Scalable Post-Training Optimization for Large Language Models

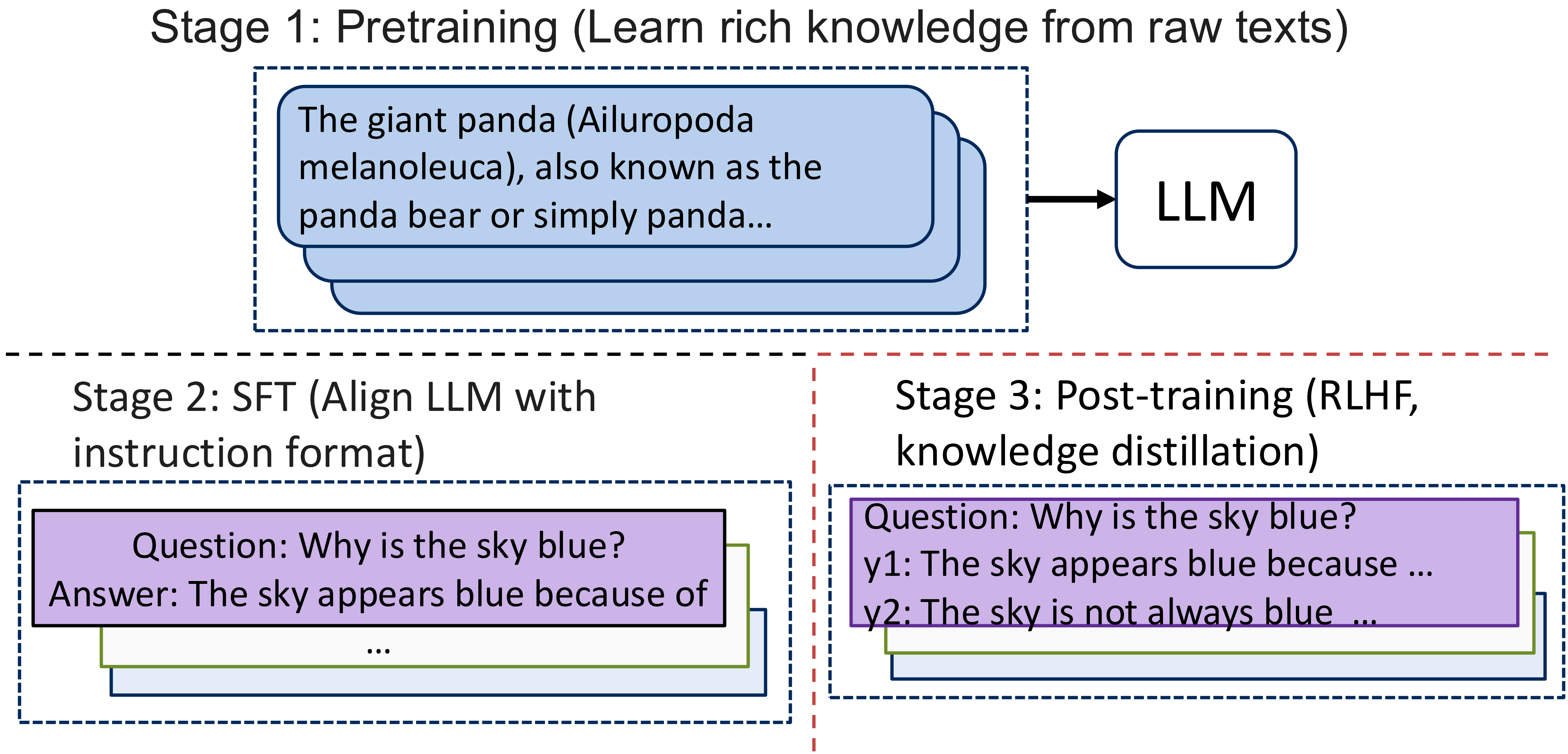
Lei Li

Language Technologies Institute

Carnegie Mellon University

June 19, 2025

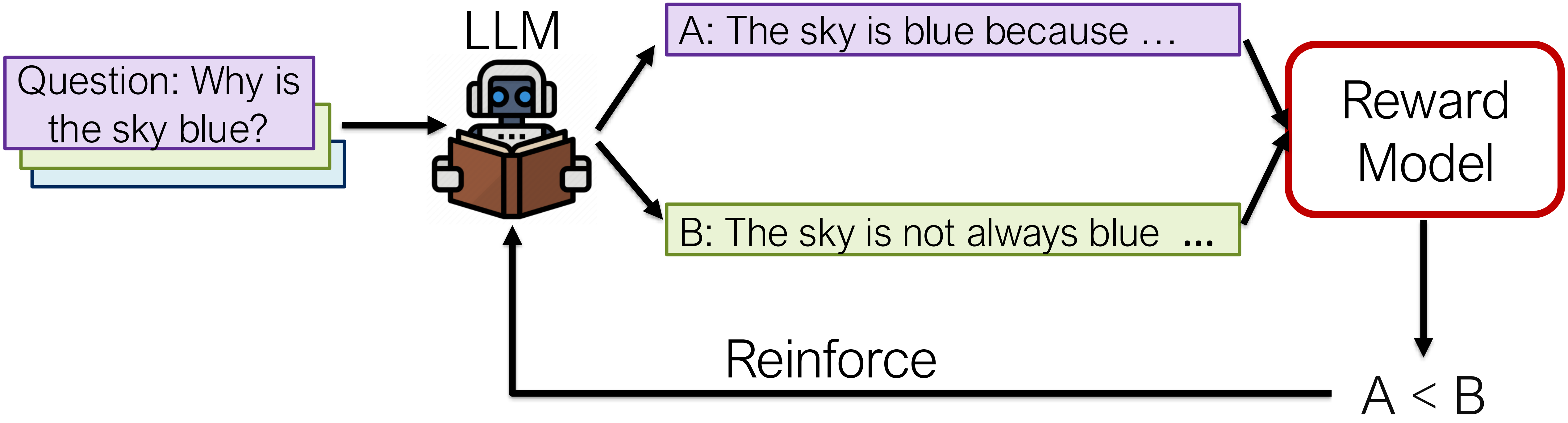
LLM training pipeline



Outline

- Aligning with online preference optimization (BPO)
- Iterative refinement with fine-grained feedback (LLMRefine)
- Learning Optimized Sample Compute Allocation (OSCA)

Learning from Reward / Quality-Estimation Metric(QE)



Ouyang et al. Training language models to follow instructions with human feedback. 2022

PPO training

RL objective:

Maximize reward

Training stability +
Avoid reward hacking

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)} \left[r_\phi(x, y) \right] - \beta \mathbb{D}_{\textbf{KL}} \left[\pi_\theta(y \mid x) \mid\mid \pi_{\text{ref}}(y \mid x) \right]$$

Constrained optimization

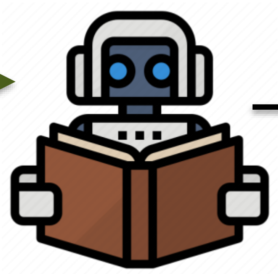
Issues with PPO:

1. Many hyperparameters to tune
2. Involve four different models: ref model, old model, optimized model, reward model

Direct Preference Optimization

Question: Why is the sky blue?

SFTed LLM



The sky appears blue because ...

The sky is not always blue ...

$$(x, y_w, y_l)$$

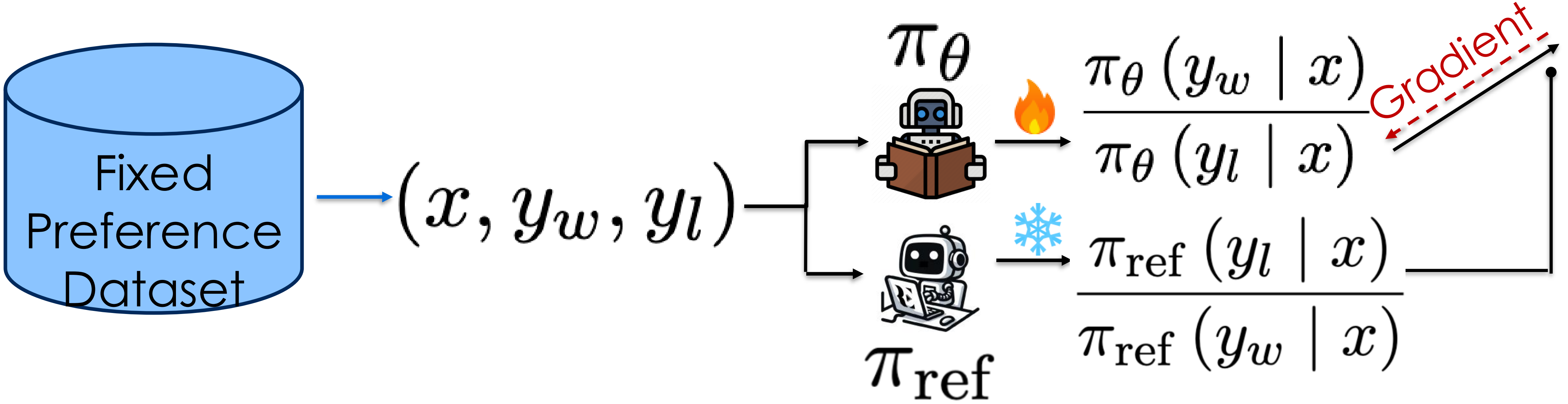
Preferred Dispreferred

We can skip reward model using DPO

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

Illustration of DPO

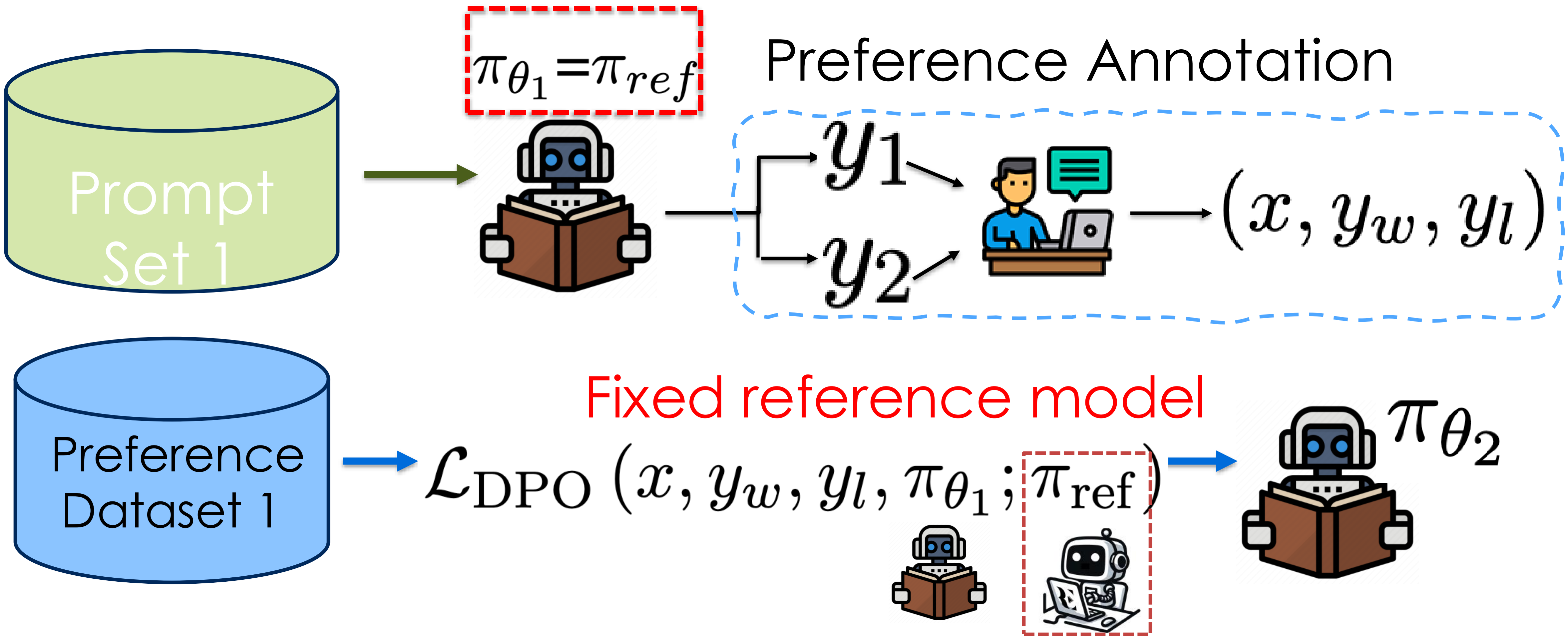
$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)} \right) \right]$$



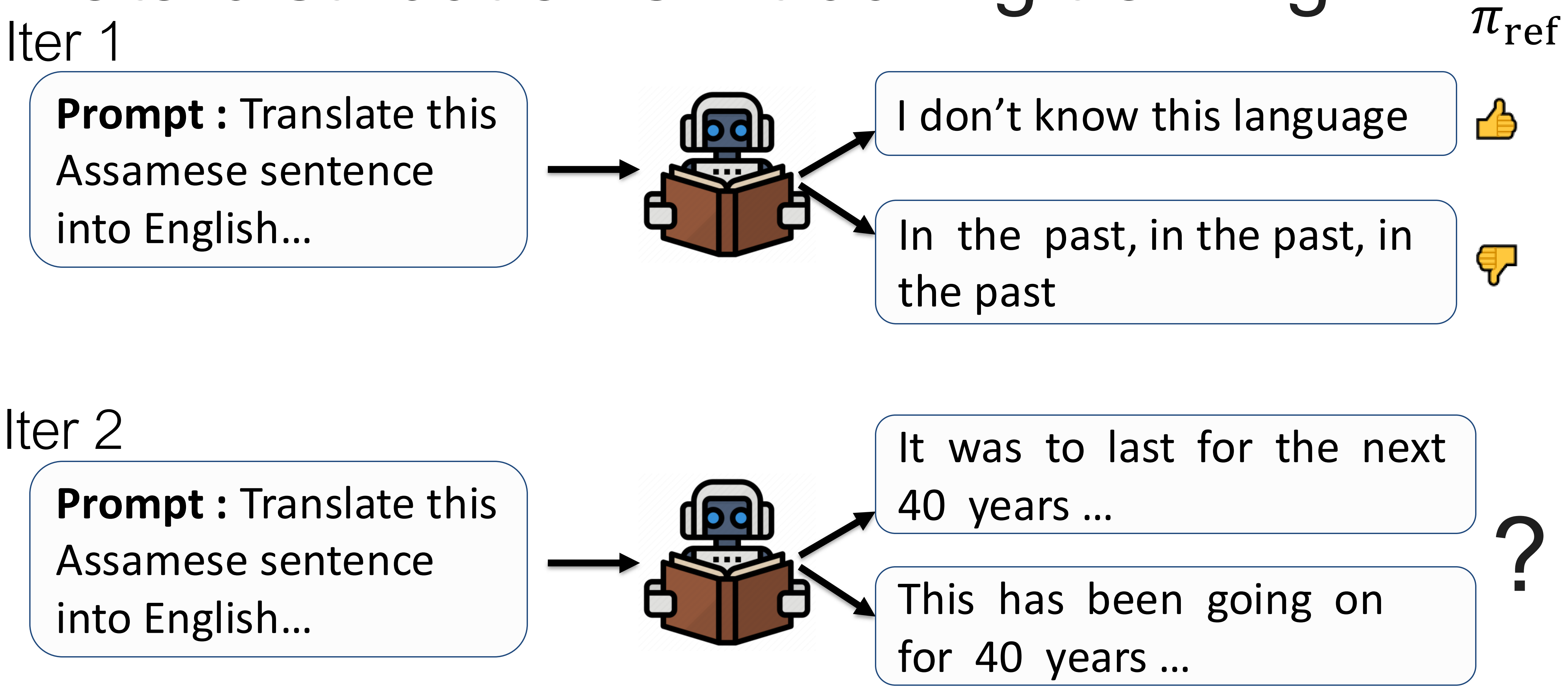
Fixed reference model

Limitation of offline DPO (and online DPO)

Synthetic data distribution shifts



Data distribution shift during training



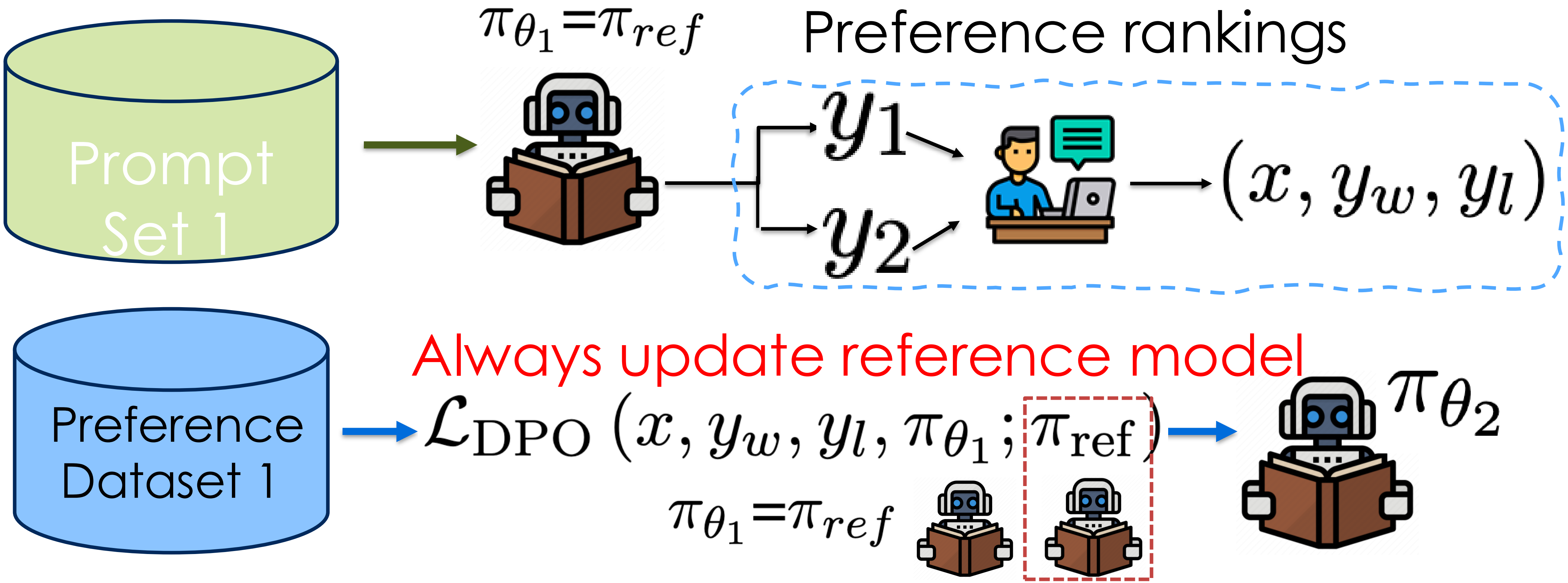
Introducing BPO (B=Behavior)

- Data collection needs to be online

- The reference model needs to be updated and has to be close to the behavior LLM

Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

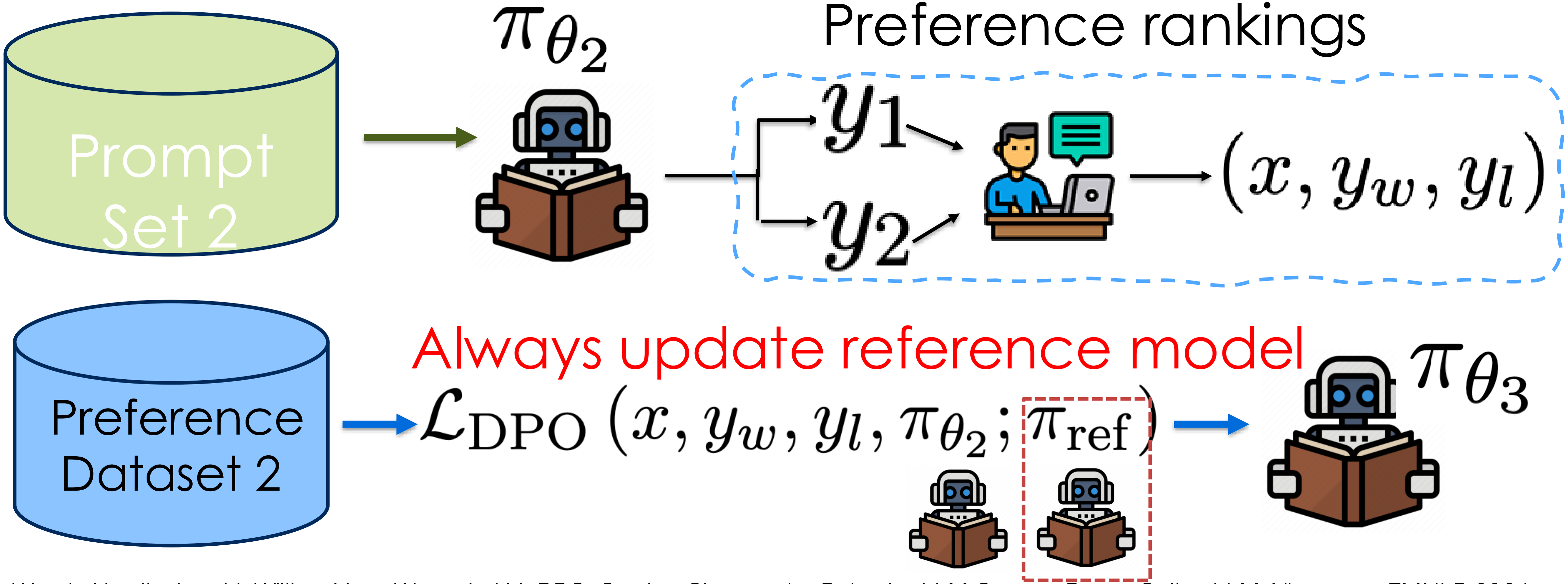
BPO



Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

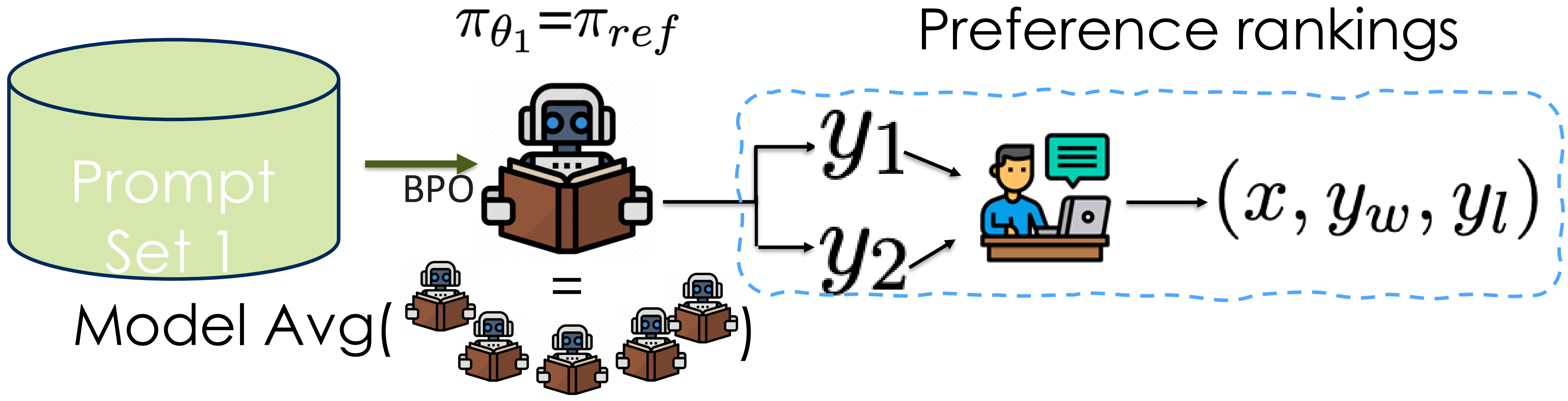
BPO

use new behavior model to generate samples



Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

Practical implementation of BPO (Lora ensemble)



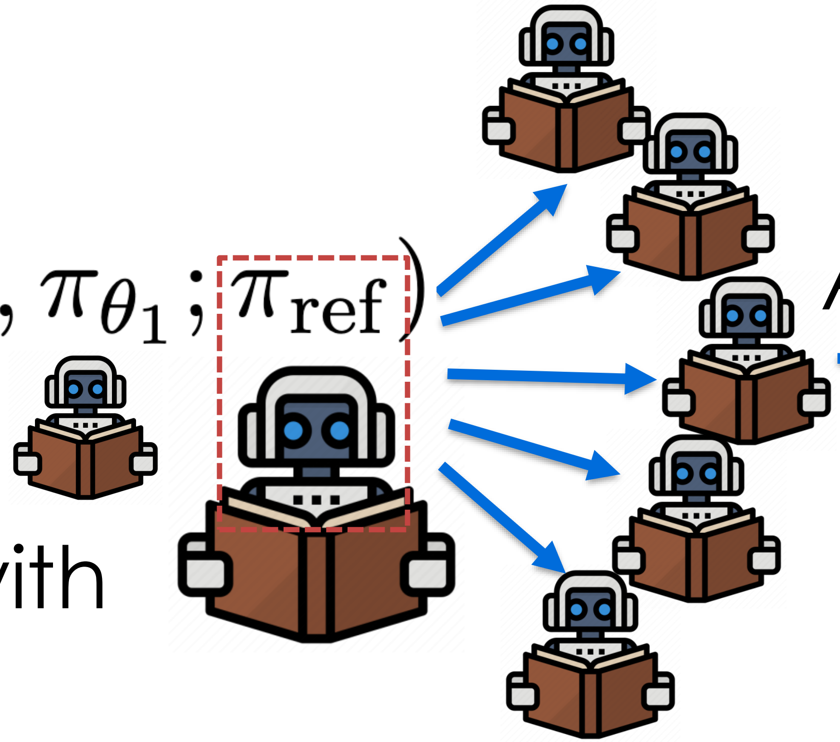
We use model averaged LoRA weights to perform sampling

Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

Practical implementation of BPO (Lora ensemble)

Preference Dataset 1

$$\mathcal{L}_{\text{DPO}}\left(x, y_w, y_l, \pi_{\theta_1}; \boxed{\pi_{\text{ref}}}\right)$$



Avg π_{θ_2}

We update reference model with Model averaged behavior LLM

Each lora weight is updated independently

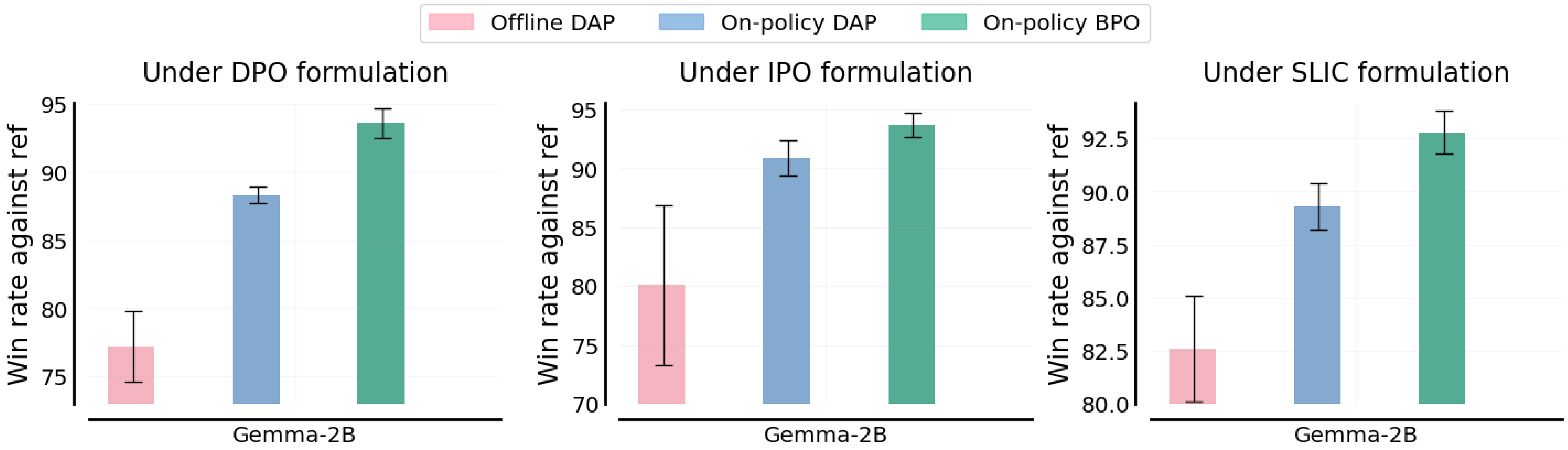
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

Evaluation on TLDR dataset

- Tasks:
  - TLDR, helpfulness and harmfulness
- Baselines:
  - DPO, SLIC, IPO in offline, online and on-policy settings
- Base model: Gemma-2B
- Preference simulator (Oracle): RM-deberta (In practice, it should be human)

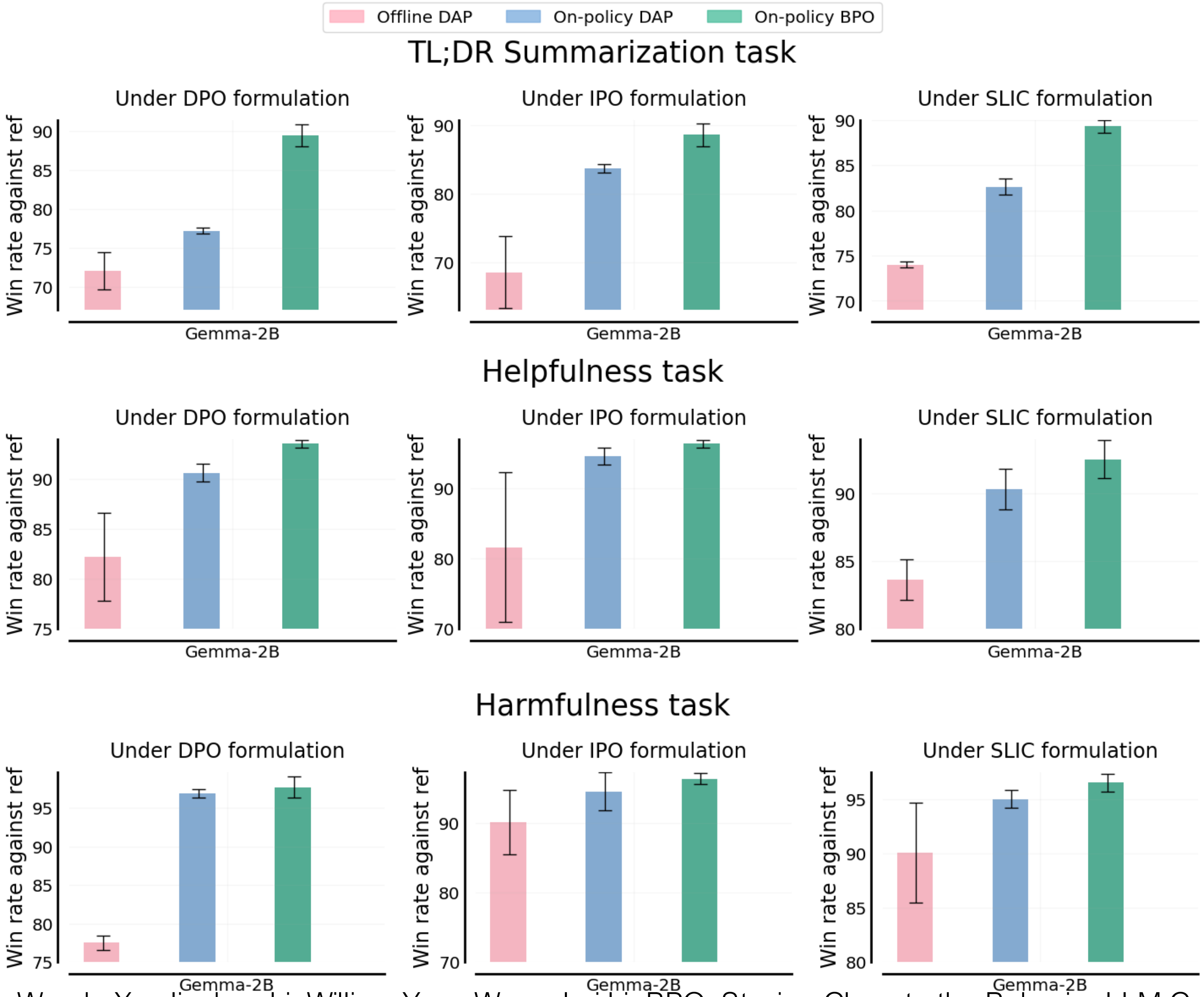
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO outperforms online and offline alignment methods



Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO outperforms baselines across three tasks



Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO Highlight

- Reference model should stay close to the behavior LLM and create better online LLM alignment

- Practical applicability: We empirically show our online BPO with >=2 data collection steps can significantly improve offline baselines

- The effectiveness of BPO stems from proximity to the behavior model, rather than improvements in the reference model's quality.

Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

Outline

- Aligning with online preference optimization (BPO)
- Iterative refinement with fine-grained feedback (LLMRefine)
- Learning Optimized Sample Compute Allocation (OSCA)

Can we use fine-grained feedback to guide LLM?

Input: Translate " 新冠疫情危机爆发 " into English.

LLM's output:

the outbreak of the new crown crisis

What feedback can we give to LLM?

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM's output:

the outbreak of the new crown crisis

Ask LLM to improve?

Source:新冠疫情危机爆发

Translation: the outbreak of the new crown crisis

Please Improve current translation.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM's output:

the outbreak of the new crown crisis

Use binary feedback to guide LLM?

Source:新冠疫情危机爆发
Translation: the outbreak of the new crown crisis
Your translation contains errors. Please improve current translation.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM's output:
the outbreak of the new crown crisis

Use scalar feedback to guide LLM?

Source:新冠疫情危机爆发
Translation: the outbreak of the new crown crisis
Your translation has score of 70/100. Please improve current translation.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM’s output:

the outbreak of the new crown crisis

Use fine-grained feedback to guide LLM!

Source:新冠疫情危机爆发
Translation: the outbreak of the new crown crisis
" new crown" is a major terminology error. Please improve current translation.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

InstructScore's Fine-grained Explanation

Input: Translate "新冠疫情危机爆发" into English.

Candidate: The outbreak of the new crown crisis

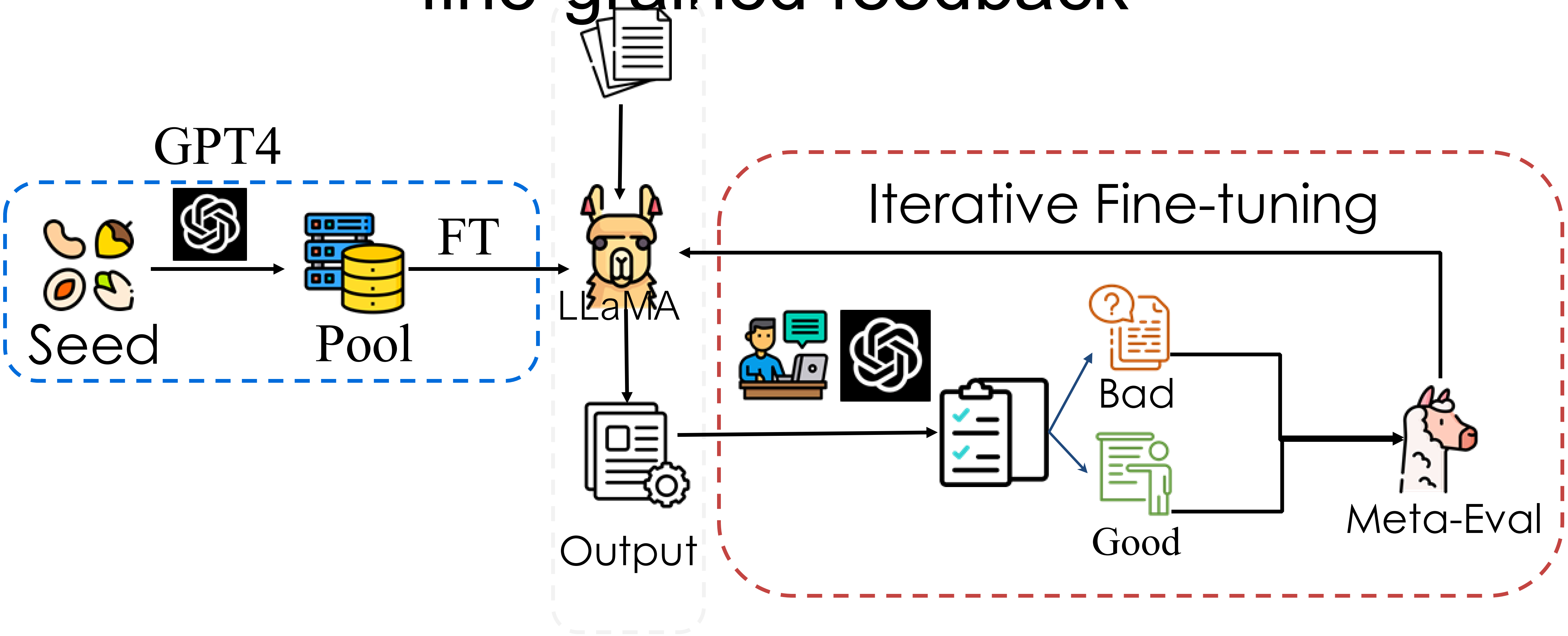
Error location: new crown

Error type: Terminology is used inconsistently

Major/Minor: Major

Explanation: The term " new crown" is not the correct term for "Covid-19".

InstructScore-QE (source-based) to provide fine-grained feedback



Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

Introducing LLMRefine

Source:新冠疫情危机爆发
Translation: the outbreak of the new crown crisis
" new crown" is a major terminology error. Please improve current translation.

LLM's proposal:
the outbreak of the new crisis

Reject

resample from LLM

Accept

Repeat above steps for n iterations

LLM's final output:
the outbreak of the Covid-19 crisis

Source Translation: 新冠疫情危机爆发

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

LLMRefine Algorithm

Repeat n times

Obtain feedback F_i from error pinpoint

Sample revision c_i based on feedback f_i and last generation y_{i-1}

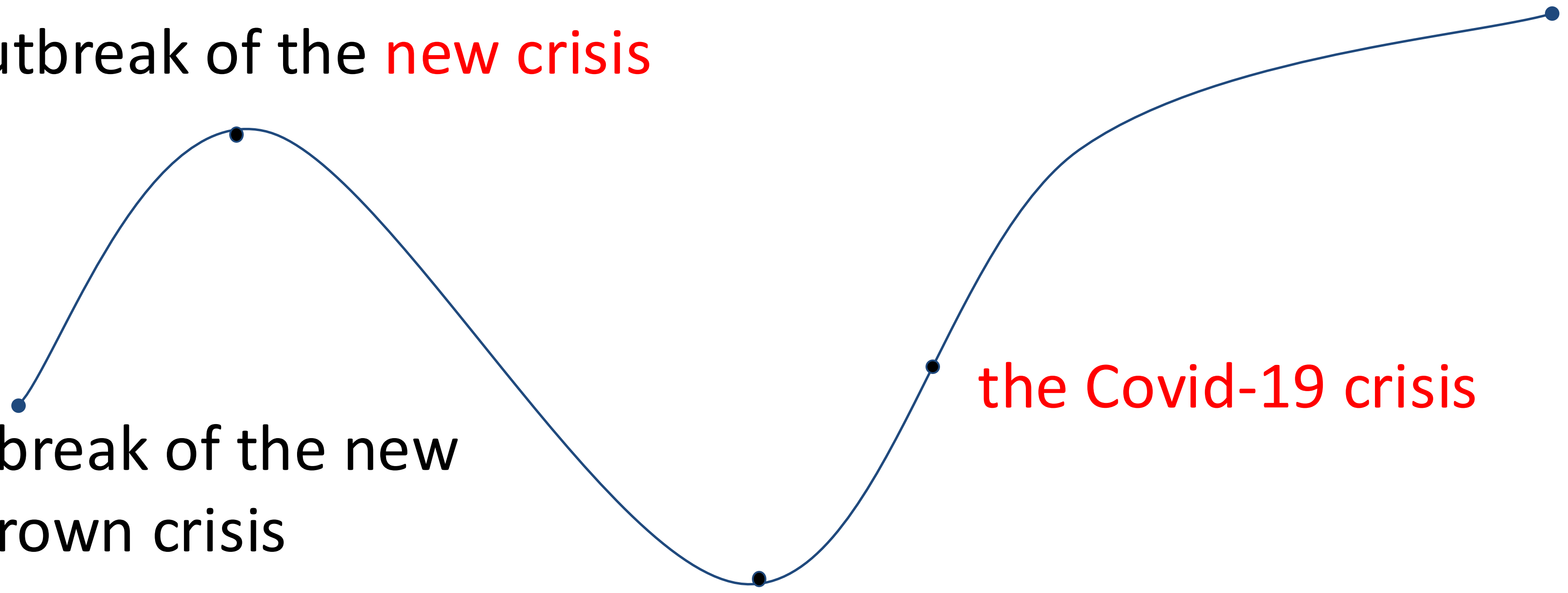
$$P_{accept} = \min(1, e^{\frac{s(F(c_i))-s(F(y\_i))}{n*T_i}})$$

Accept new revision

Keep the last step candidate

$$T_{i+1} = max(T_i - c * T_i, 0)$$

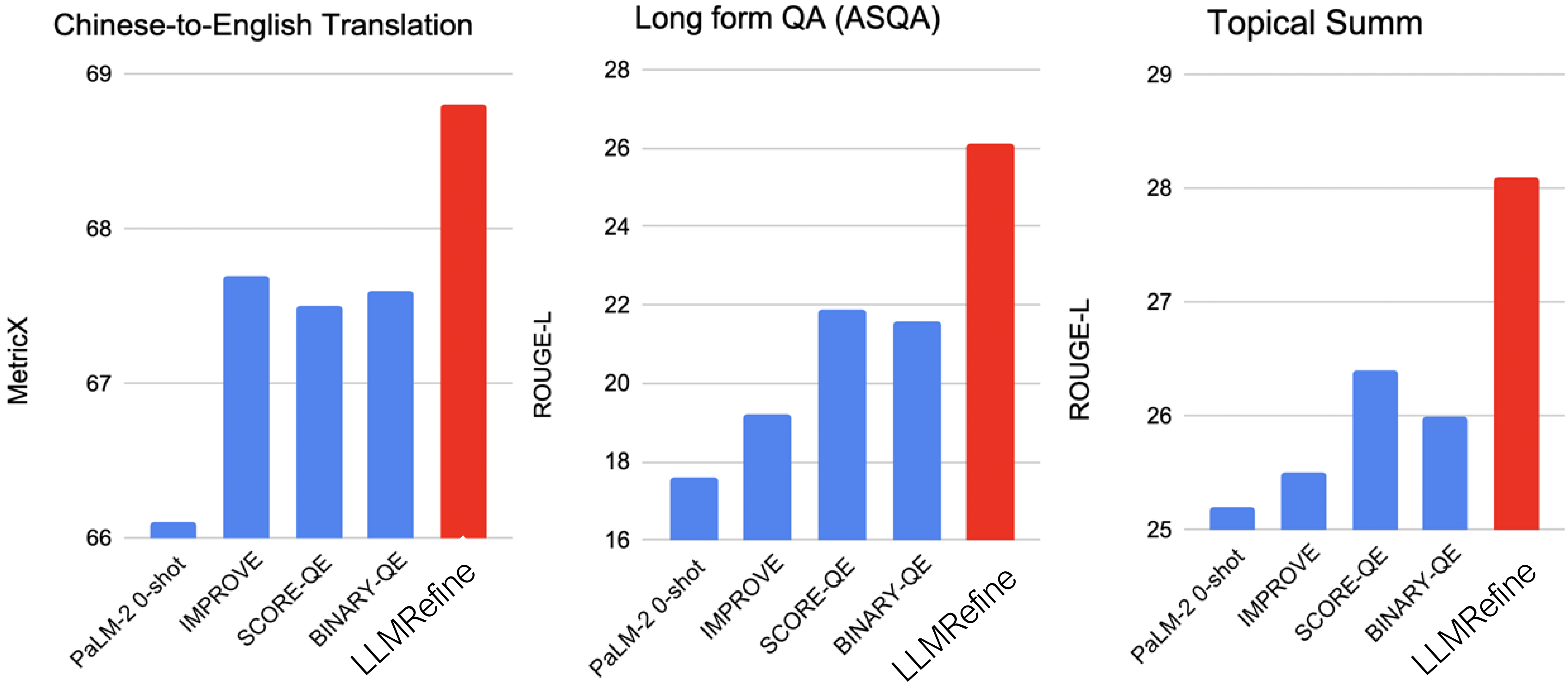
Source Translation: 新冠疫情危机爆发



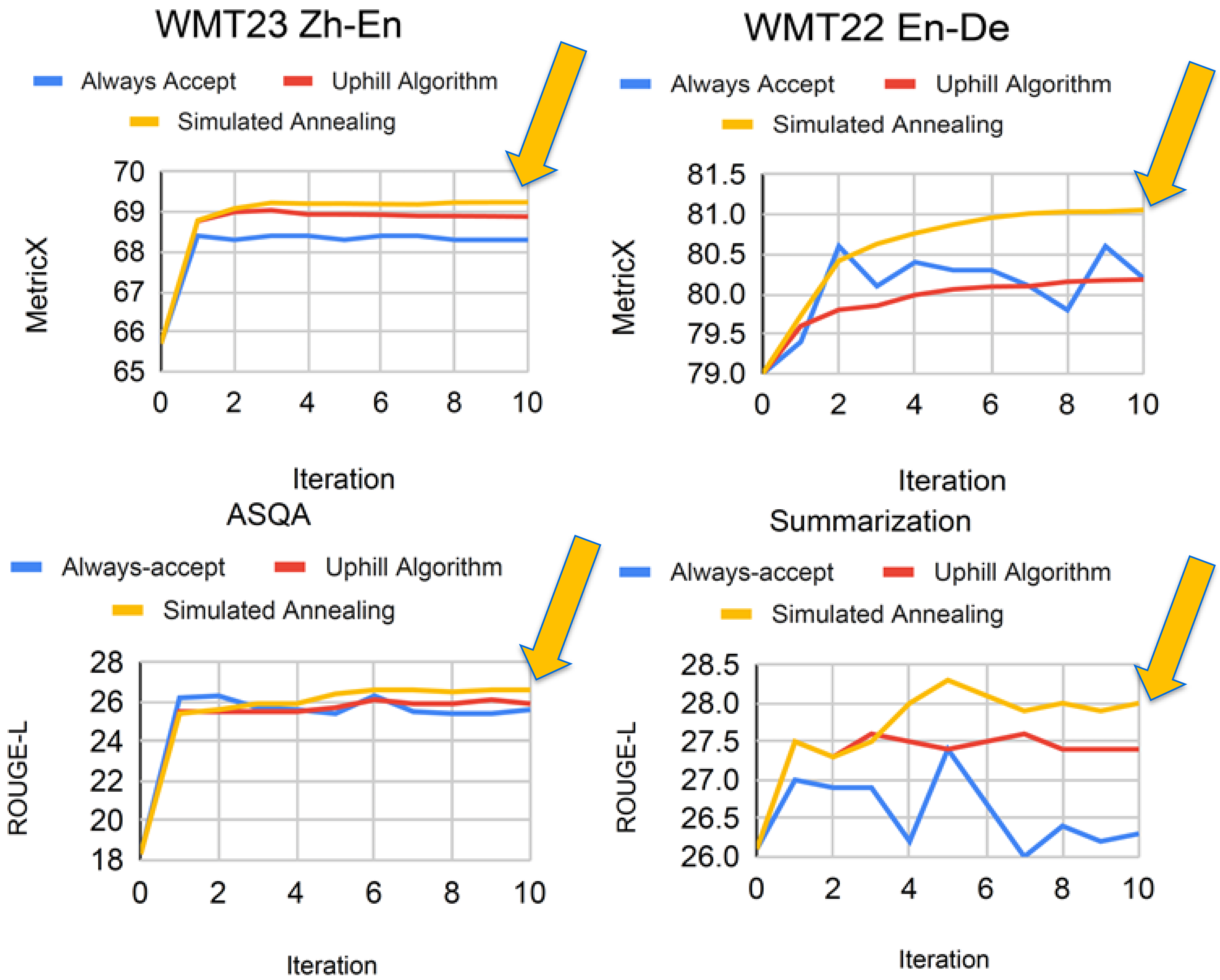
"the new crisis" is a major mistranslation error. The correct translation should be: " the Covid-19 crisis"

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

LLMRefine results in better translations than coarse feedback



Simulated Annealing in LLMRefine



Translation
Summarization
Long form QA

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL24

Key insights of LLMRefine

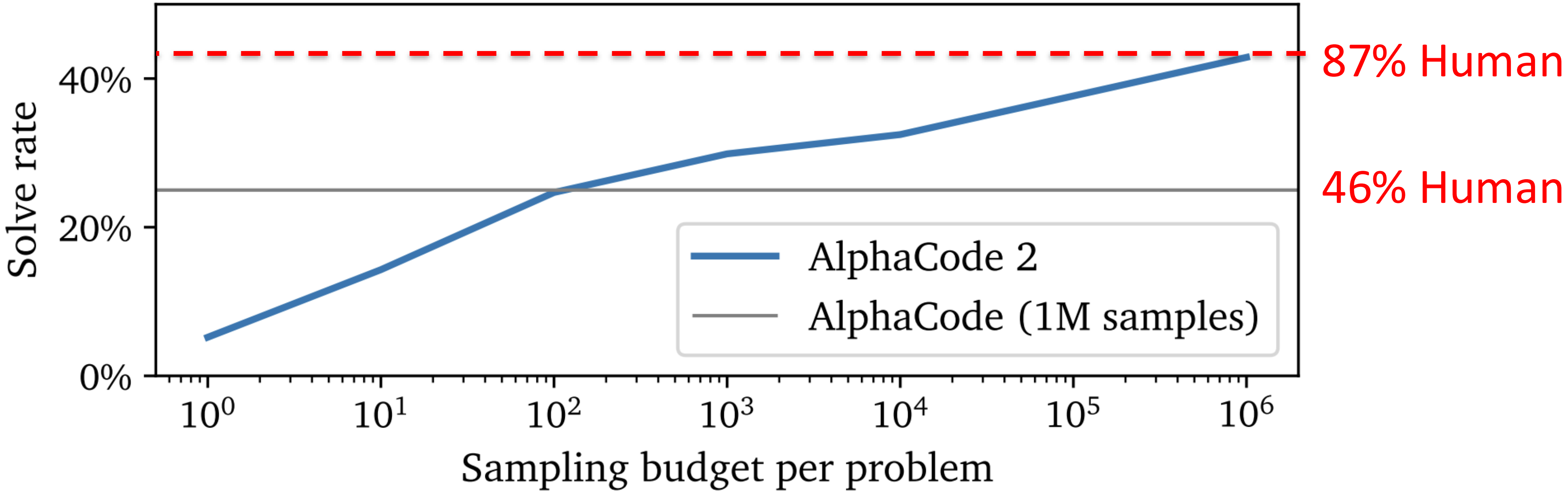
- Binary feedback is not enough
- Fine-grained feedback is better
- Algorithmic iterative refinement is superb

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL24

Outline

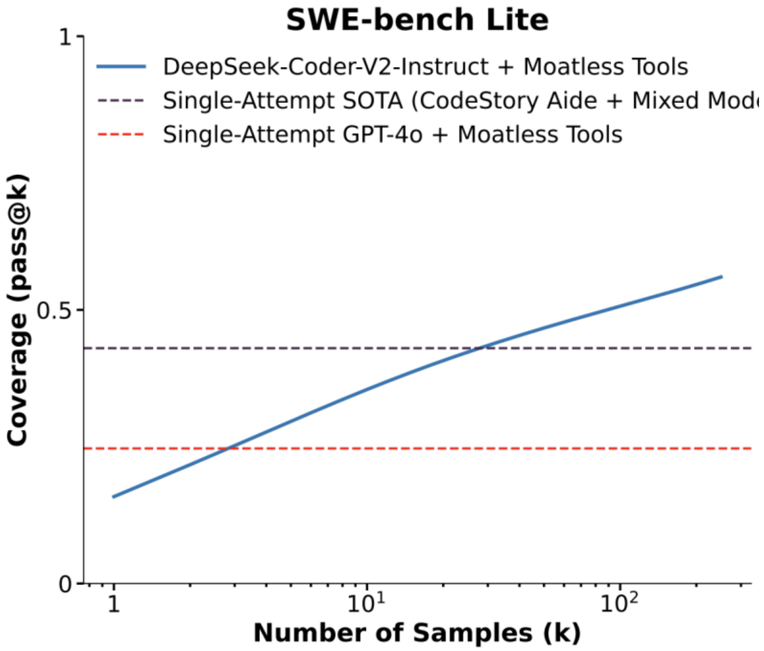
- Aligning with online preference optimization (BPO)
- Iterative refinement with fine-grained feedback (LLMRefine)
- Learning Optimized Sample Compute Allocation (OSCA)

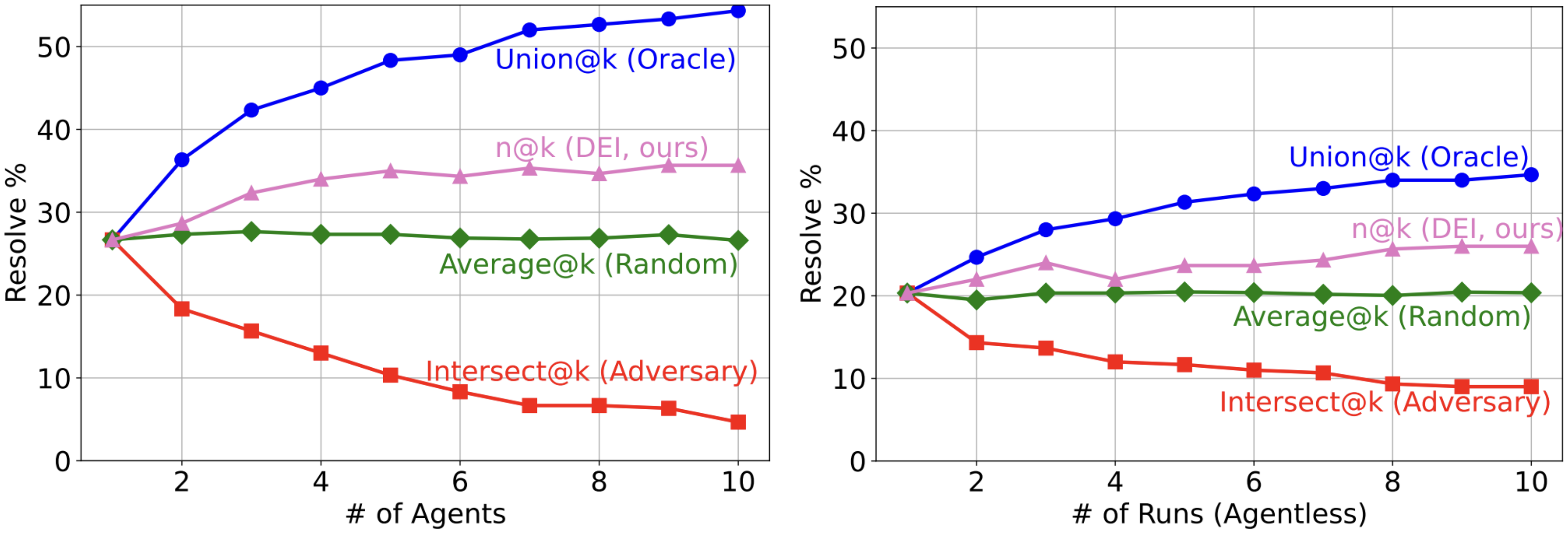
Inference-Time Scaling Law



AlphaCode2 Tech Report, https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2_Tech_Report.pdf

Inference-Time Scaling Law



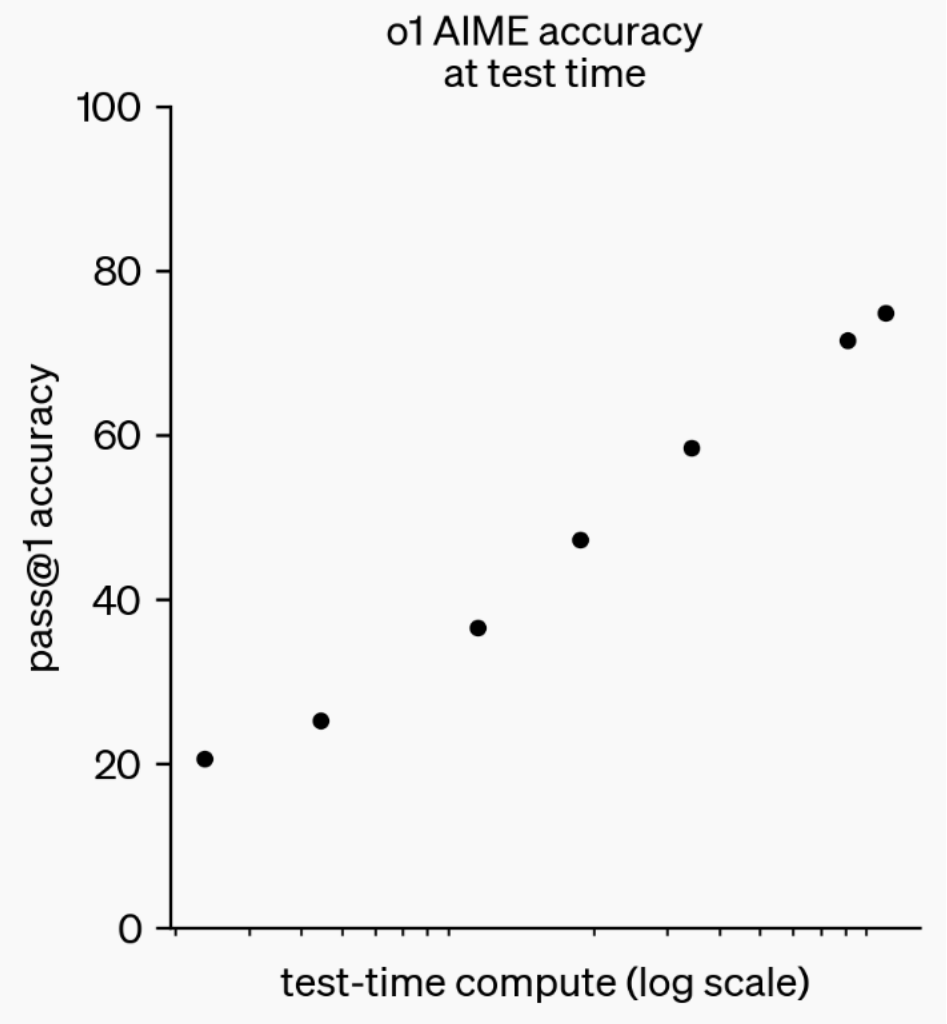


same for Agentic Tasks like SWE-Bench.
More agents, more runs ➔ Better solve rate.

Large Language Monkeys: Scaling Inference Compute with Repeated Sampling, https://arxiv.org/abs/2407.21787
Diversity Empowers Intelligence: Integrating Expertise of Software Engineering Agents, https://arxiv.org/abs/2408.07060

Inference-Time Scaling Law

- Solve rates scale log-linearly with longer CoT.



Learning to Reason with LLMs, https://openai.com/index/learning-to-reason-with-llms/

Sample Compute Allocation Problem for LLM Inference

- Allocate the total amount of compute (# samples, # tokens, FLOPs) C.

- Sampling configurations (i.e. inference hyperparameters):
  - Model to use: gpt-4o, gemini, deepseek, qwen, …
  - Temperature:
  - Output language: python / C++ / Chinese / English

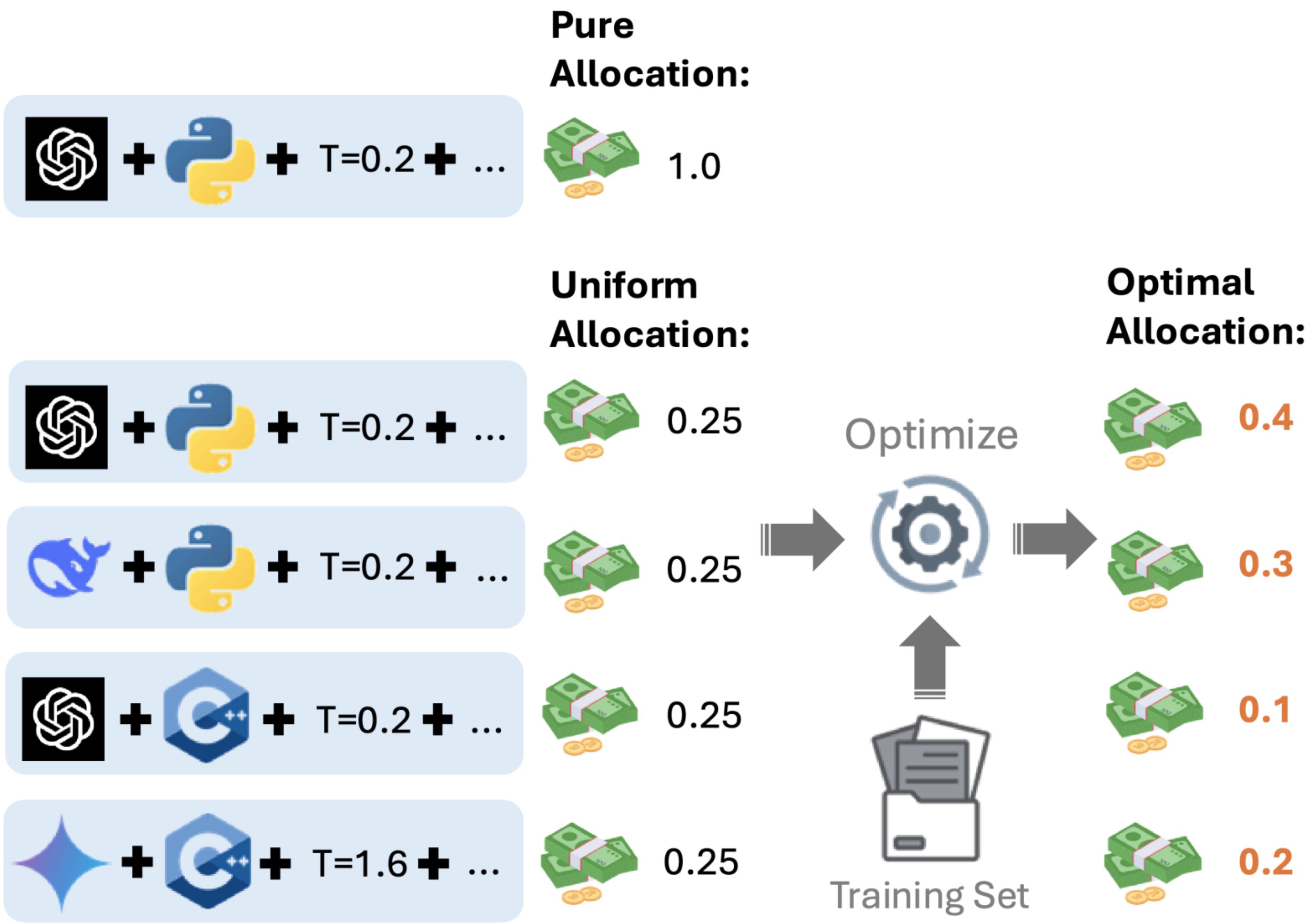
USE CASE	TEMPERATURE
Coding / Math	0.0
Data Cleaning / Data Analysis	1.0
General Conversation	1.3
Translation	1.3
Creative Writing / Poetry	1.5

- Pure Strategy
  - one config uses all compute

- Mixed Strategy

Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

Sample Compute Allocation



Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

Mixed Strategy could be better

- Two problems p1 and p2, two inference settings d1 and d2.
- P(d1 solving p1) = 10%, P(d2 solving p1) = 1%.
- P(d1 solving p2) = 1%, P(d2 solving p2) = 10%.
- Expected number of problems solved given 10 samples:
  - Pure strategy (select either d1 or d2): 37.3%
  - Mixed strategy (5 samples for d1 & d2): 43.8%
- Better to use mixed strategy!

Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

Optimizing Sample Compute Allocation

- The task:
  - Given a set of sampling configurations.
  - Given a training problem set i.i.d. with the test.
  - Given a compute budget C.
  - Find the optimal allocation π that maximizes pass@C.

$$\max_{\pi} \mathbb{E}[\textbf{pass@}C]$$

$$= \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \left(1 - \prod_{i=1}^{|\mathcal{H}|} (1 - p_{ij})^{\pi_i} \right),$$

$$\text{s.t. } 0 \leq \pi_i \leq C,$$

$$\sum_{i=1}^{|\mathcal{H}|} \pi_i = C, \pi_i \in \mathbb{N}.$$

Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

OSCA: Learning to Scale Inference Optimally

- If we ignore the integral constraints, this is a convex problem.
- We run hill climbing algorithm to find the solution.

$$\max_{\pi} \mathbb{E}[\textbf{pass@}C]$$

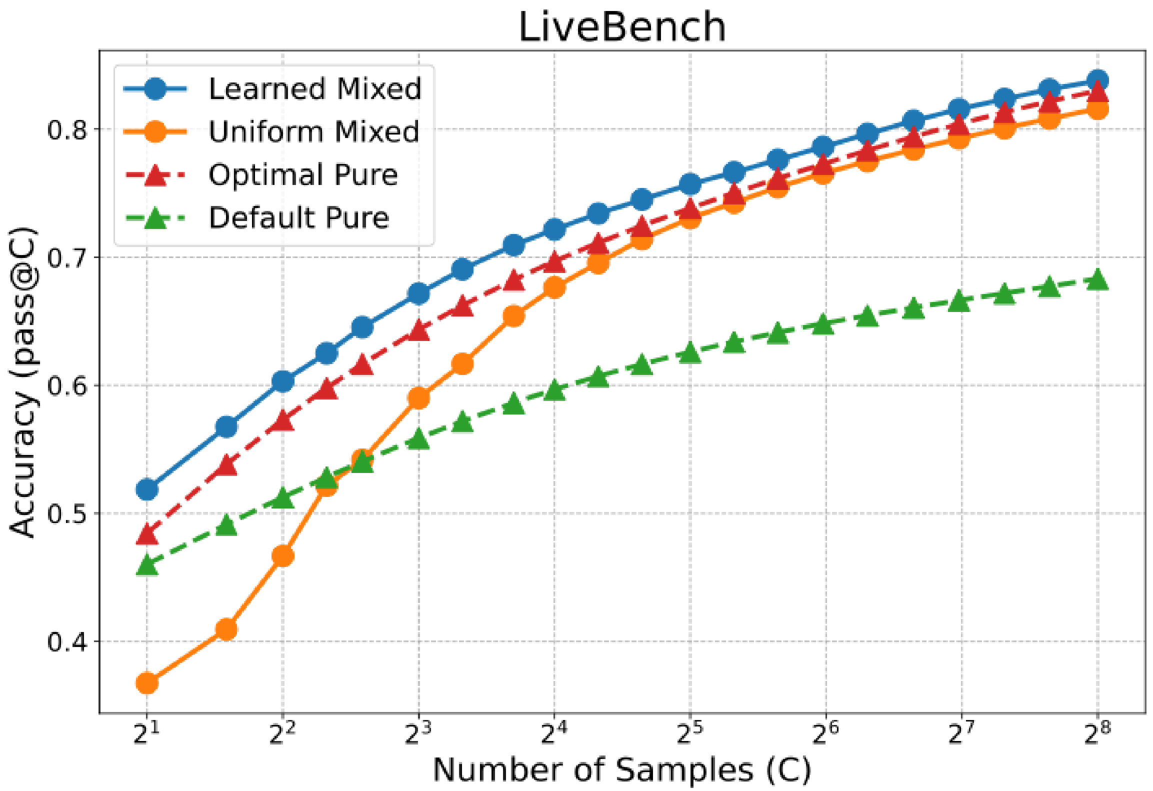
$$= \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \left(1 - \prod_{i=1}^{|\mathcal{H}|} (1 - p_{ij})^{\pi_i} \right),$$

$$\text{s.t. } 0 \leq \pi_i \leq C,$$

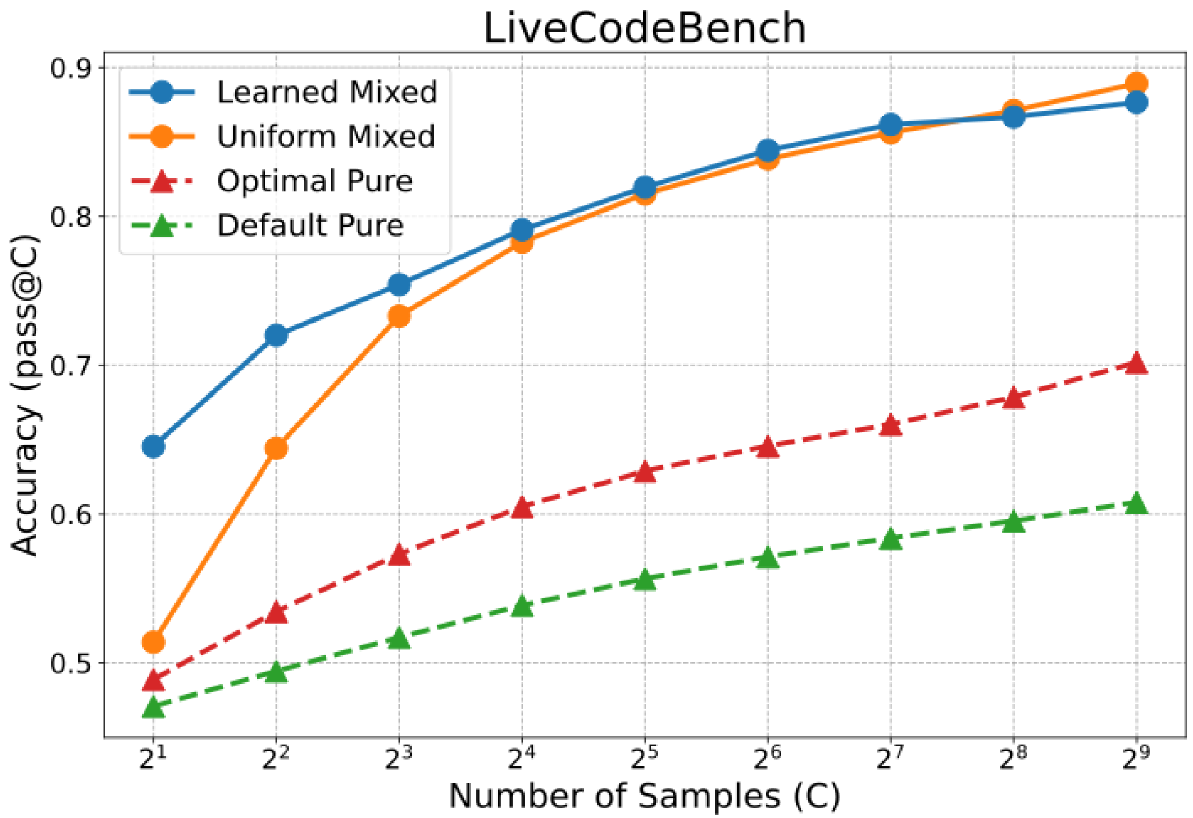
$$\sum_{i=1}^{|\mathcal{H}|} \pi_i = C, \pi_i \in \mathbb{N}.$$

Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

OSCA learned strategies excel!



Qwen2, LLaMA3, Deepseek-70B



GPT4o, Gemini, Deepseek

Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

OSCA Learns to Scale Inference Optimally

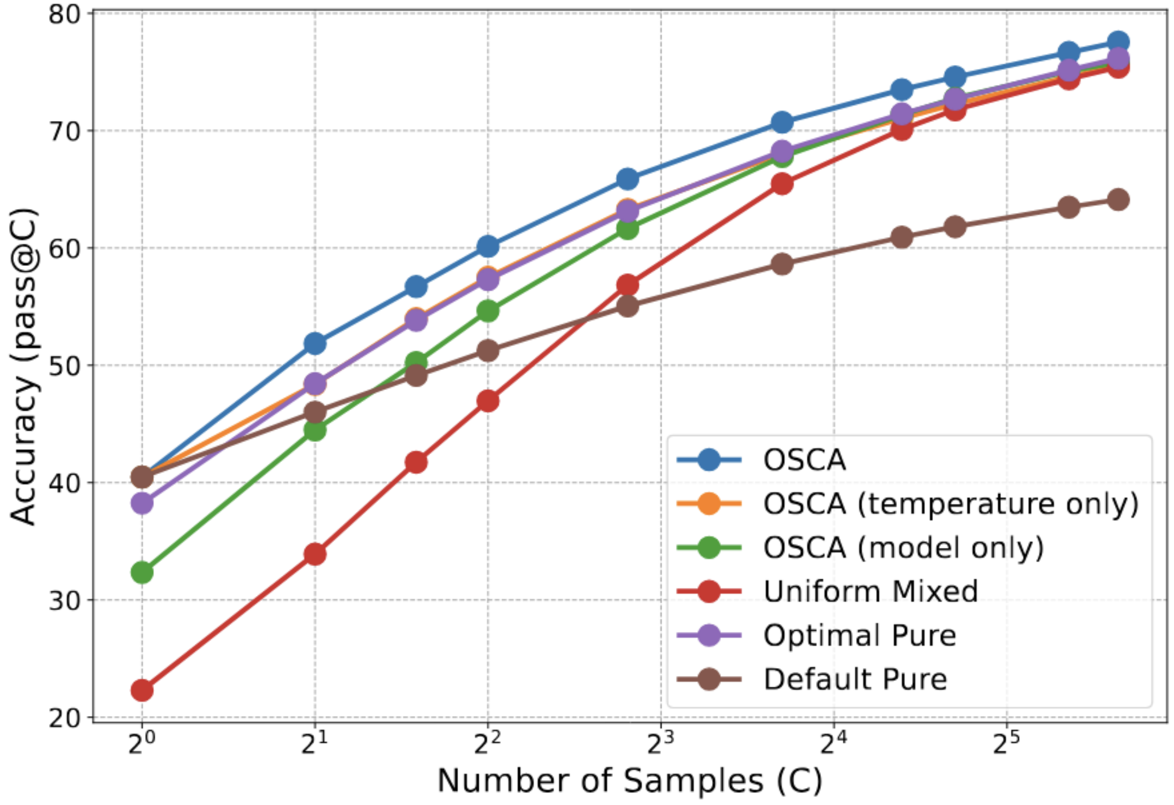


Figure 5: OSCA's pass rates on LiveBench when it is banned from allocating compute to multiple temperatures or multiple models.

Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

Highlight of OSCA

- LLM's problem solve-rate grows log-linearly with # of samples.
- Allocating compute to different inference settings could lead to huge improvement
- Estimating the passing rate for each problem and each configuration
- Hill-climbing to find the optimal allocation

Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.

Summary

- Aligning with online preference optimization (BPO)
    - online and on-policy alignment is better
- Iterative refinement with fine-grained feedback (LLMRefine)
    - simulated annealing with fine-grained feedback improves LLM
- Learning Optimized Sample Compute Allocation (OSCA)
    - sample compute configuration as hyperparameters to optimize

Reference

- Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

- Xu, Deutsch, Finkelstein, Juraska, Zhang, Liu, Wang, Li, Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024.

- Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, Lei Li. Scaling LLM inference with optimized sample compute allocation. NAACL 2025.