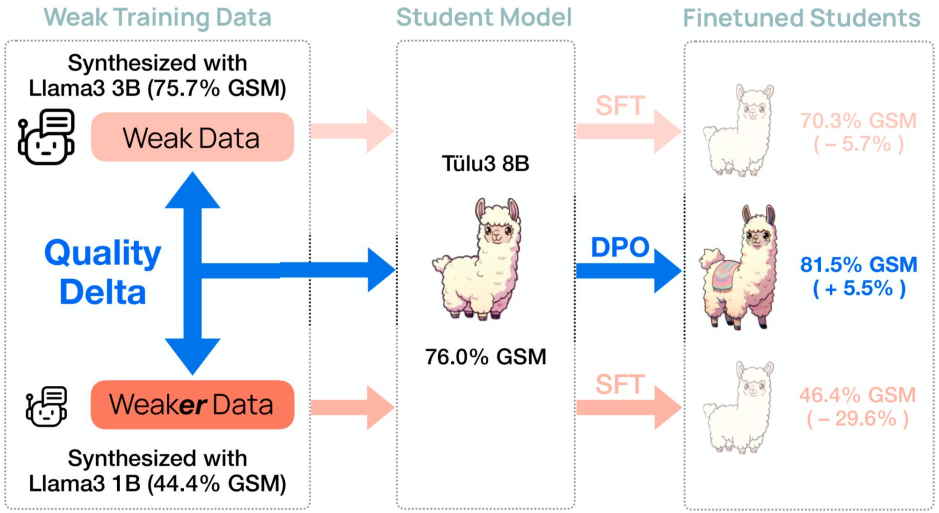


Limitation: **<u>High collection cost of strong data.</u>**

- May require expert annotations.
- May exceed current human expertise.

This paper: **Can we instead train on the Δ between two weaker models?**

Introduction



Spoiler: **Good performance with vastly less supervision.**

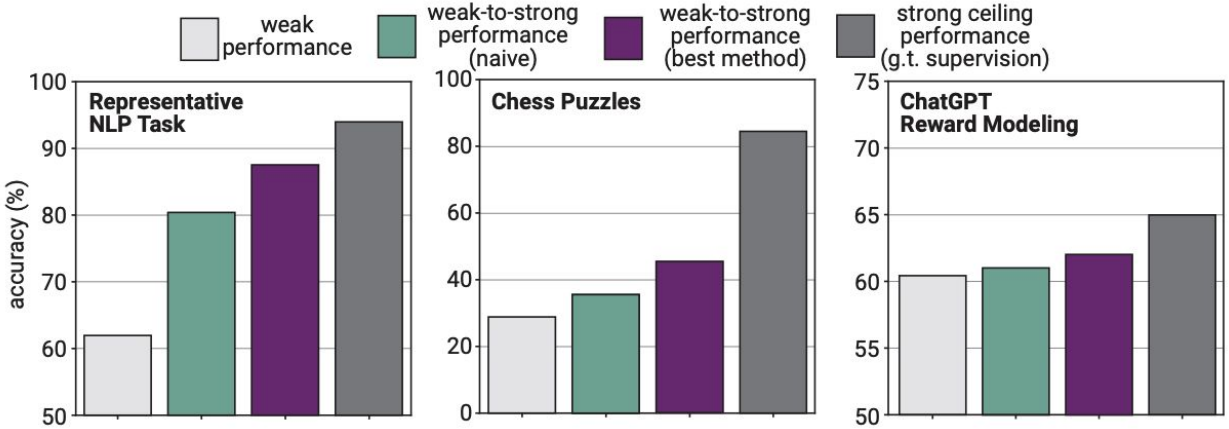
Limitation: <u>High collection cost of strong data.</u>

- May require expert annotations.
- May exceed current human expertise.

This paper: **Can we instead train on the Δ between two weaker models?**

Background work

Weak-to-strong Generalization



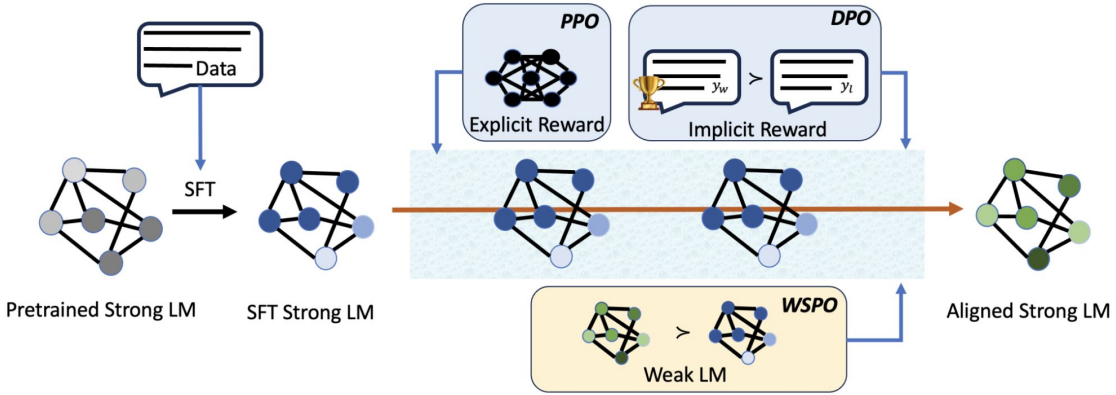
Conclusions:

1. Strong pretrained models naturally generalize beyond their weak supervisors.
2. Naively finetuning on weak supervision is not enough.
3. Improving weak-to-strong generalization is tractable.

Collin Burns et al, "Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision," 2023.

Background work

Weak-to-strong Preference Optimization (WSPO)



First, align a weak model (SFT/DPO). Then,

Define: a new WSPO objective that minimizes a new $\mathcal{L}_{\text{WSPO}}$: leverage the **change in the weak model's outputs** before and after alignment as a **supervisory signal** to guide the alignment of the stronger reference model.

Wenhong Zhu et al, "Weak-to-Strong Preference Optimization: Stealing Reward from Weak Aligned Model," 2025.

The Delta Learning Hypothesis

"Data with high *absolute quality* is not strictly necessary to improve language models."

Intuition:

- Learn to extrapolate on the **relative quality difference** (Δ) between paired samples.
- Can work even if **neither sample alone is stronger** than the model being trained.

The Delta Learning Hypothesis

"Data with high *absolute quality* is not strictly necessary to improve language models."

x : Prompt y_c : Response chosen by LM. y_r : Response rejected by LM.

Training a model *M* on paired responses (x, y_c, y_r) enables learning from the **relative quality difference** between y_c and y_r .

Let $\mu(x, y)$ represent the utility of a response y to prompt x.

The Δ Learning Hypothesis posits that $\exists\ (x, y_c, y_r)$ where $\mu(x, y_c) > \mu(x, y_r)$ and:

1. **Low absolute utility:**

   $\mu(x, y_c) \leq M$'s capability (SFT would hurt performance)

2. **Extrapolated gain:**

   Preference tuning improves M beyond $\mu(x, y_c)$.

Warm-up Case Study

- Tuning Llama-3-IT checkpoints.
- UltraFeedback-Weak preference dataset.
- DPO to prefer "weak" vs "weak***er***" responses yields gains.
- SFT on the preferred responses hurts.

Model / Training	MMLU	AE2	Full Avg.
LLAMA-3.2-3B-INST.	62.9	18.7	57.8
+ UF-WEAK SFT	61.8	12.3	54.0
+ UF-WEAK DPO	**64.0**	**22.4**	**59.0**
LLAMA-3.1-8B-INST.	71.8	24.9	63.9
+ UF-WEAK SFT	65.7	8.9	56.1
+ UF-WEAK DPO	**72.0**	**26.3**	**64.5**

Blue indicates gain over baseline.

Orange indicates degradation.

[see Table 1]

Warm-up Case Study

- Tuning Llama-3-IT checkpoints.
- UltraFeedback-Weak preference dataset.
- DPO to prefer "weak" vs "weak***er***" responses yields gains.
- SFT on the preferred responses hurts.

Blue indicates gain over baseline.

Orange indicates degradation.

Model / Training	MMLU	AE2	Full Avg.
LLAMA-3.2-3B-INST.	62.9	18.7	57.8
+ UF-WEAK SFT	61.8	12.3	54.0
+ UF-WEAK DPO	**64.0**	**22.4**	**59.0**
LLAMA-3.1-8B-INST.	71.8	24.9	63.9
+ UF-WEAK SFT	65.7	8.9	56.1
+ UF-WEAK DPO	**72.0**	**26.3**	**64.5**

DPO: *both y_c and y_r responses derived from models weaker than Llama 3.*

[see Table 1]

Warm-up Case Study

- Tuning Llama-3-IT checkpoints.
- UltraFeedback-Weak preference dataset.
- DPO to prefer "weak" vs "weak***er***" responses yields gains.
- SFT on the preferred responses hurts.

Blue indicates gain over baseline.

Orange indicates degradation.

Model / Training	MMLU	AE2	Full Avg.
LLAMA-3.2-3B-INST.	62.9	18.7	57.8
+ UF-WEAK SFT	61.8	12.3	54.0
+ UF-WEAK DPO	**64.0**	**22.4**	**59.0**
LLAMA-3.1-8B-INST.	71.8	24.9	63.9
+ UF-WEAK SFT	65.7	8.9	56.1
+ UF-WEAK DPO	**72.0**	**26.3**	**64.5**

SFT: *only on y_c.*

[see Table 1]

Results: Controlled Experiments 1

μ1: Number of **bold sections in *y***

- We expect Δ learning to leverage *relative differences* in *y_c* & *y_r* to improve μ1.
- **SFT only helps when** *y_c* > *y_r* and otherwise hurts.
- Δ Learning shows much more extrapolation.

Model/Algorithm	Chosen Res.	Rejected Res.	Section Δ	# Sections Generated
LLAMA-3.2-3B-INST. (Baseline)	—	—	—	5.9
+ SFT	9 sections	—	—	24.6 (+ 18.7)
+ SFT	3 sections	—	—	4.4 (- 1.5)
+ SFT	2 sections	—	—	2.9 (- 3.0)
+ DPO	3 sections	2 sections	+1	81.1 (+ 75.2)
+ DPO	2 sections	3 sections	−1	1.1 (- 4.8)
+ DPO	3 sections	3 sections	0	6.1 (+ 0.2)

[see Table 2]

Results: Controlled Experiments 2

- Can Δ Learning teach *M* from itself?
  - What if *y_c* was from *M* directly? Not learning from a stronger teacher, by construction.
  - *y_r* is obtained from a smaller model in the same family.
  -

Model/Training Setup	MMLU	MATH	GSM	AEval2	IFEval	BBH	TQA	HEval+	Avg.
LLAMA-3.2-3B-INSTRUCT (Weaker)	62.9	39.6	75.7	18.7	76.5	61.6	50.6	76.8	57.8
LLAMA-3.1-8B-INSTRUCT (Baseline)	71.8	43.0	83.7	24.9	78.2	72.7	55.1	81.6	63.9
+ SFT (self-generated responses)	**72.2**	42.2	82.9	24.3	76.3	72.1	53.4	78.0	62.7
+ DPO (self-generated over weaker)	72.0	**43.5**	**84.2**	**25.7**	**80.0**	71.4	**55.6**	**82.2**	**64.3**
+ DPO (weaker over self-generated)	70.9	42.3	83.4	22.9	78.6	72.1	54.6	80.5	63.2

Blue: gain over baseline.

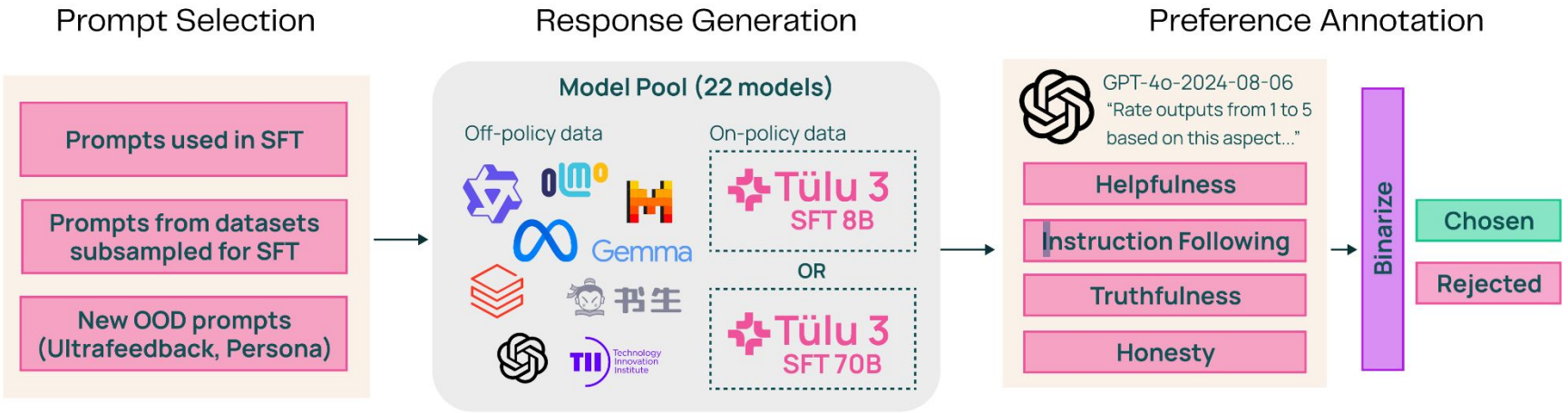
Orange: degradation.

[see Table A3]

Post-training LMs with Δ Learning

Tülu 3: SoTA Open Post-training recipe.

Baseline pipeline for generating and scaling preference data:

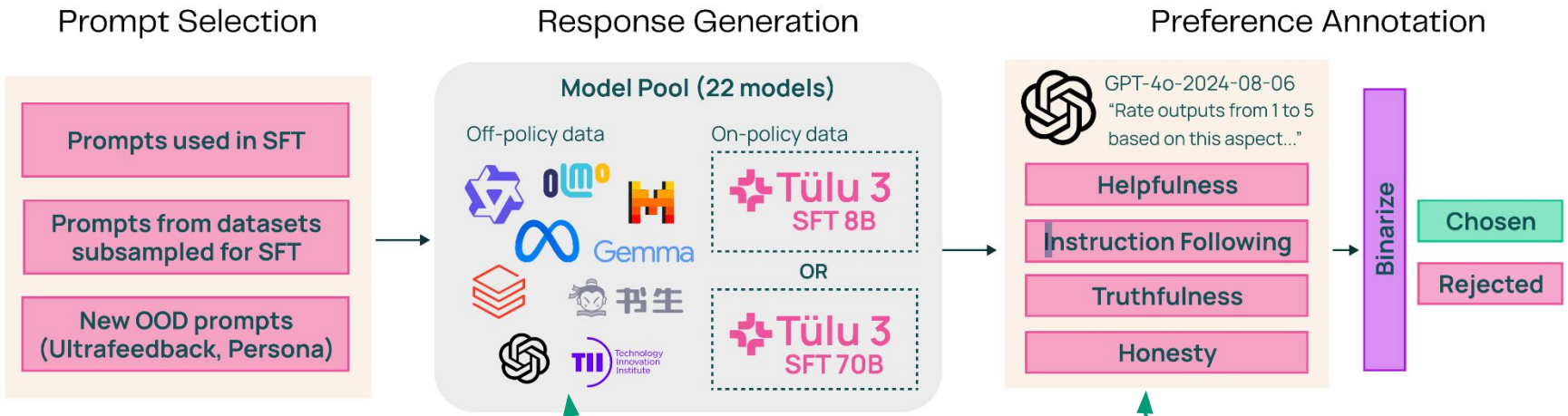


[Nathan Lambert et al, "Tulu 3: Pushing Frontiers in Open Language Model Post-Training," 2025; see figure 7.]

Post-training LMs with Δ Learning

Tülu 3: SoTA Open Post-training recipe.

Baseline pipeline for generating and scaling preference data:

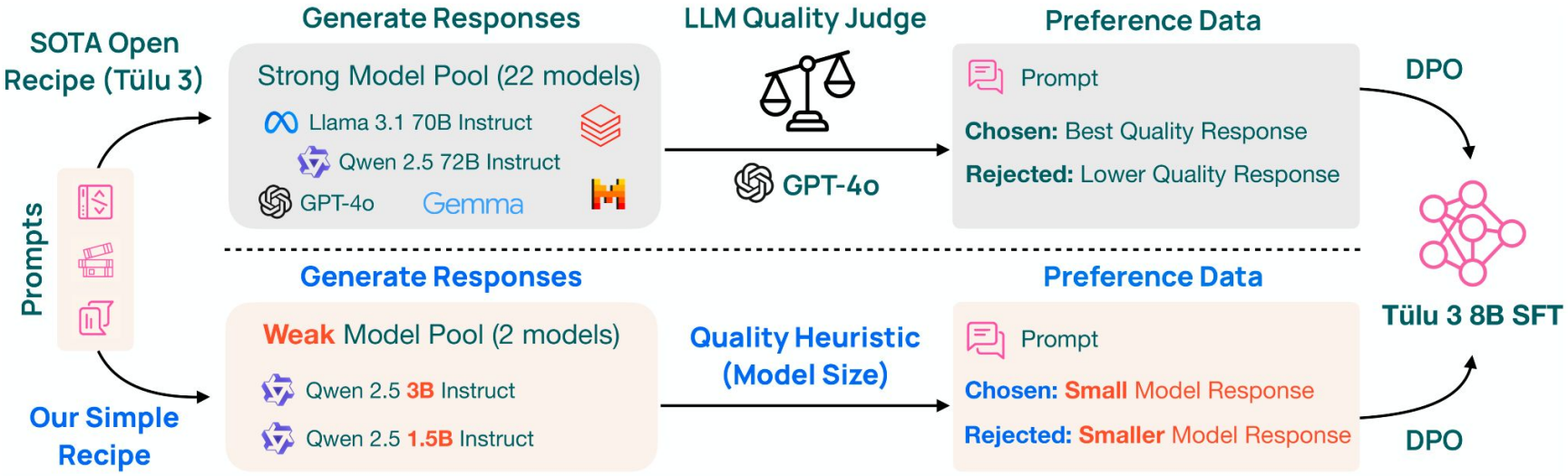


Requires** access to **strong supervisors:

[Nathan Lambert et al, "Tulu 3: Pushing Frontiers in Open Language Model Post-Training," 2025; see figure 7.]

Post-training LMs with Δ Learning

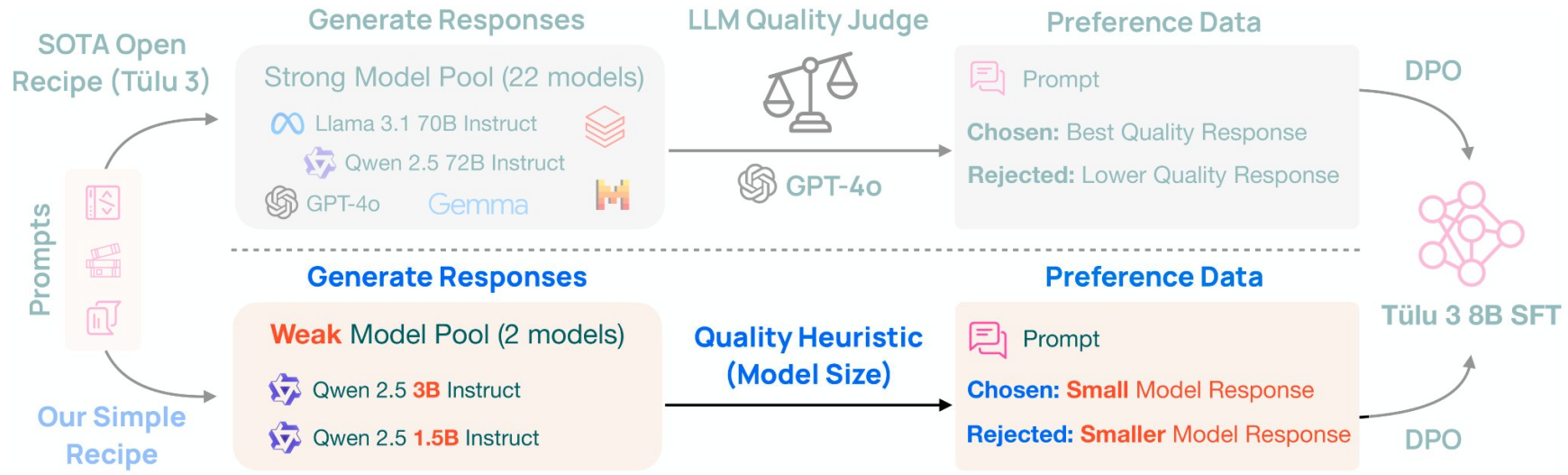
Instead,



[Figure 2]

Post-training LMs with Δ Learning

Instead,



[Figure 2]

Post-training LMs with Δ Learning

Data Generation:

Generate **all chosen responses** using a small model (below Tülu-3-8B-SFT.) **6% of original FLOPs** required.

Forming Preference Pairs:

Remove GPT-4o quality annotations. Use **model size as a proxy** for quality.

Pair each small model generation (y_c) with a small***er*** model from the same family (y_r).

[Figure 2]

Results: Post-training LMs with Δ Learning

Model/Preference Data	MMLU	PopQA	MATH	GSM	AE2	IFEval	BBH	DROP	TQA	HEval	HEval+	Avg.
LLAMA-3.2-1B-INSTRUCT	46.1	13.9	21.1	44.4	8.8	54.5	40.2	32.2	40.0	64.8	60.0	38.7
LLAMA-3.2-3B-INSTRUCT	62.9	19.4	39.6	75.7	18.7	76.5	61.6	48.5	50.6	79.7	76.8	55.5
QWEN-2.5-0.5B-INSTRUCT	46.2	10.1	27.2	39.2	3.3	28.8	32.2	25.3	45.4	60.5	58.9	34.3
QWEN-2.5-1.5B-INSTRUCT	59.7	15.4	41.6	66.2	7.2	44.2	45.9	14.1	46.5	83.0	79.8	45.8
QWEN-2.5-3B-INSTRUCT	69.5	15.7	63.1	77.7	17.8	64.0	57.6	31.5	57.2	90.5	87.4	57.5
TÜLU-3-8B-SFT	66.1	29.6	31.2	76.0	12.2	71.3	69.2	61.2	46.8	**86.2**	79.8	57.2
+ Llama 3.2 3B over 1B	68.8	30.3	40.9	81.5	24.9	75.0	70.0	60.7	54.2	84.7	81.1	61.1
+ Qwen 2.5 1.5B over 0.5B	67.4	29.9	39.9	79.8	15.8	72.5	**70.8**	61.8	52.1	83.7	78.1	59.3
+ Qwen 2.5 3B over 1.5B	69.4	**31.7**	**42.6**	83.4	**36.1**	78.6	69.4	62.0	**57.7**	84.4	**81.7**	**63.4**
+ Tülu 3 Preference Dataset	**69.8**	30.3	**42.6**	**84.2**	32.8	**80.4**	69.2	**62.5**	56.1	84.7	80.8	63.0

Blue: best setup.

Pink: Tülu 3 preference data.

Results: Post-training LMs with Δ Learning

Model/Preference Data	MMLU	PopQA	MATH	GSM	AE2	IFEval	BBH	DROP	TQA	HEval	HEval+	Avg.
LLAMA-3.2-1B-INSTRUCT	46.1	13.9	21.1	44.4	8.8	54.5	40.2	32.2	40.0	64.8	60.0	38.7
LLAMA-3.2-3B-INSTRUCT	62.9	19.4	39.6	75.7	18.7	76.5	61.6	48.5	50.6	79.7	76.8	55.5
QWEN-2.5-0.5B-INSTRUCT	46.2	10.1	27.2	39.2	3.3	28.8	32.2	25.3	45.4	60.5	58.9	34.3
QWEN-2.5-1.5B-INSTRUCT	59.7	15.4	41.6	66.2	7.2	44.2	45.9	14.1	46.5	83.0	79.8	45.8
QWEN-2.5-3B-INSTRUCT	69.5	15.7	63.1	77.7	17.8	64.0	57.6	31.5	57.2	90.5	87.4	57.5
TÜLU-3-8B-SFT	66.1	29.6	31.2	76.0	12.2	71.3	69.2	61.2	46.8	**86.2**	79.8	57.2
+ Llama 3.2 3B over 1B	68.8	30.3	40.9	81.5	24.9	75.0	70.0	60.7	54.2	84.7	81.1	61.1
+ Qwen 2.5 1.5B over 0.5B	67.4	29.9	39.9	79.8	15.8	72.5	**70.8**	61.8	52.1	83.7	78.1	59.3
+ Qwen 2.5 3B over 1.5B	69.4	**31.7**	**42.6**	83.4	**36.1**	78.6	69.4	62.0	**57.7**	84.4	**81.7**	**63.4**
+ Tülu 3 Preference Dataset	**69.8**	30.3	**42.6**	**84.2**	32.8	**80.4**	69.2	**62.5**	56.1	84.7	80.8	63.0

Blue: best setup.

Pink: Tülu 3 preference data.

Tuning with Δ Learning *matches Tülu 3 recipe,* ***+0.4 avg gain .***

[Table 3]

Results: Post-training LMs with Δ Learning

Model/Preference Data	MMLU	PopQA	MATH	GSM	AE2	IFEval	BBH	DROP	TQA	HEval	HEval+	Avg.
LLAMA-3.2-1B-INSTRUCT	46.1	13.9	21.1	44.4	8.8	54.5	40.2	32.2	40.0	64.8	60.0	38.7
LLAMA-3.2-3B-INSTRUCT	62.9	19.4	39.6	75.7	18.7	76.5	61.6	48.5	50.6	79.7	76.8	55.5
QWEN-2.5-0.5B-INSTRUCT	46.2	10.1	27.2	39.2	3.3	28.8	32.2	25.3	45.4	60.5	58.9	34.3
QWEN-2.5-1.5B-INSTRUCT	59.7	15.4	41.6	66.2	7.2	44.2	45.9	14.1	46.5	83.0	79.8	45.8
QWEN-2.5-3B-INSTRUCT	69.5	15.7	63.1	77.7	17.8	64.0	57.6	31.5	57.2	90.5	87.4	57.5
TÜLU-3-8B-SFT	66.1	29.6	31.2	76.0	12.2	71.3	69.2	61.2	46.8	**86.2**	79.8	57.2
+ Llama 3.2 3B over 1B	68.8	30.3	40.9	81.5	24.9	75.0	70.0	60.7	54.2	84.7	81.1	61.1
+ Qwen 2.5 1.5B over 0.5B	67.4	29.9	39.9	79.8	15.8	72.5	**70.8**	61.8	52.1	83.7	78.1	59.3
+ Qwen 2.5 3B over 1.5B	69.4	**31.7**	**42.6**	83.4	**36.1**	78.6	69.4	62.0	**57.7**	84.4	**81.7**	**63.4**
+ Tülu 3 Preference Dataset	**69.8**	30.3	**42.6**	**84.2**	32.8	**80.4**	69.2	**62.5**	56.1	84.7	80.8	63.0

Blue: best setup.

Pink: Tülu 3 preference data.

Increased avg. gain in all 3 weak preference dataset settings.

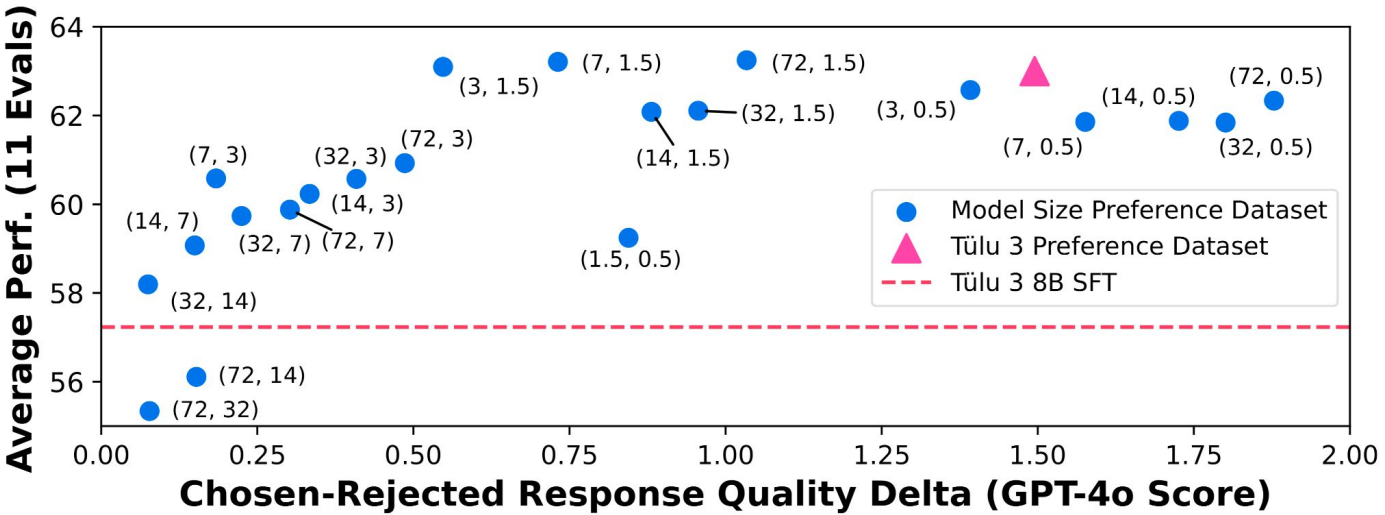
[Table 3]

Analysis

Four potential factors affecting tuning performance:

1. **Magnitude of** the quality **Δ** between y_c & y_r.
2. **Absolute quality** of y_c.
3. Model size-based **reward heuristic**.
4. Choice of **Tülu-3-8B-SFT** as the base model.

Analysis



- Magnitude of Δ strongly predicts tuning performance; strong correlation until Δ ≈ 0.55.
- Not all positive Δ drive learning. Hypothesis: both y_c & y_r are stronger than M, hurts gains?

Analysis

Model/Preference Data	MMLU	PopQA	MATH	GSM	AE2	IFEval	BBH	DROP	TQA	HEval	HEval+	Avg.
TÜLU-3-8B-SFT	66.1	29.6	31.2	76.0	12.2	71.3	69.2	61.2	46.8	**86.2**	79.8	57.2
+ Model size heuristic	69.4	**31.7**	**42.6**	83.4	**36.1**	78.6	**69.4**	**62.0**	57.7	84.4	**81.7**	**63.4**
+ GPT-4o judge reward	**69.9**	31.5	40.6	**83.9**	29.9	**79.5**	66.5	61.2	**62.4**	85.7	80.8	62.9
OLMO-2-7B-SFT	61.4	**23.6**	25.3	73.5	8.4	66.5	49.3	59.6	48.6	70.0	63.8	50.0
+ Qwen 2.5 3B over 1.5B	**62.9**	**23.6**	30.0	80.6	**31.0**	71.5	**50.9**	59.3	**56.3**	**72.6**	**66.6**	**55.0**
+ OLMo 2 Preference Dataset	61.9	23.5	**30.3**	**83.1**	27.7	**72.3**	**50.9**	**60.2**	56.0	70.7	66.2	54.8

- Model size as a heuristic: 80.5% agreement rate with GPT-4o.
- Choice of SFT model: Δ learning generalizes to tuning OLMo-2-7B-SFT.

References

1. Wenhong Zhu et al, "Weak-to-Strong Preference Optimization: Stealing Reward from Weak Aligned Model," 2025.
2. Collin Burns et al, "Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision," 2023.
3. Scott Geng et al, "The Delta Learning Hypothesis: Preference Tuning on Weak Data can Yield Strong Gains," 2025.
4. Nathan Lambert et al, "Tulu 3: Pushing Frontiers in Open Language Model Post-Training," 2025.

also DPO/PPO/