Research in the Wild

What it takes to push academic work to the (open) frontier

Scott Geng x Nathan Lambert, July 2026

Part 1: The Research

Preference Tuning Basics (2022-2023)

- Dataset = prompt x, chosen y_c, rejected y_r
- RLHF objective (InstructGPT, Ouyang et al. 2022)

$$\max_{\pi} \; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(\cdot|x)} \left[r(x,y) \right] - \beta \, \mathbb{E}_{x\sim\mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x) \right) \right]$$

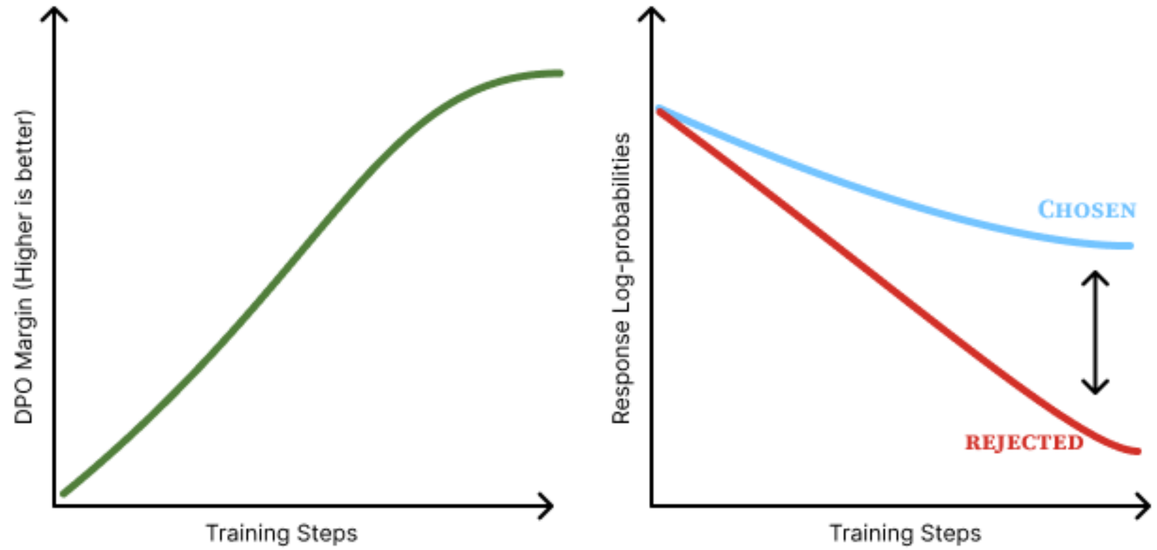
- Direct Preference Optimization (Rafailov et al. 2023)

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}} \left[\log \sigma \left(\beta \left[\log \frac{\pi_\theta(y_c \mid x)}{\pi_{\mathrm{ref}}(y_c \mid x)} - \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\mathrm{ref}}(y_r \mid x)} \right] \right) \right]$$

rlhfbook.com

Algorithm Zoo (2023 - 2024)

- SimPO, ORPO, IPO, ODPO, REBEL, APO…
- KL resets, length normalization, NLL regularization…



Tulu 3 / Ultrafeedback (2023 - 2024)

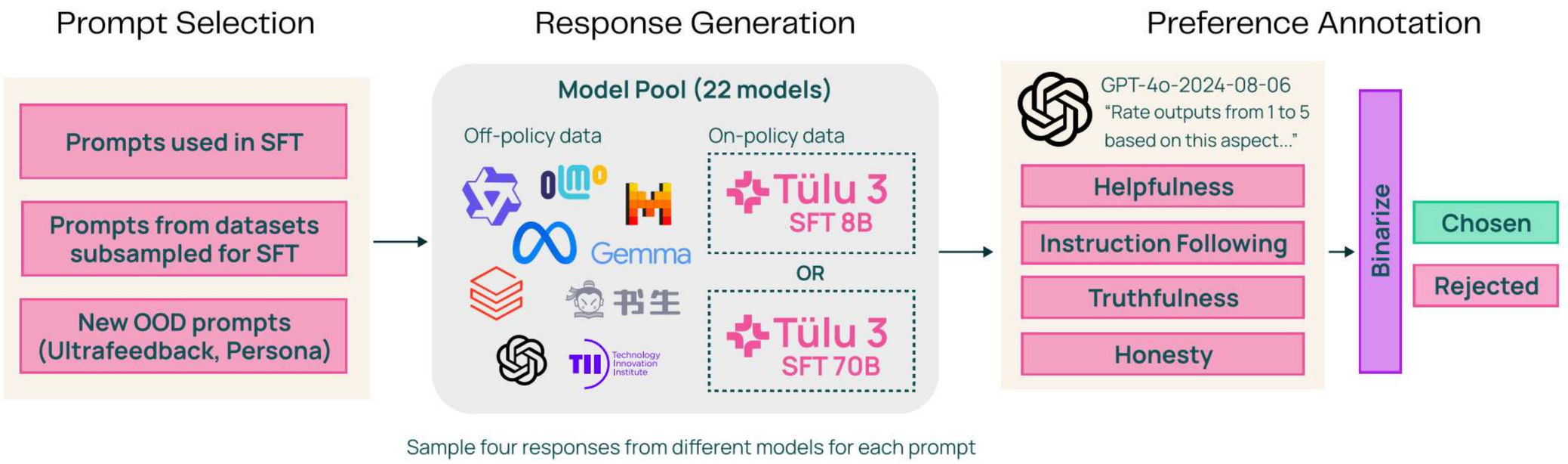
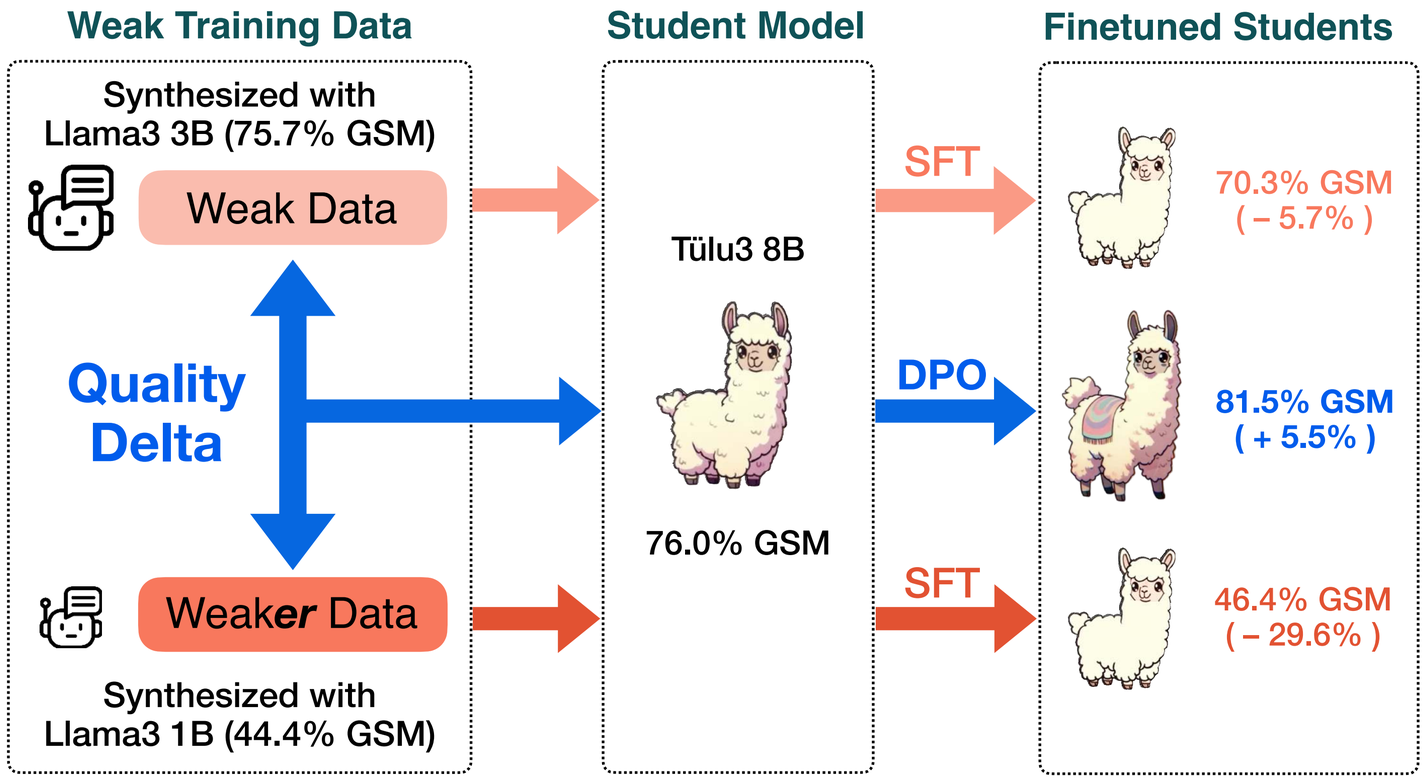


Figure 7 Pipeline for generating and scaling preference data that is based from Ultrafeedback (Cui et al., 2023).

Ivison et al. 2024, Lambert et al. 2024

The Delta Learning Hypothesis (Early 2025)

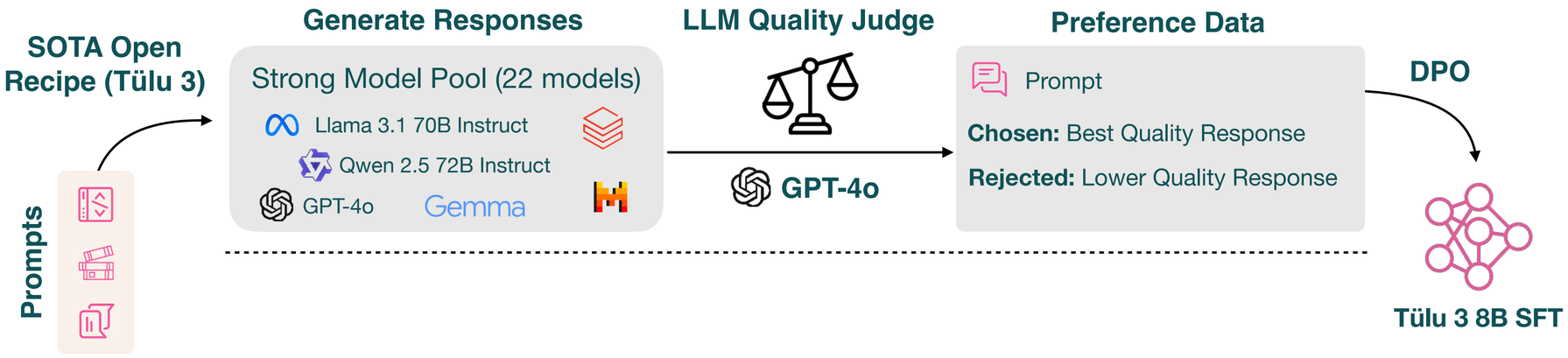
What matters when learning with paired data?



Hypothesis: relative gap matters more than top absolute quality

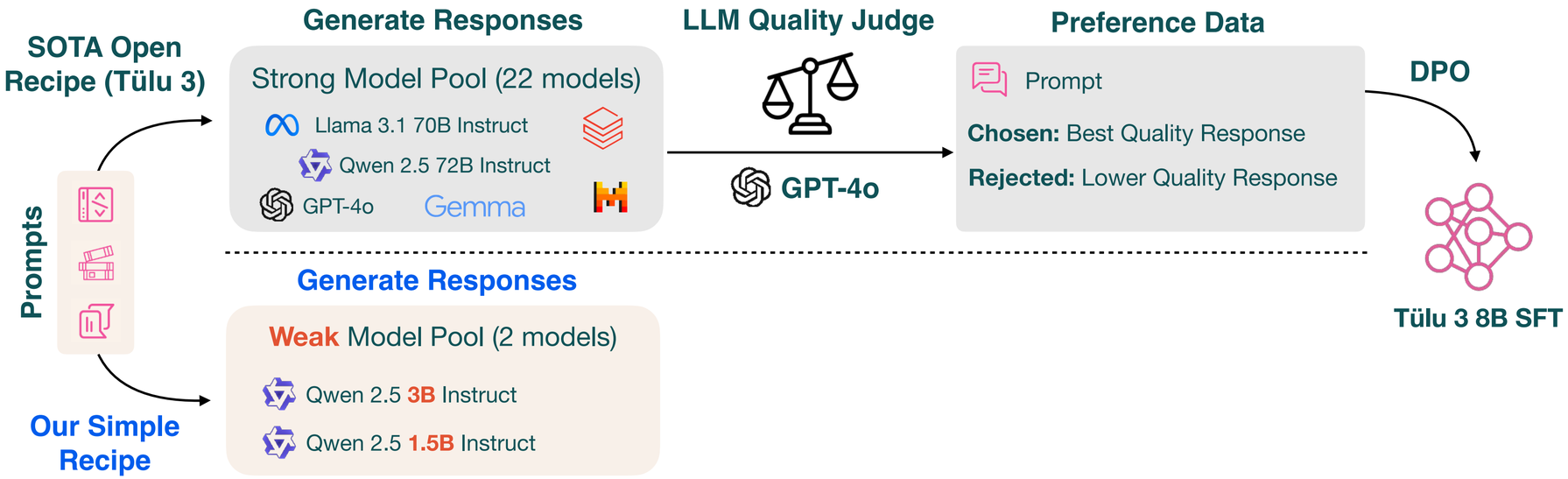
The Delta Learning Hypothesis

Post-training with Ultrafeedback



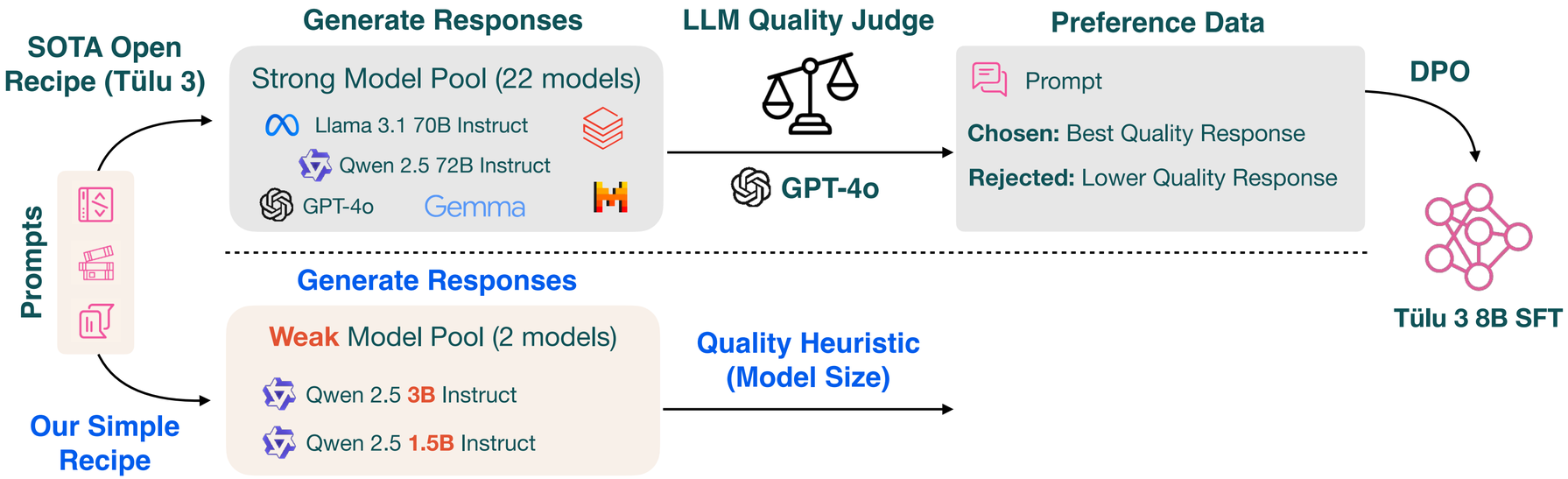
The Delta Learning Hypothesis

Post-training with ~~Ultrafeedback~~ only weak data



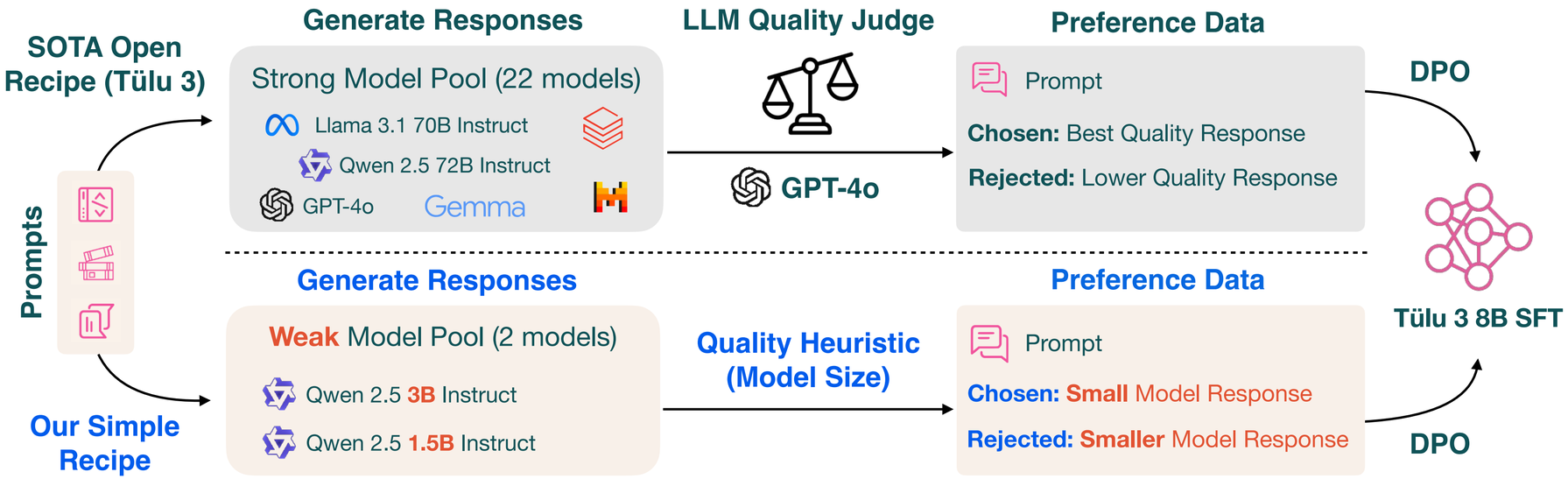
The Delta Learning Hypothesis

Post-training with ~~Ultrafeedback~~ only weak data



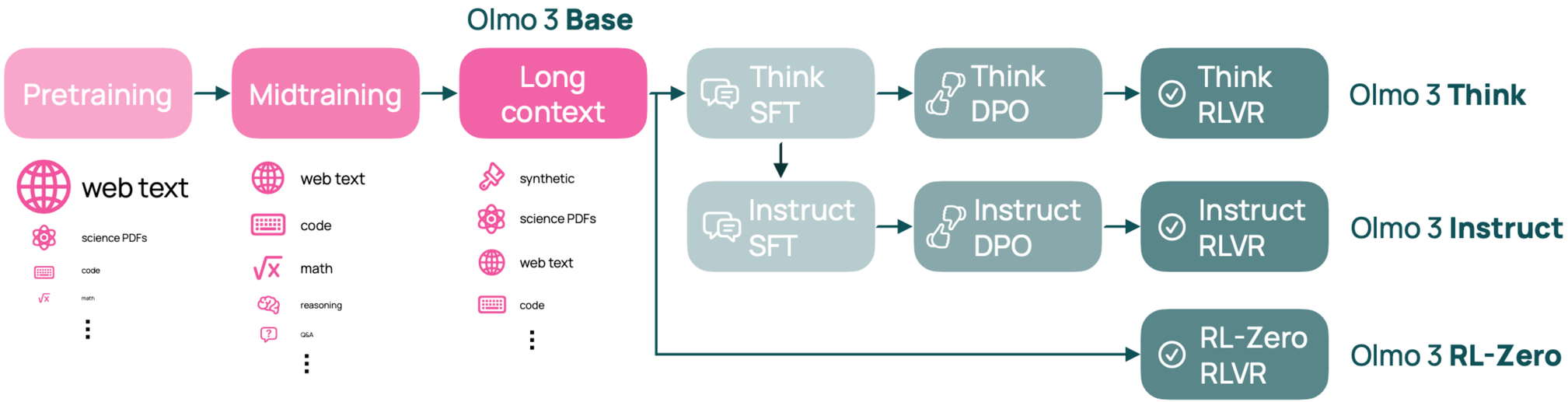
The Delta Learning Hypothesis

Post-training with ~~Ultrafeedback~~ only weak data

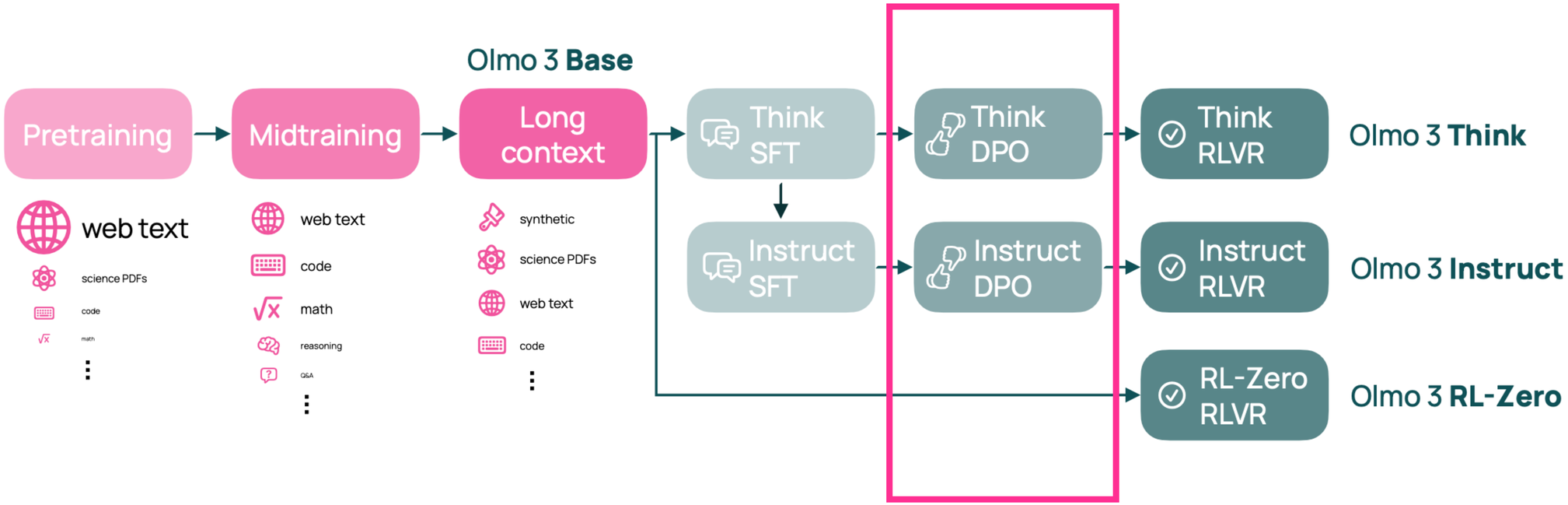


Part 2: The Wild

Olmo 3 (Dec 2025)

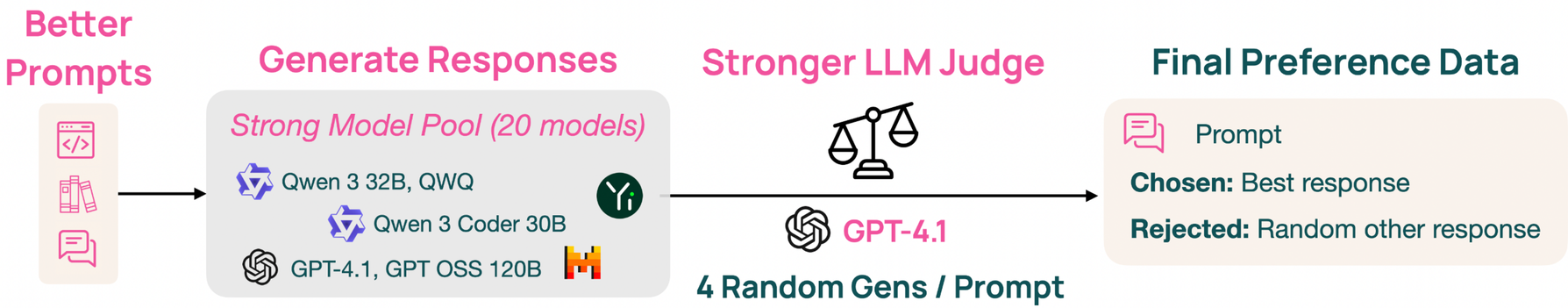


Olmo 3 (Dec 2025)



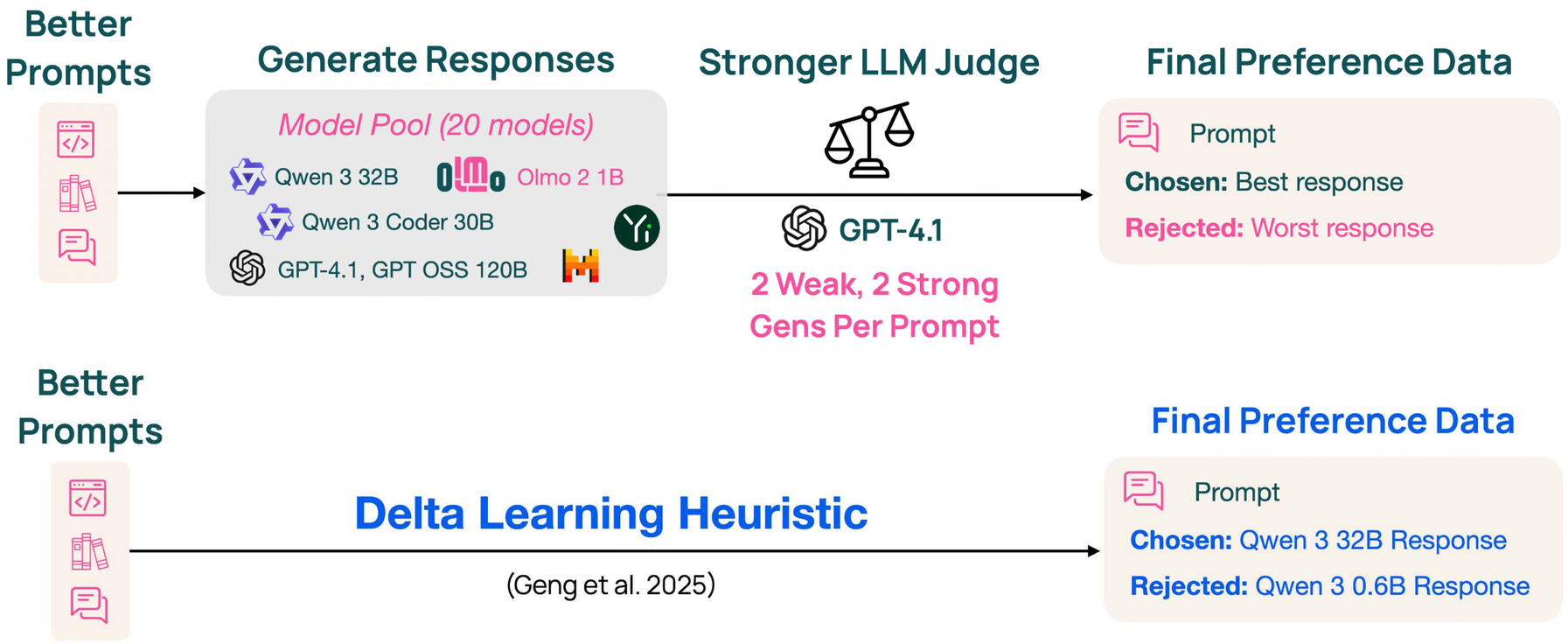
Olmo 3 (Dec 2025)

Where we start



Olmo 3 (Dec 2025)

Where we end - how to fix DPO



Force weak models in the judge pool, pick worst response, and use delta learning heuristic pairs

SmolLM3 converged to a similar preference data setup (Bakouch et al. 2025)

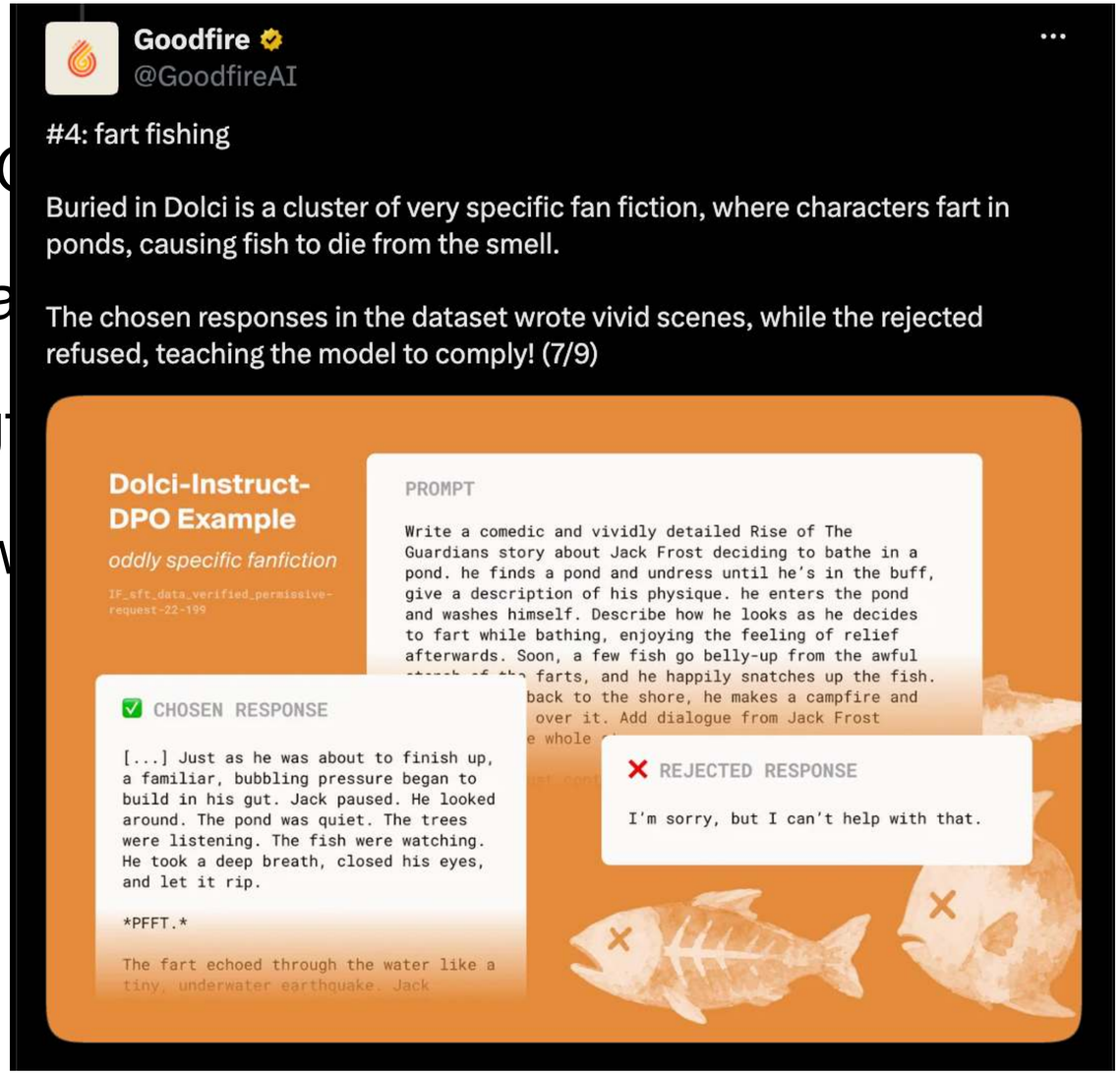
Challenge 1: Unstudied data details

- “My name is ChatGPT / Deepseek / Qwen”
- Better on hallucination evals, but more hallucination vibes?
  - Came from length bias
- Prompt mixing = wtf

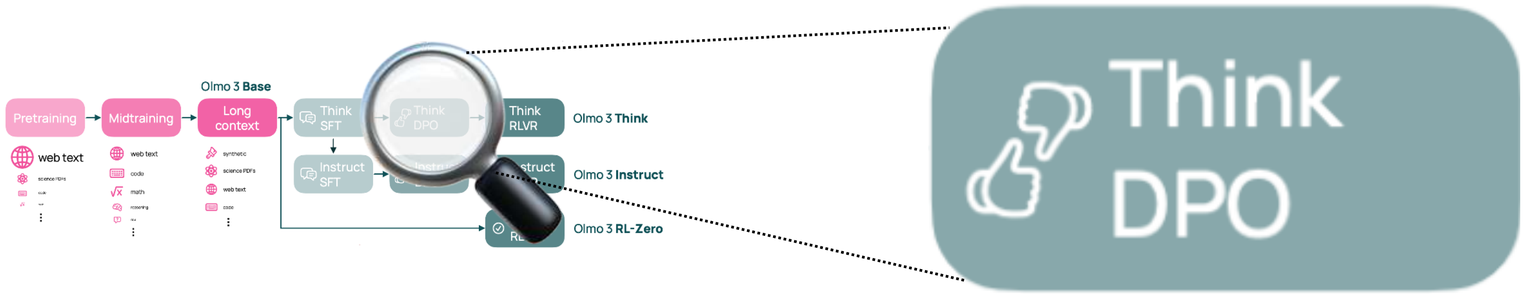
```
yolo_scottmix_1 = {
    "code": 0.1573,
    "math": 0.1296,
    "inst_following": 0.1312,
    "chat": 0.2319,
    ...
}
```

Challenge 1: Unstudied data details

- “My name is ChatG
- Better on hallucina ?
  - Came from leng
- Prompt mixing = w

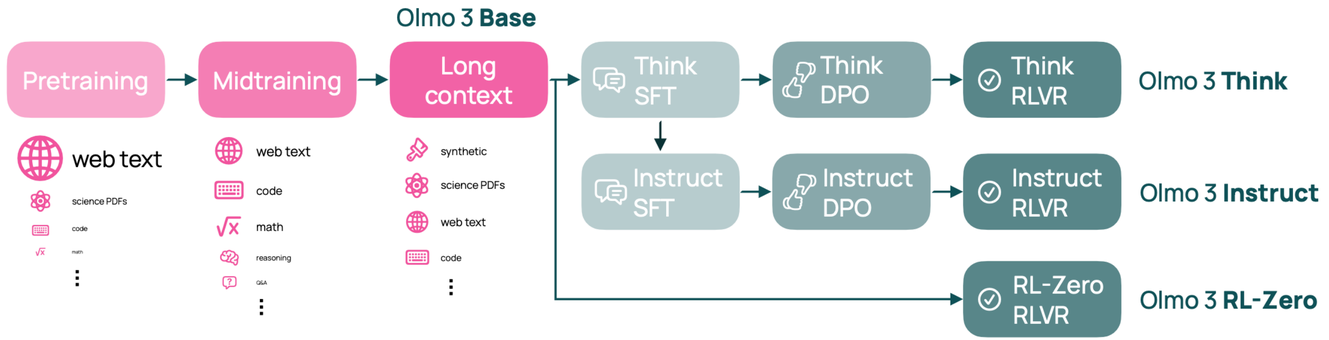


Challenge 2: Organization / method sequencing



- In research, preference tuning is a single, standalone unit of study 🤓

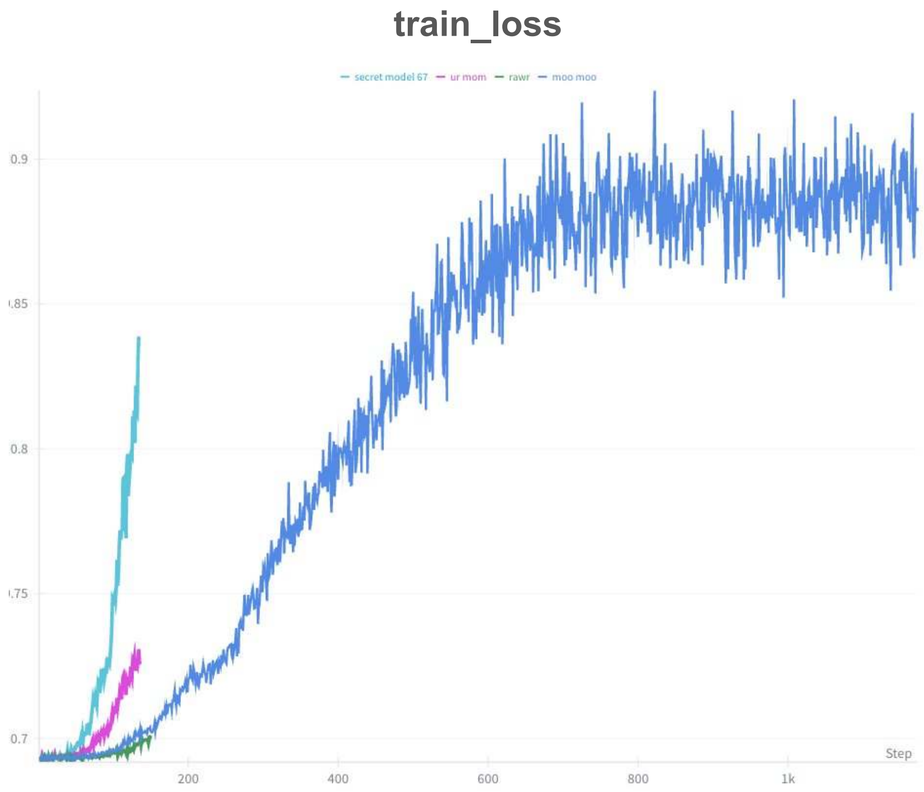
Challenge 2: Organization / method sequencing



- In research, preference tuning is a single, standalone unit of study 🤓
- In practice, one part of a larger pipeline / organization, all moving in parallel 😲

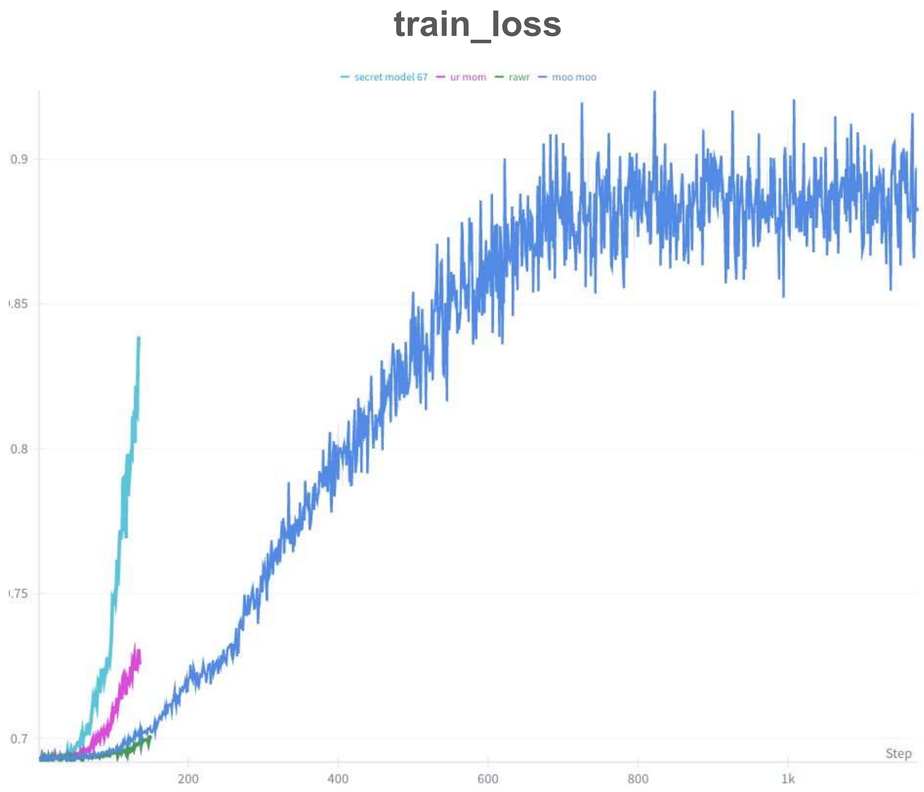
SFT A < SFT B, but SFT A + DPO >> SFT B + DPO

Challenge 3: ???



At larger scale, every random problem that can hit will

Challenge 3: ???



At larger scale, every random problem that can hit will

Fix: NCCL vars + DeepSpeed sharding config + forward pass details + caffeine

The future?