Scalable Post-Training Optimization for Large Language Models

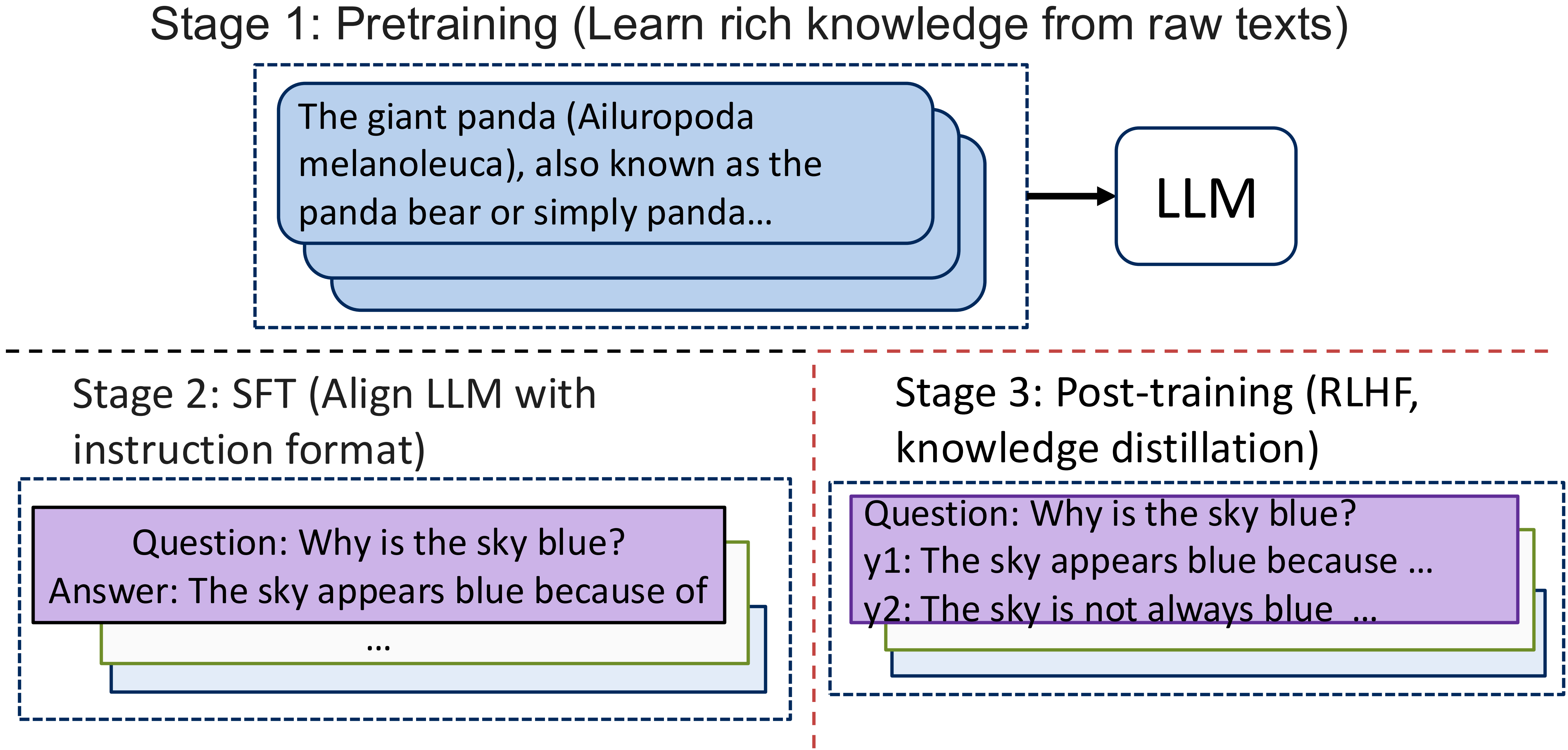
Lei Li

Language Technologies Institute

Carnegie Mellon University

December 11, 2025

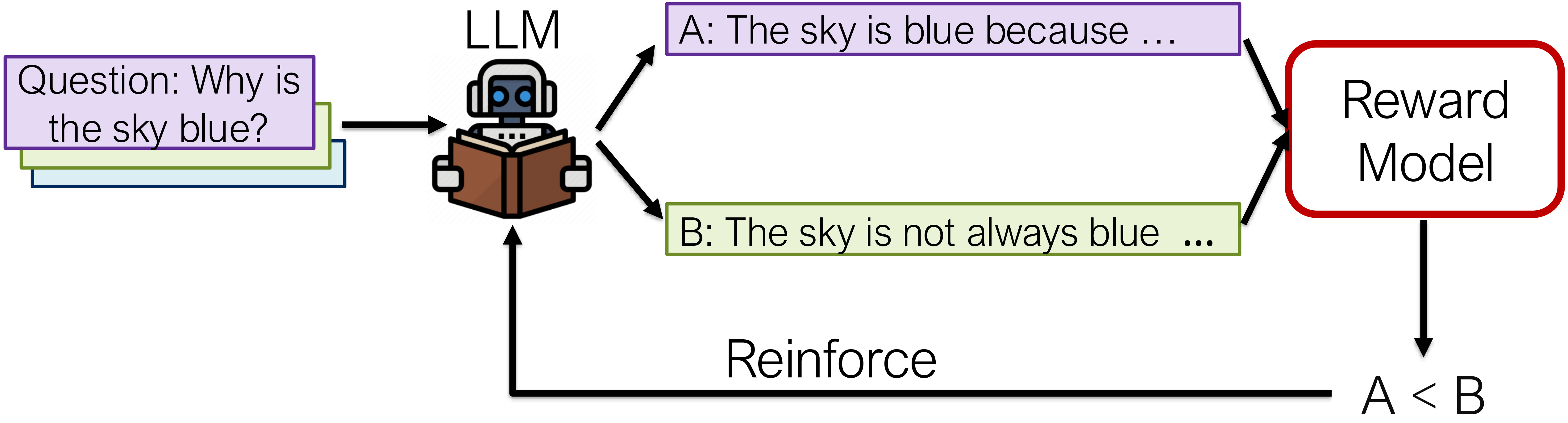
LLM training pipeline



Outline

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

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

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

Constrained optimization

Issues with PPO:

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

Reward modeling in RLHF

$$(x, y_w, y_l) \longrightarrow \boxed{\text{Reward Model}}$$

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$$ Bradley-Terry Model

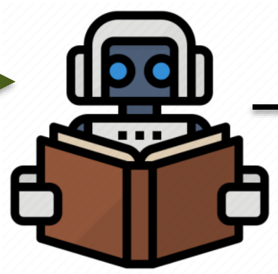
$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\big[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\big]$$

Training language models to follow instructions with human feedback

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

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

Derivation of DPO

$$r(x,y) = \beta \log \frac{\pi_r(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x)$$

Represent reward r using optimal policy π_r

$$p^*(y_1 > y_2 |\, x) = \frac{\exp(r^*(x,y_1))}{\exp\left(r^*(x,y_1)\right)+\exp\left(r^*(x,y_2)\right)} = \frac{1}{1+\frac{\exp\left(r^*(x,y_2)\right)}{\exp\left(r^*(x,y_1)\right)}} =$$

$$\frac{1}{1+\exp\left(\beta \log \frac{\pi_r(y_2|x)}{\pi_{ref}(y_2|x)} - \beta \frac{\pi_r(y_1|x)}{\pi_{ref}(y_1|x)}\right)}$$

Plug reward function into Bradley Terry model

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

Offline DPO variants

All DPO variants follow this

$$r_\phi(y_w) - r_\phi(y_l) = \beta \left(\log \frac{\pi^*_\theta(y_w)}{\pi_{\mathrm{ref}}(y_w)} - \log \frac{\pi^*_\theta(y_l)}{\pi_{\mathrm{ref}}(y_l)} \right).$$

DPO loss:

$$-\log \sigma \left(\beta \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^-|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^-|\boldsymbol{x})} \right)$$

IPO loss:

$$\left(\log \left(\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^-|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^-|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^+|\boldsymbol{x})} \right) - \frac{1}{2\beta} \right)^2$$

Avoids the overfitting from DPO (Squared loss)

SLiC loss:

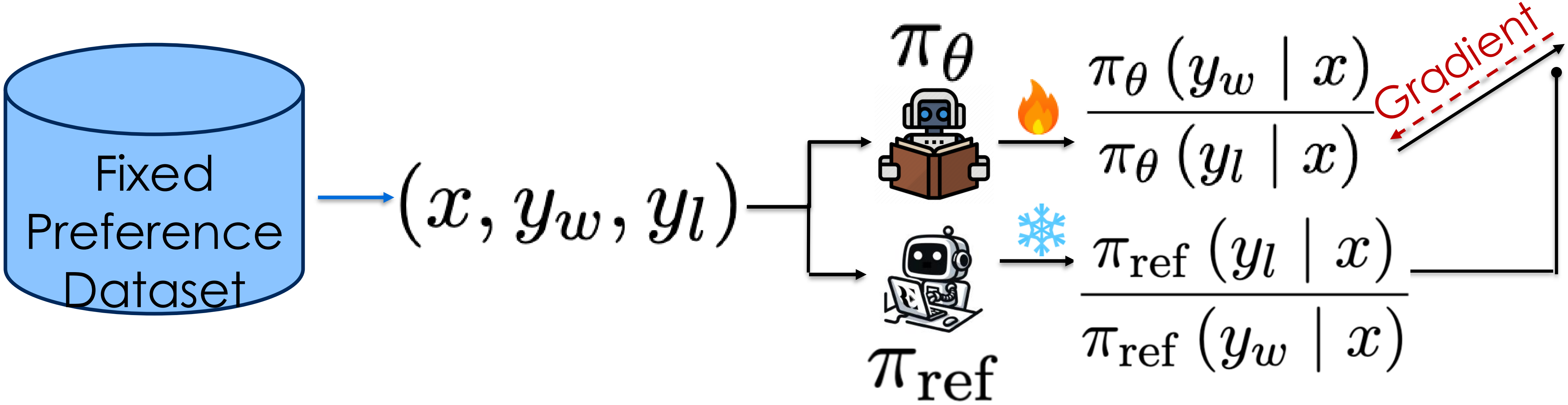
$$\max \left(0, 1 - \beta \log \left(\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^-|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^-|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^+|\boldsymbol{x})} \right) \right)$$

Hinge loss

Generalized Preference Optimization: A Unified Approach to Offline Alignment

Illustration of DPO

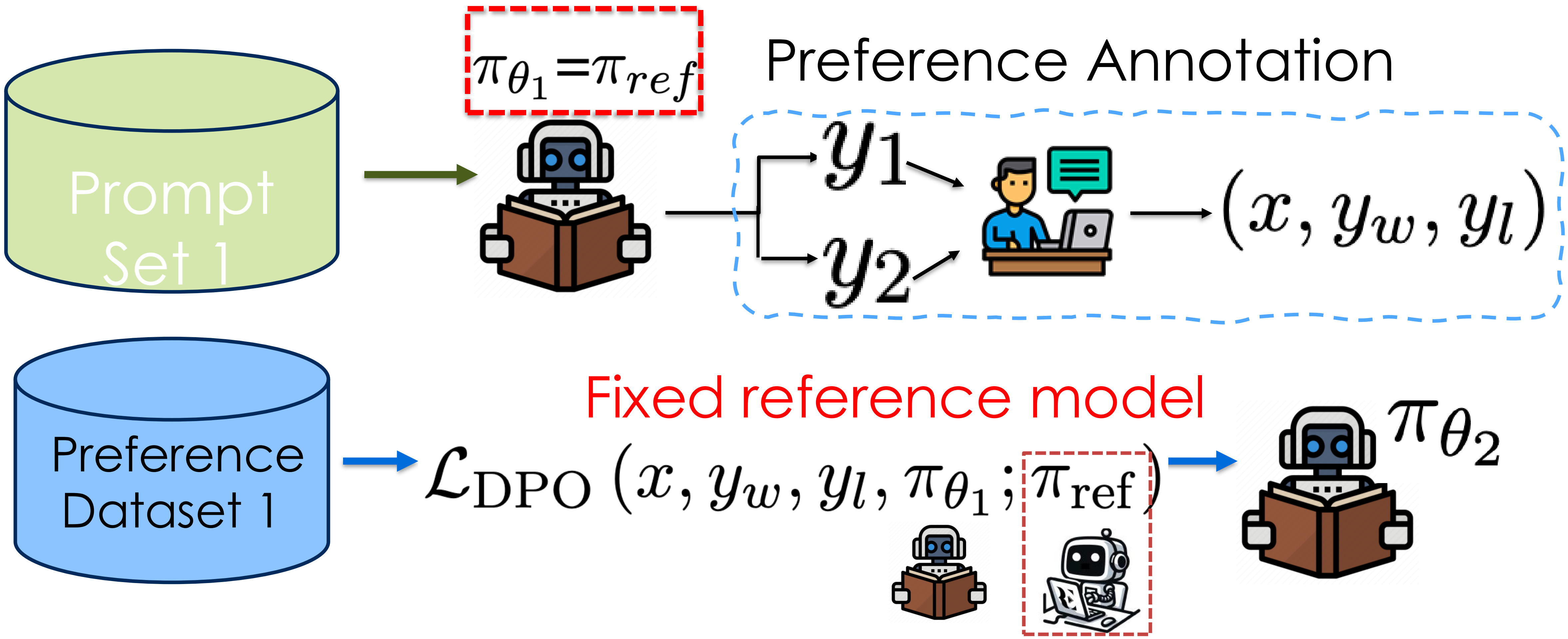
$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$



Fixed reference model

Limitation of offline DPO (and online DPO)

Synthetic data distribution shifts

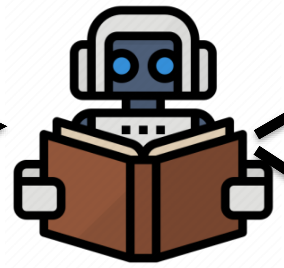


Data distribution shift during training

π_{ref}

Iter 1

Prompt : Translate this Assamese sentence into English…

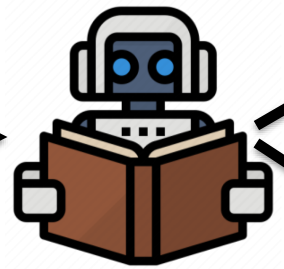


I don't know this language

In the past, in the past, in the past

Iter 2

Prompt : Translate this Assamese sentence into English…



It was to last for the next 40 years …

This has been going on for 40 years …

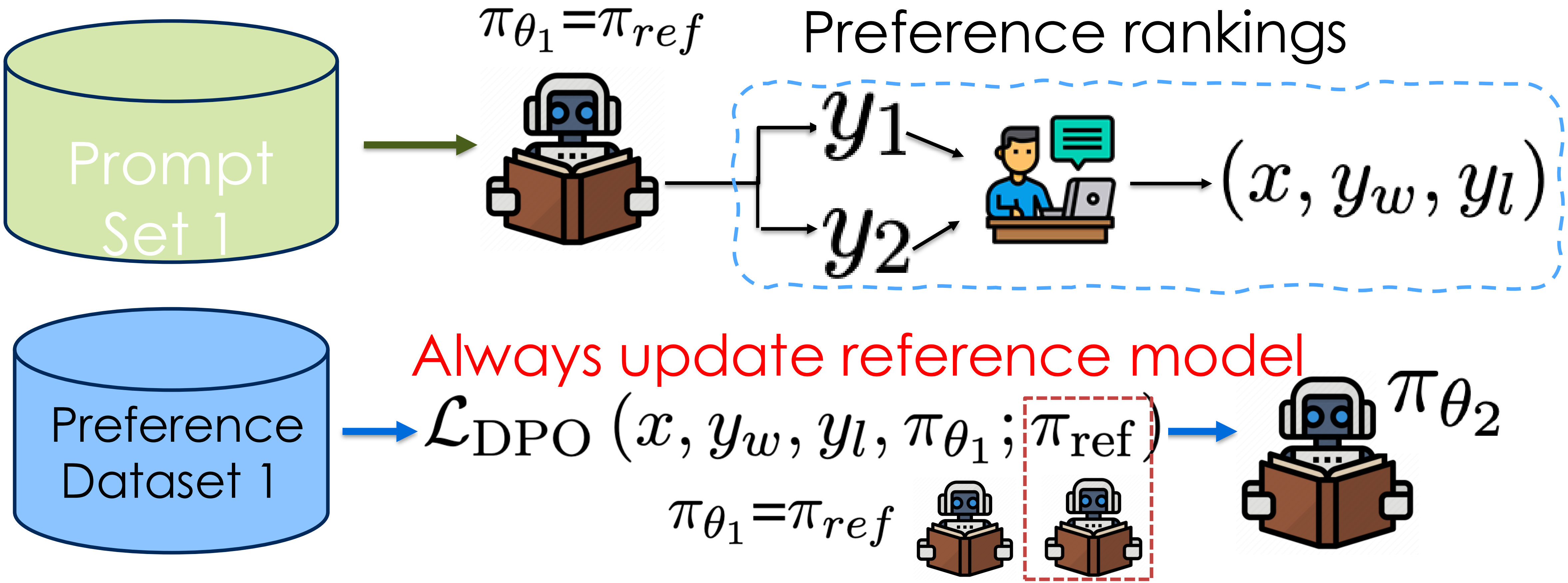
?

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

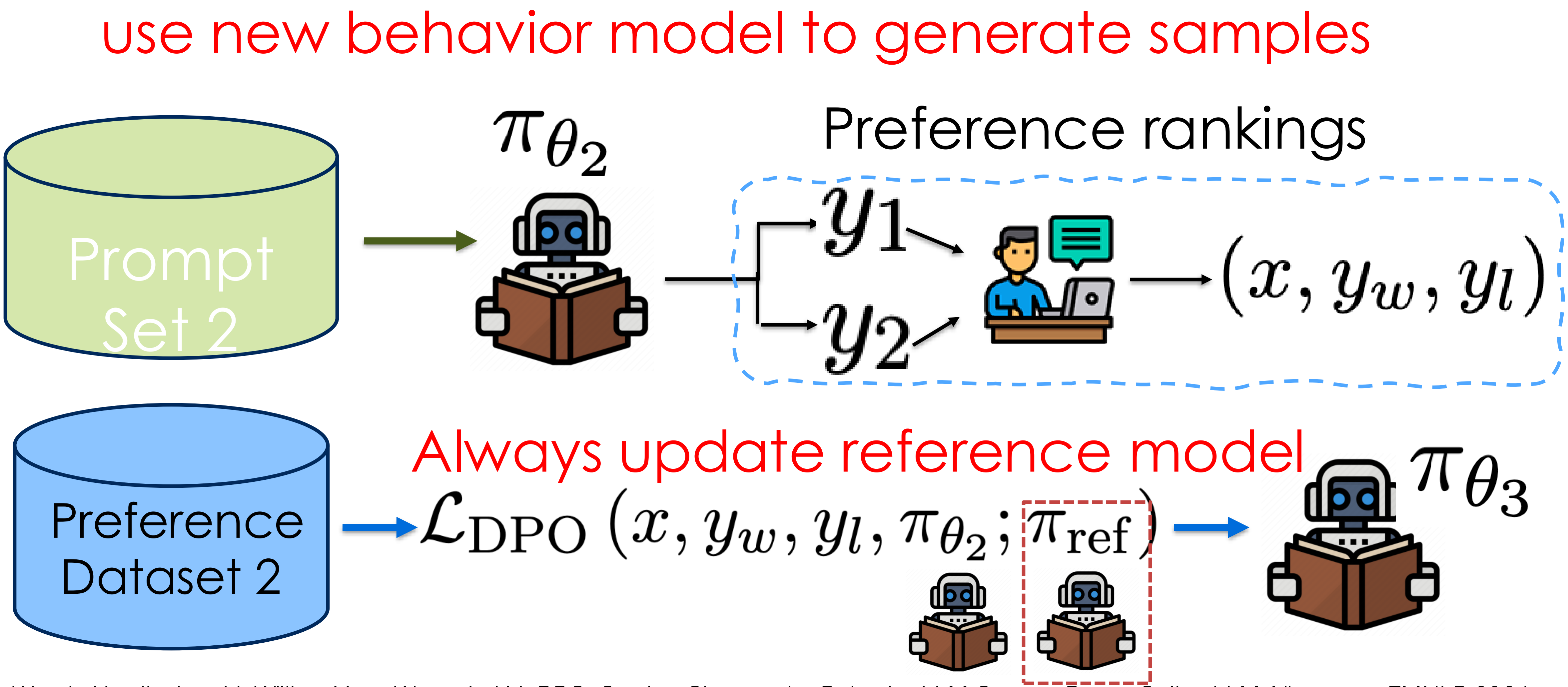
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO



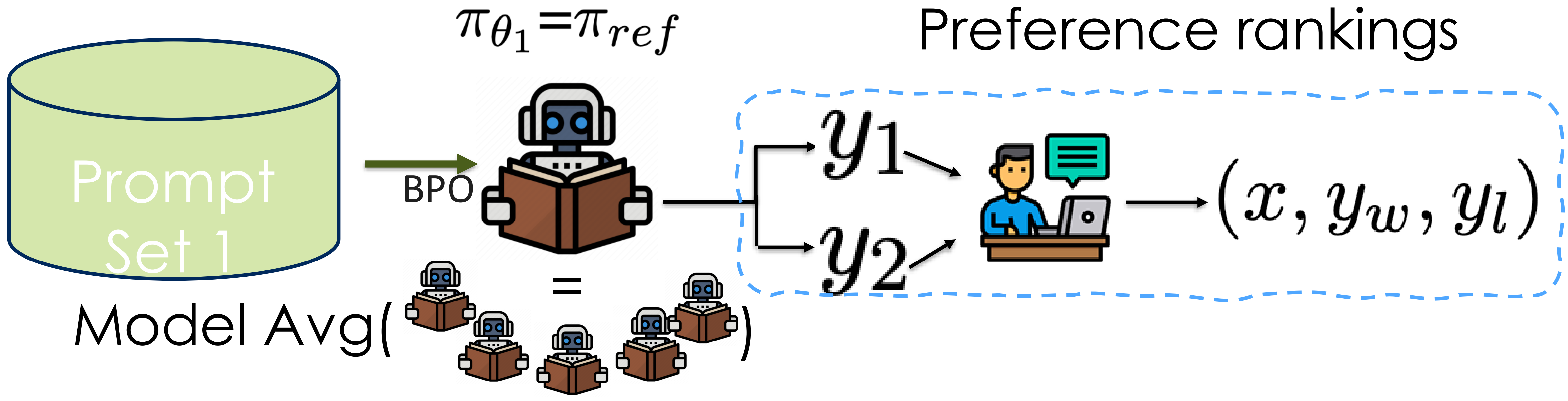
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO



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

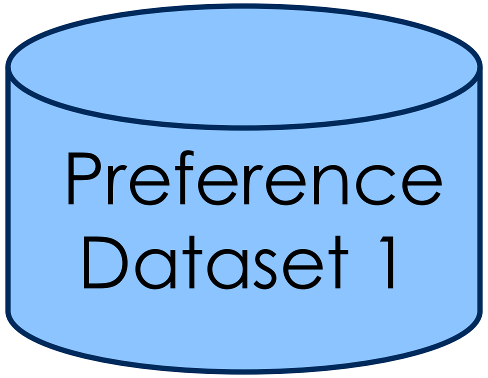
Practical implementation of BPO (Lora ensemble)



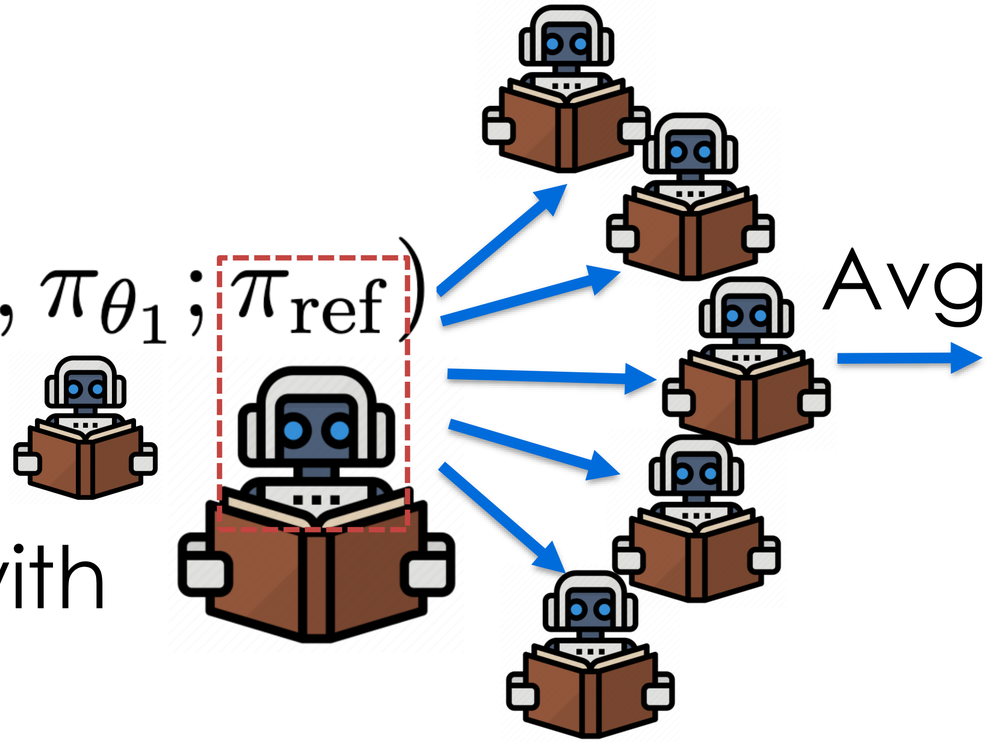
We use model averaged LoRA weights to perform sampling

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)



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



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.

Why ref=behavior may not work?

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

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$
$$-\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \bigg] \bigg]$$

$$\text{where } \hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

This term degenerates to constant 0.5

When ref=behavior

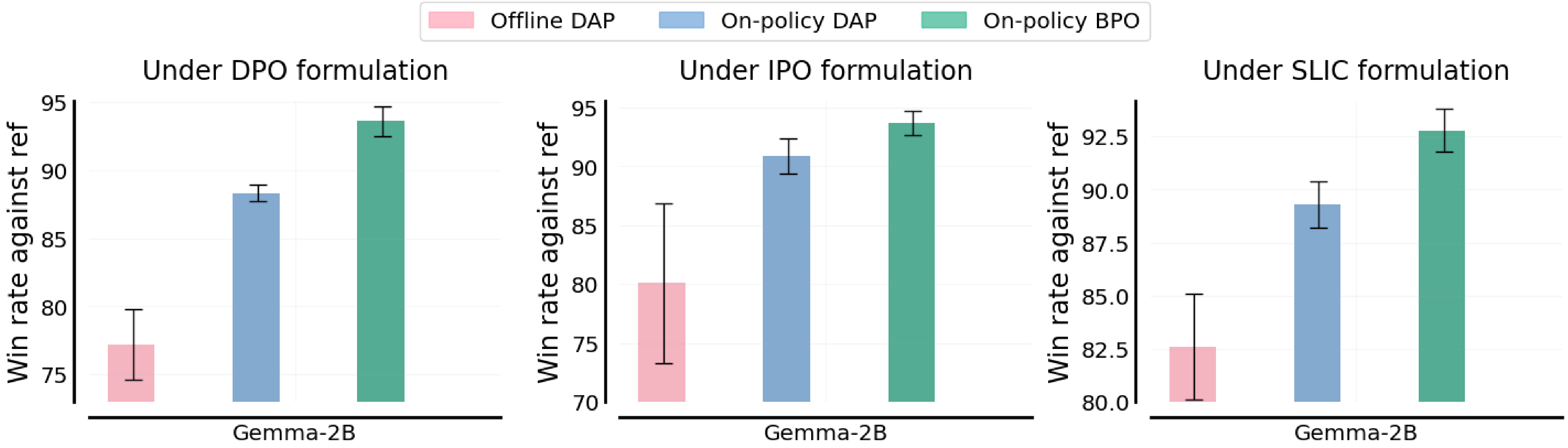
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

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)

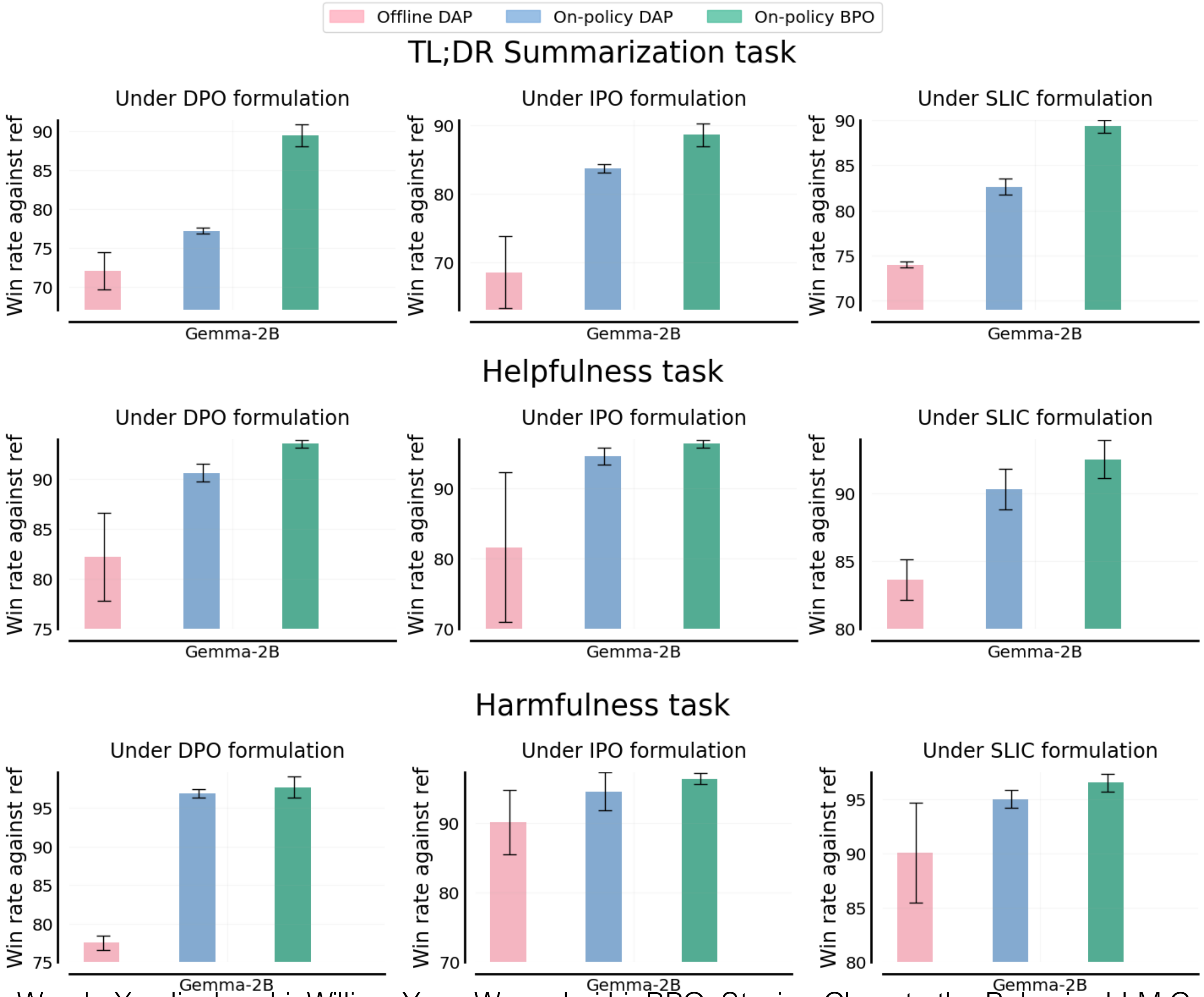
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO outperforms online and offline alignment methods



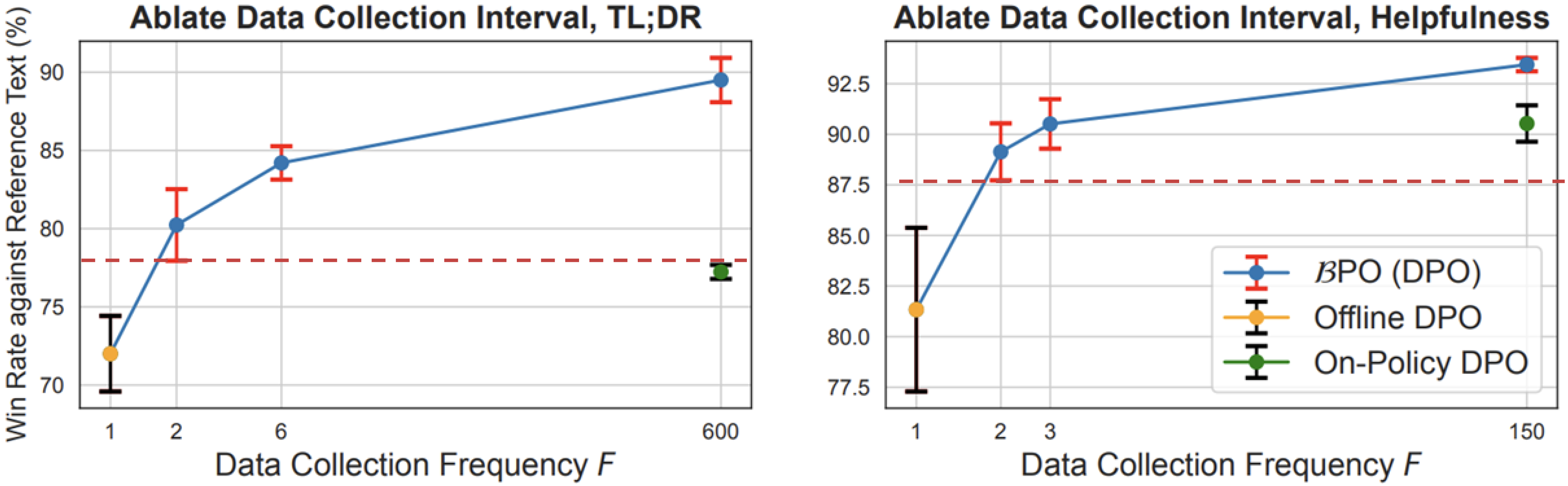
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

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 adapts to different data collection frequencies



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)
- Speculative Knowledge Distillation

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

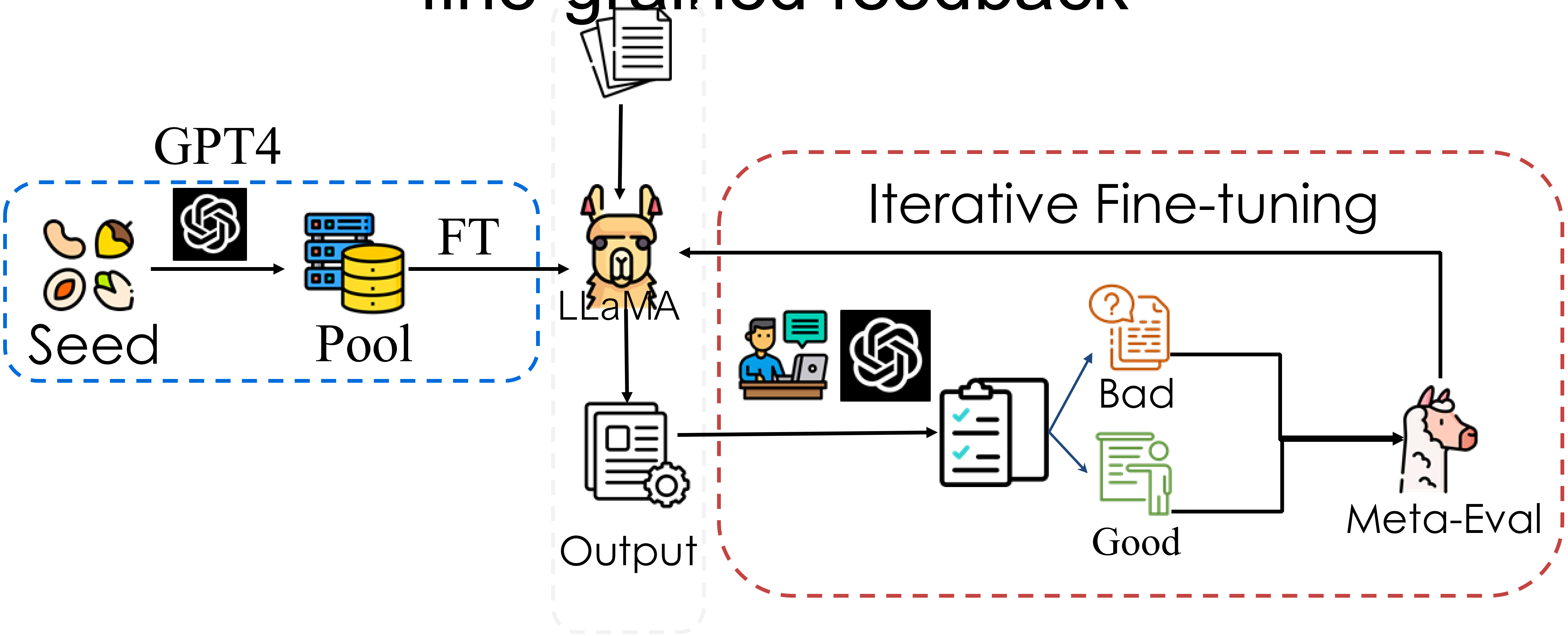
Error type: Terminology is used inconsistently

Major/Minor: Major

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

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

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

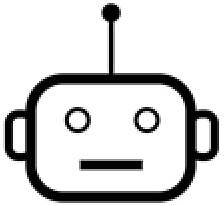
Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

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}



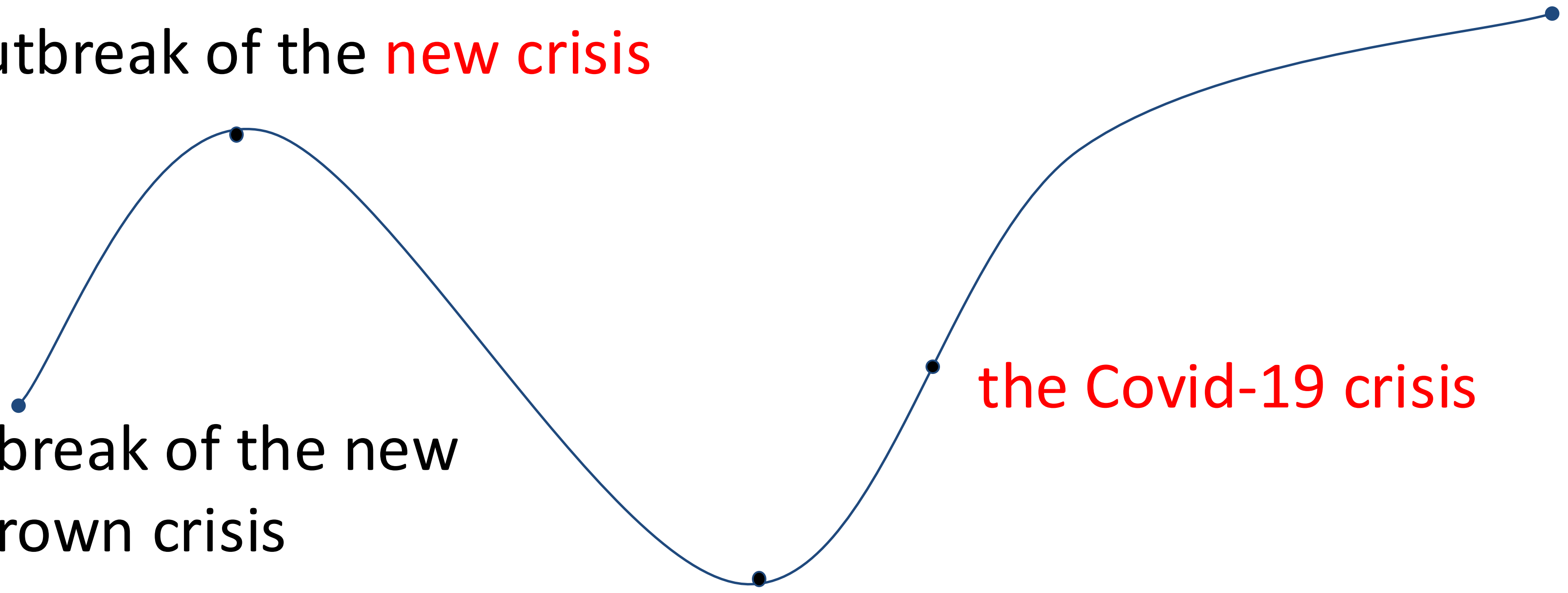
$$P_{accept} = \min\left(1, e^{\frac{s(F(c_i)) - s(F(y\_i))}{n * T_i}}\right)$$

Accept new revision

Keep the last step candidate

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

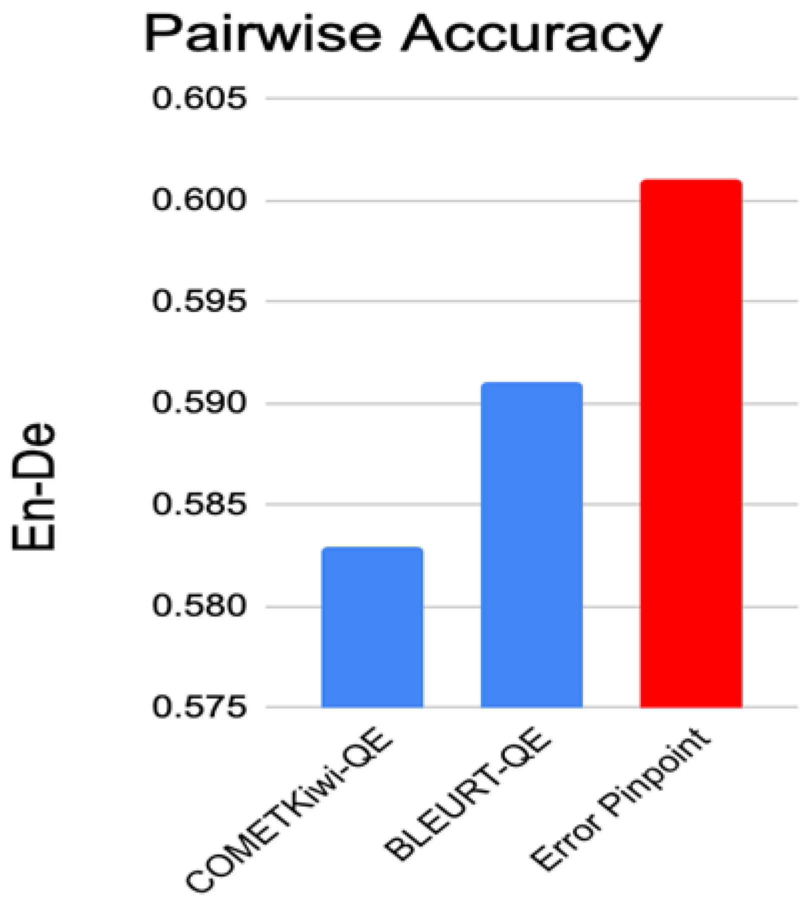
Source Translation: 新冠疫情危机爆发

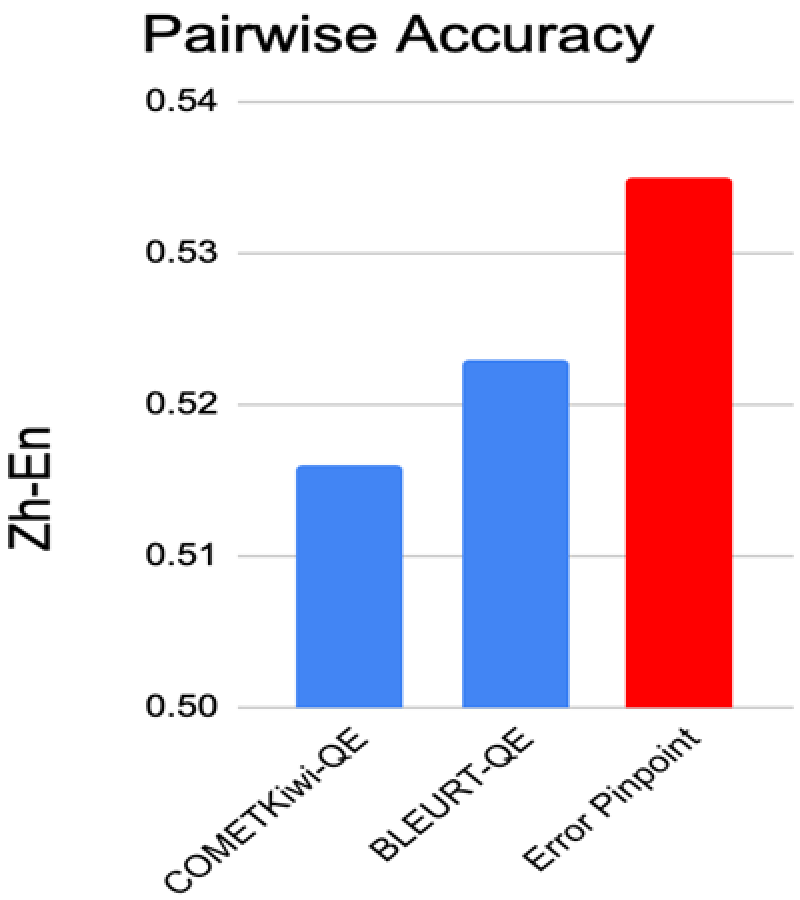


"the new crisis" is a major mistranslation error. The correct translation should be: " the Covid-19 crisis"

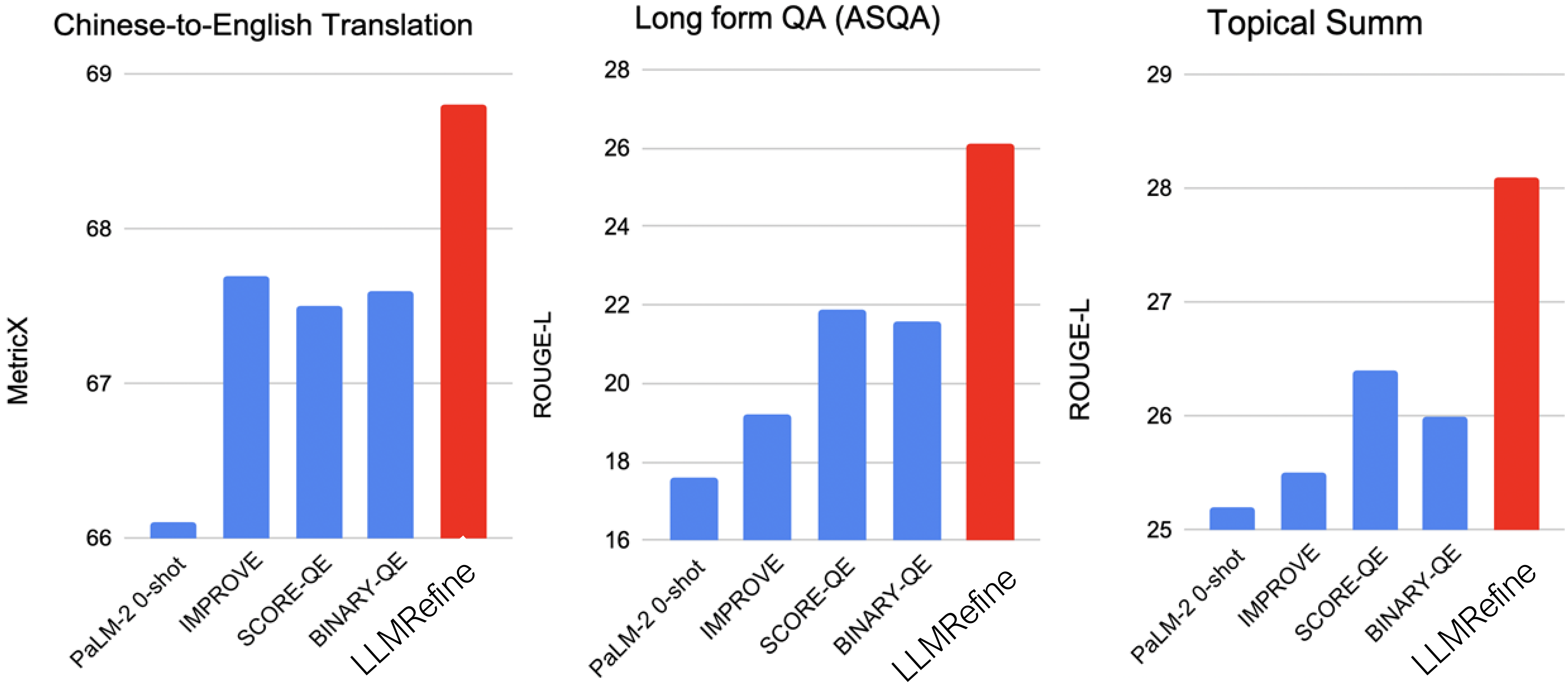
Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

RQ1: How well does our error pinpoint model align with human annotations of generation quality?

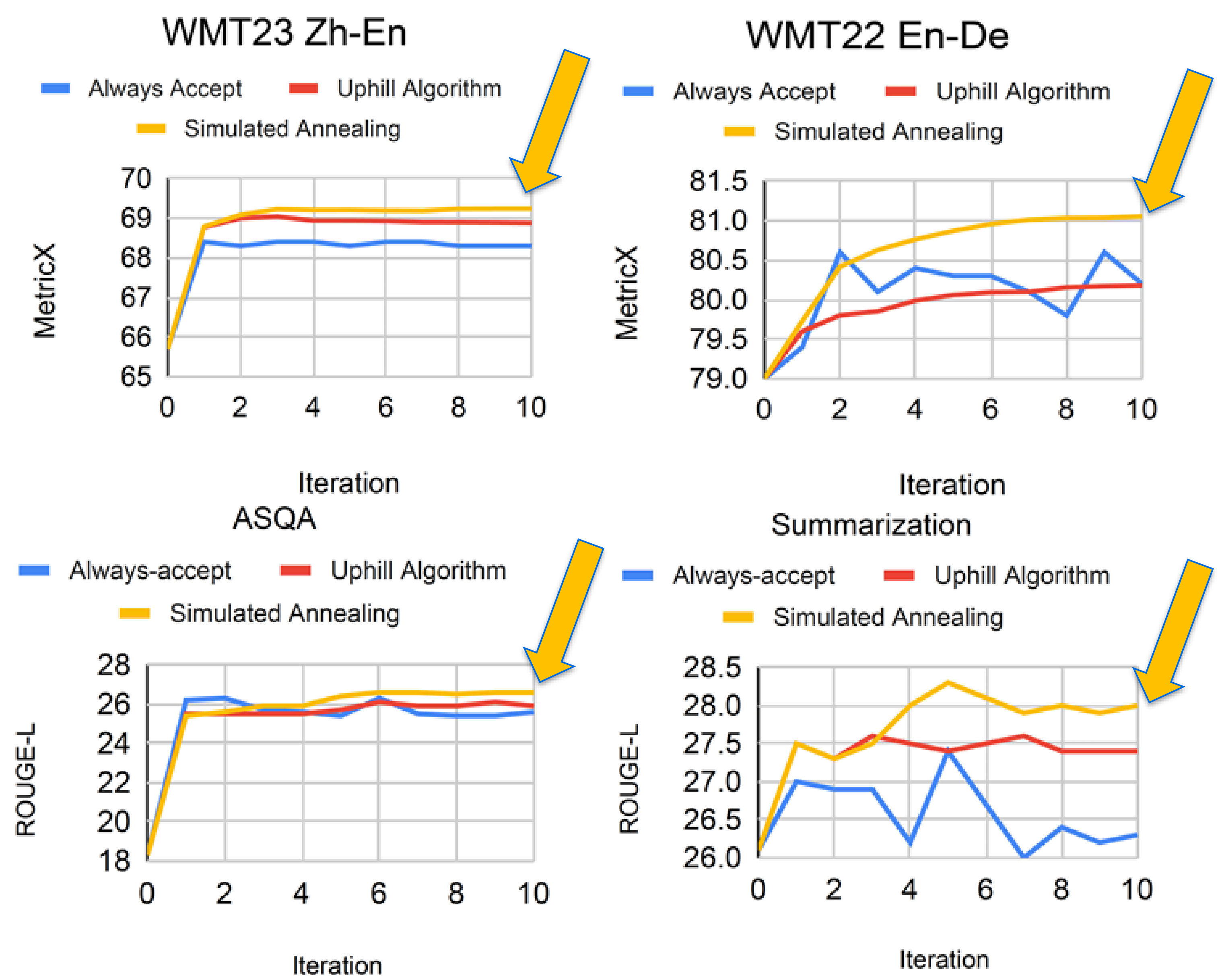




LLMRefine results in better translations than coarse feedback



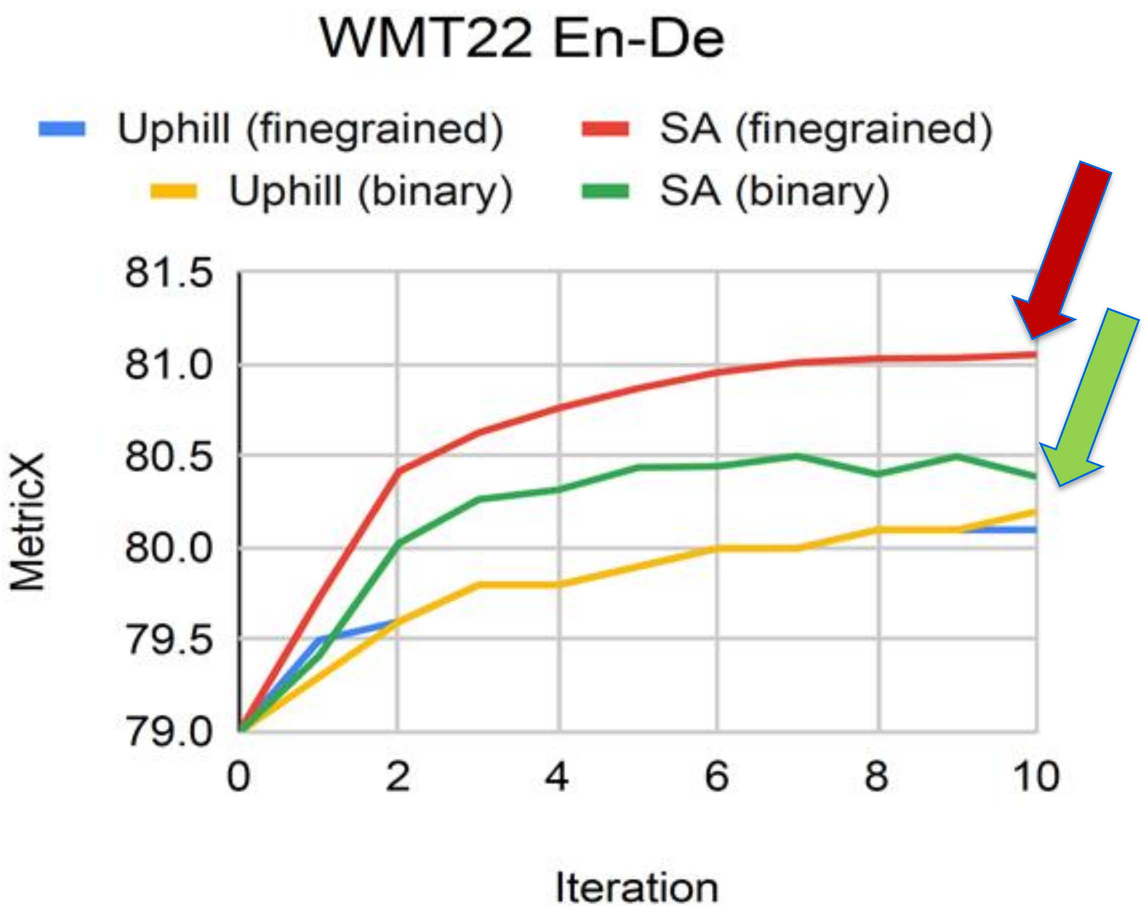
Simulated Annealing works in LLMRefine

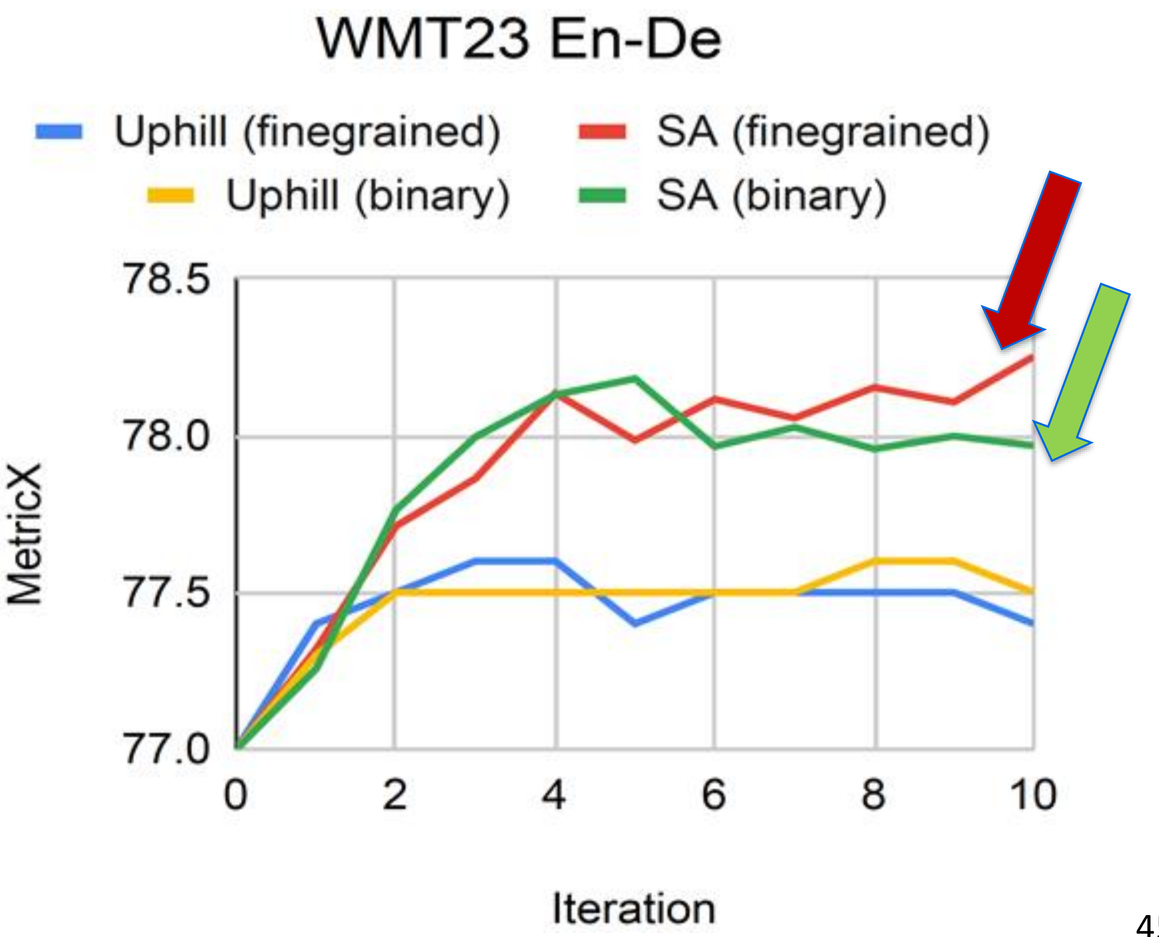


Translation
Summarization
Long form QA

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

Simulated annealing can boost performance of both coarse and fine-grained feedback





Human Evaluation further validates our results

Our fine-grained has all win/lose ratios greater than 1

WMT22 En-De	Win/lose ratio
0-shot	2.34
Improve	2.44
BLEURT-Score-QE	2.79
BLEURT-Binary-QE	1.76
Score-QE	1.23
Binary-QE	1.84

Our SA has all win/lose ratios greater than 1

WMT22 En-De	Win/lose ratio
Always-Accept	1.56
Greedy Uphill	1.38

Key insights of LLMRefine

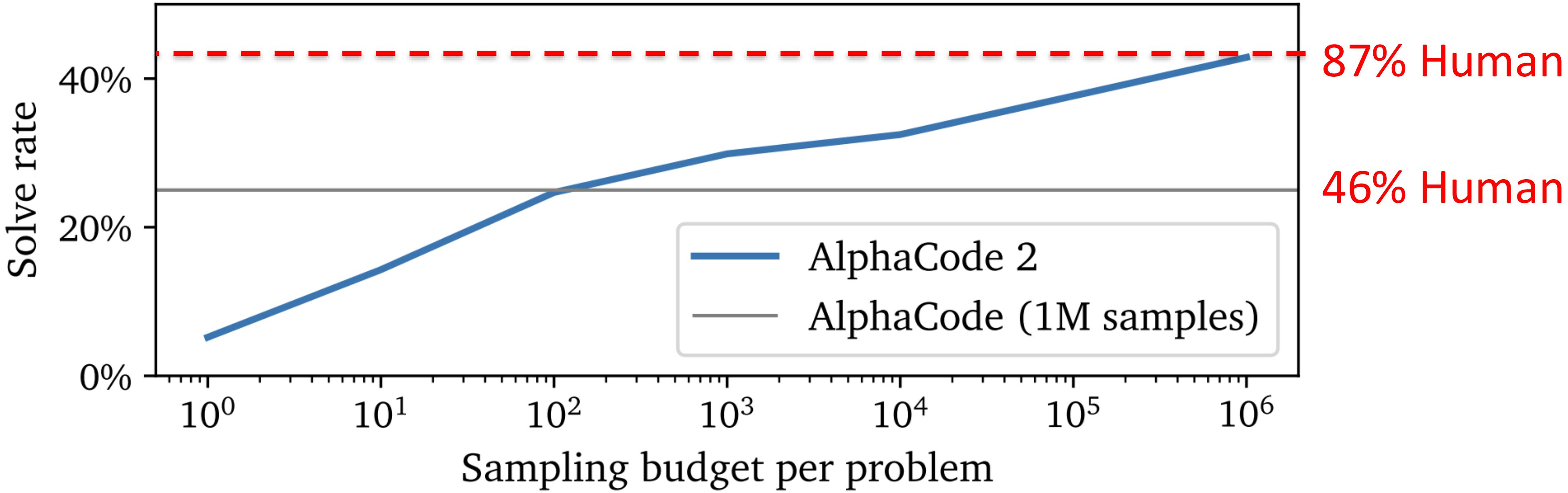
- Binary feedback is not enough
- Fine-grained feedback is better
- Algorithmic iterative refinement is superb

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

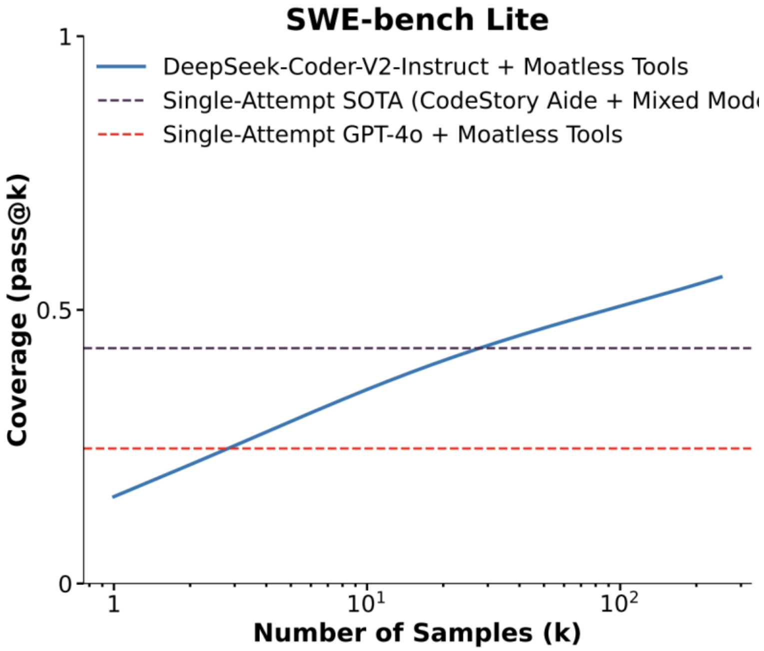
- Aligning with online preference optimization (BPO)
- Iterative refinement with fine-grained feedback (LLMRefine)
- Learning Optimized Sample Compute Allocation (OSCA)
- Speculative Knowledge Distillation

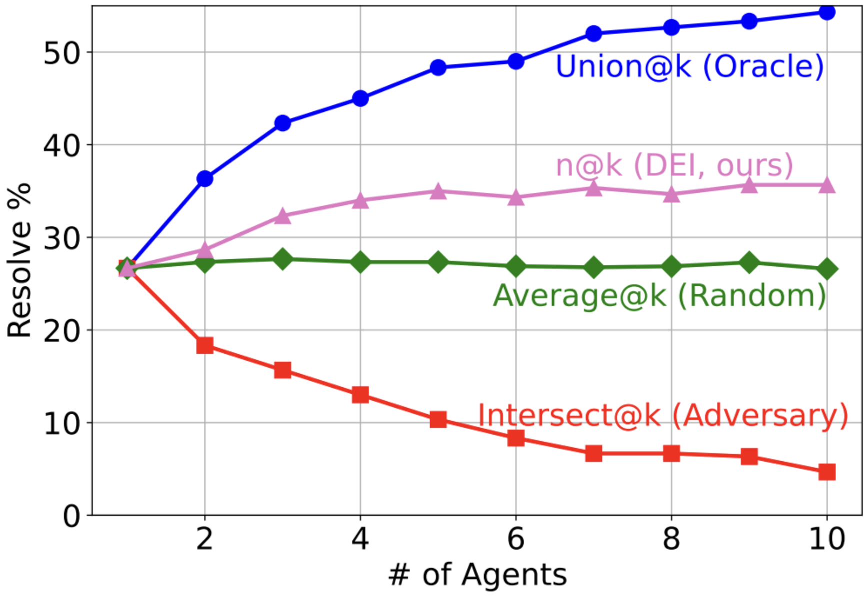
Inference-Time Scaling Law

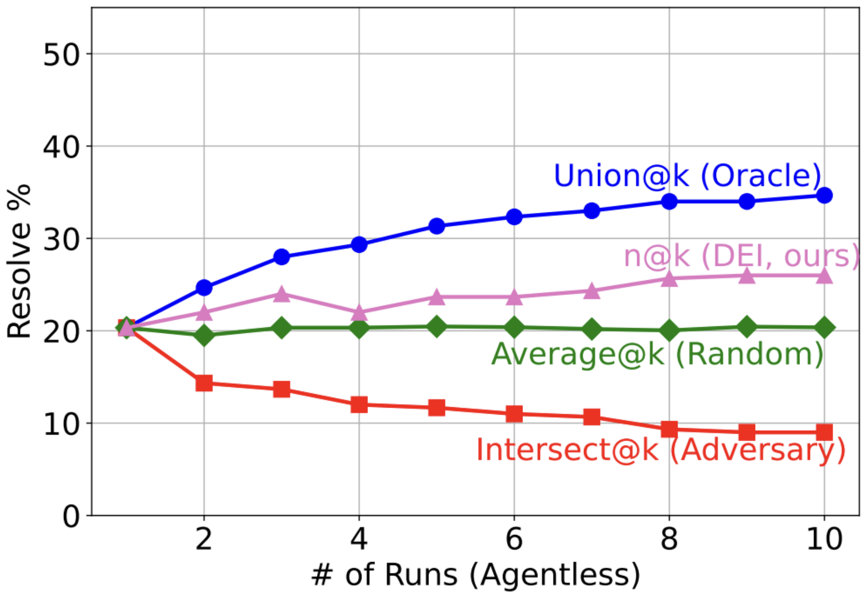


AlphaCode2 Tech Report, https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2_Tech_Report.pdf

Inference-Time Scaling Law







same for Agentic Tasks like SWE-Bench.

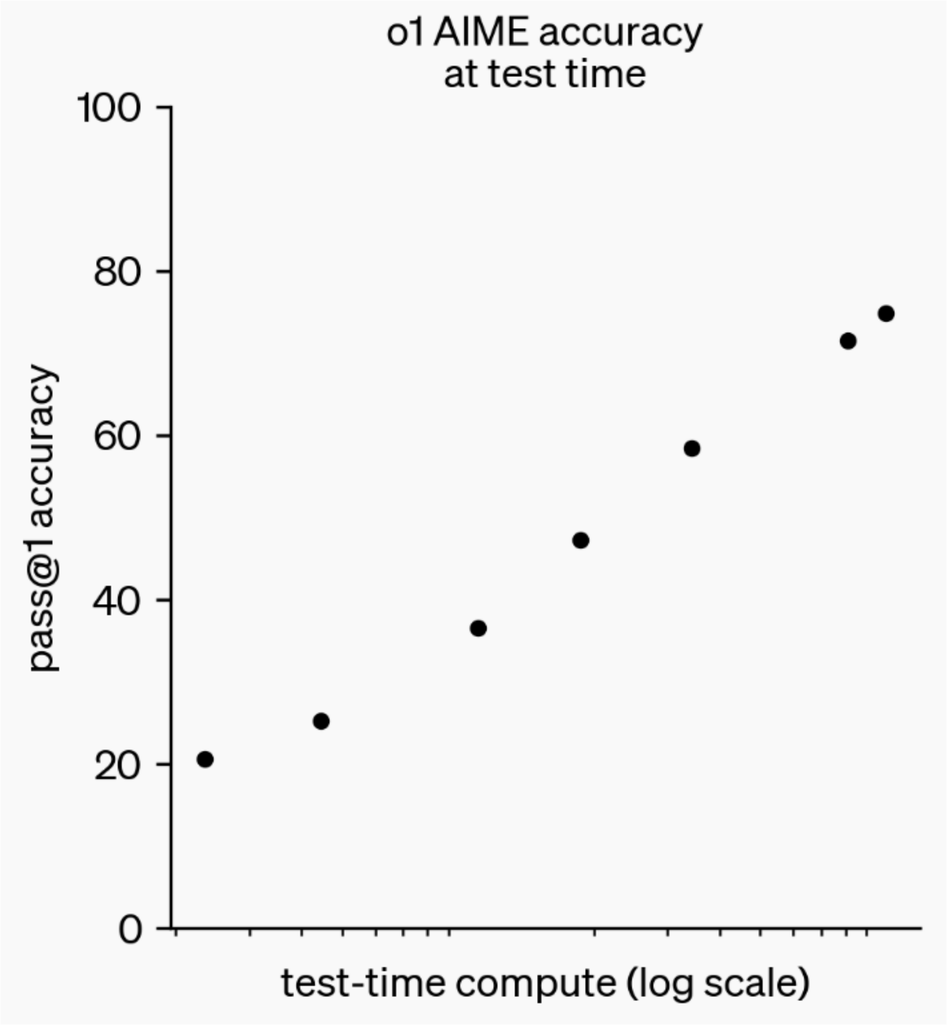
More agents, more runs ➔ Better solve rate.

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/

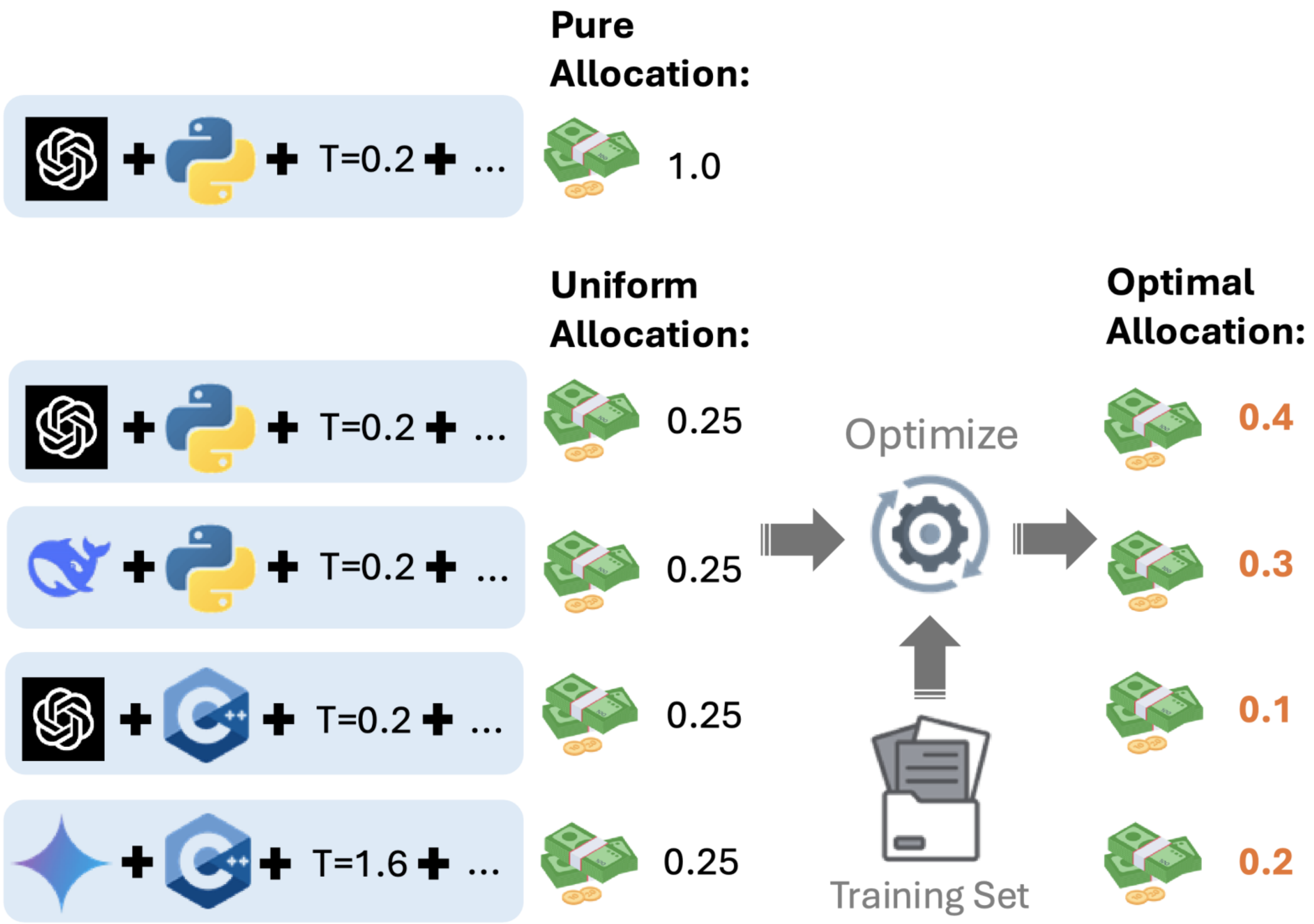
Sample Compute Allocation Problem for LLM Inference

- 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
- Pure Strategy
  - one config uses all compute
- Mixed Strategy

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

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.

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

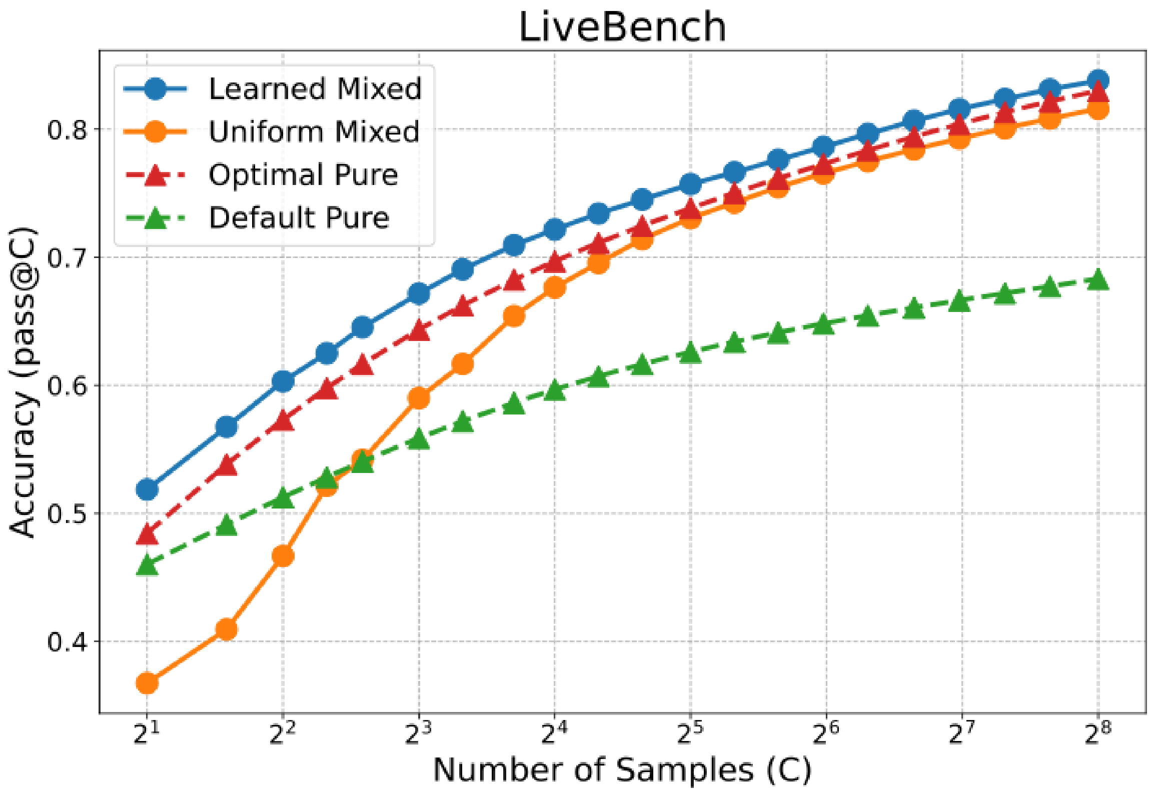
$$= \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \left(1 - \prod_{i=1}^{|\mathcal{H}|} (1 - p_{ij})^{\pi_i} \right),$$

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

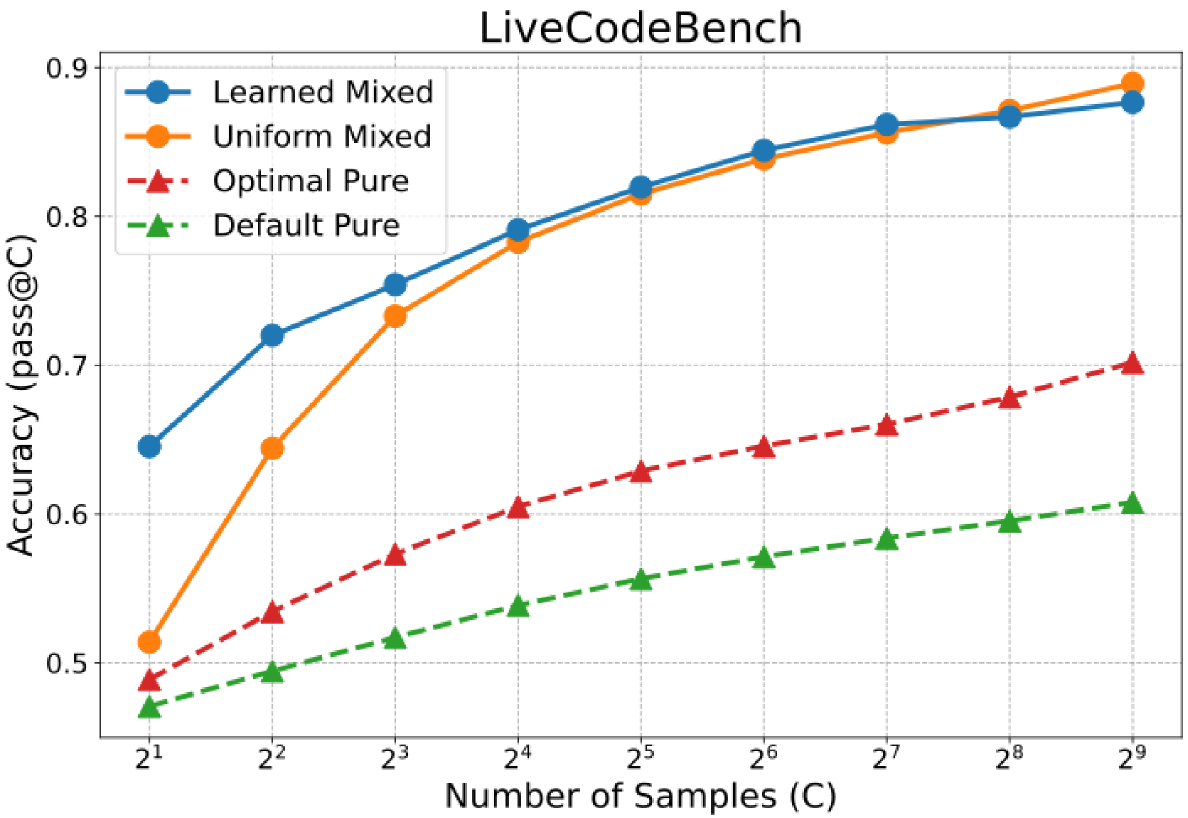
$$\sum_{i=1}^{|\mathcal{H}|} \pi_i = C, \pi_i \in \mathbb{N}.$$

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

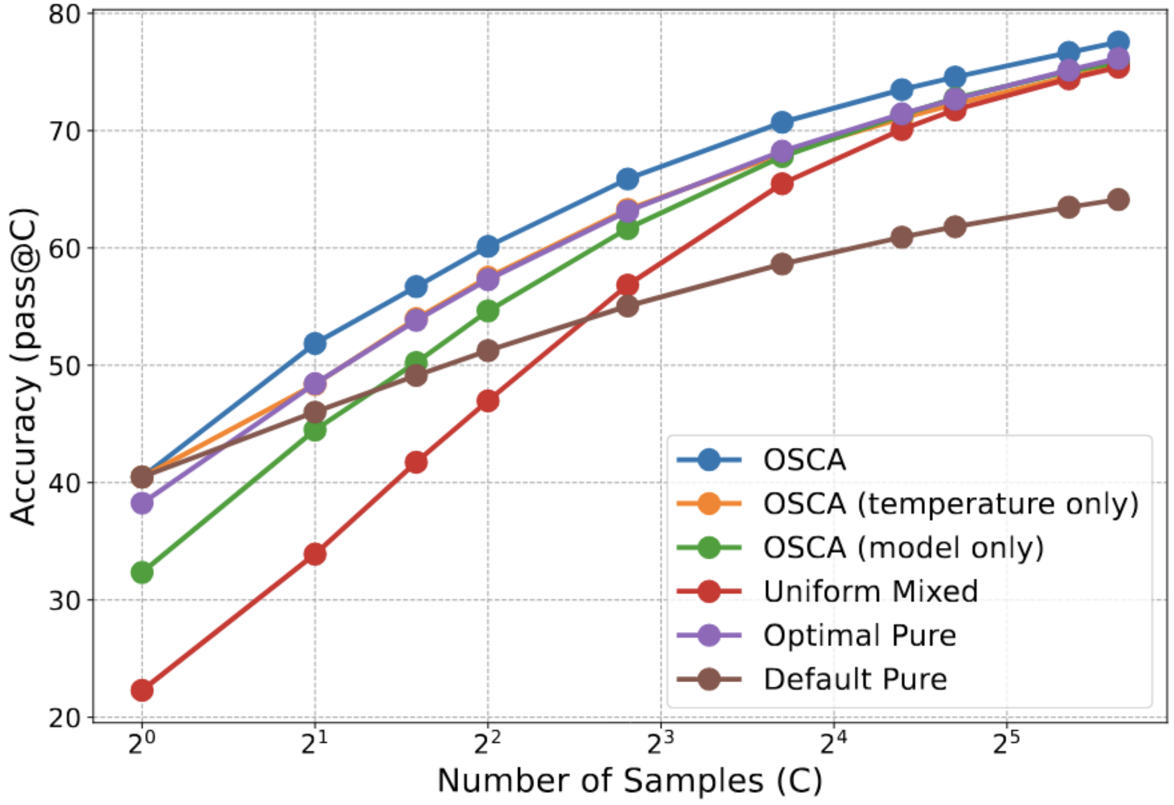


Figure 5: OSCA's pass rates on LiveBench when it is banned from allocating compute to multiple temperatures or multiple models.

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

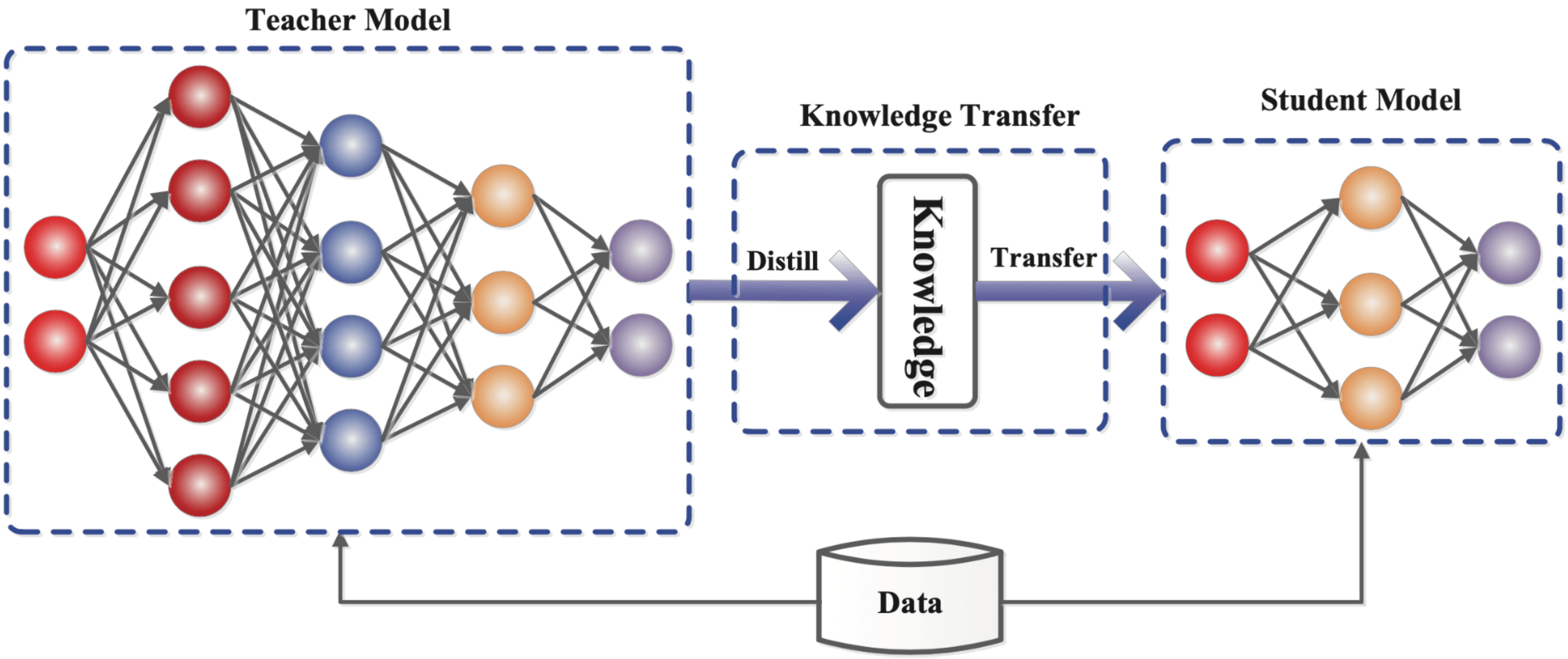
Highlight of OSCA

- 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

Outline

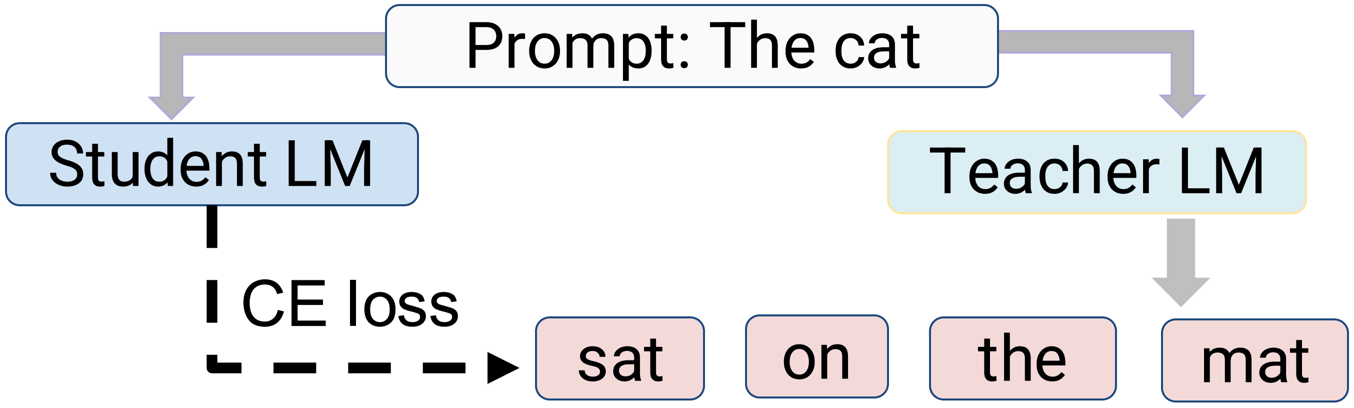
- Aligning with online preference optimization (BPO)
- Iterative refinement with fine-grained feedback (LLMRefine)
- Learning Optimized Sample Compute Allocation (OSCA)
- Speculative Knowledge Distillation

Knowledge Distillation

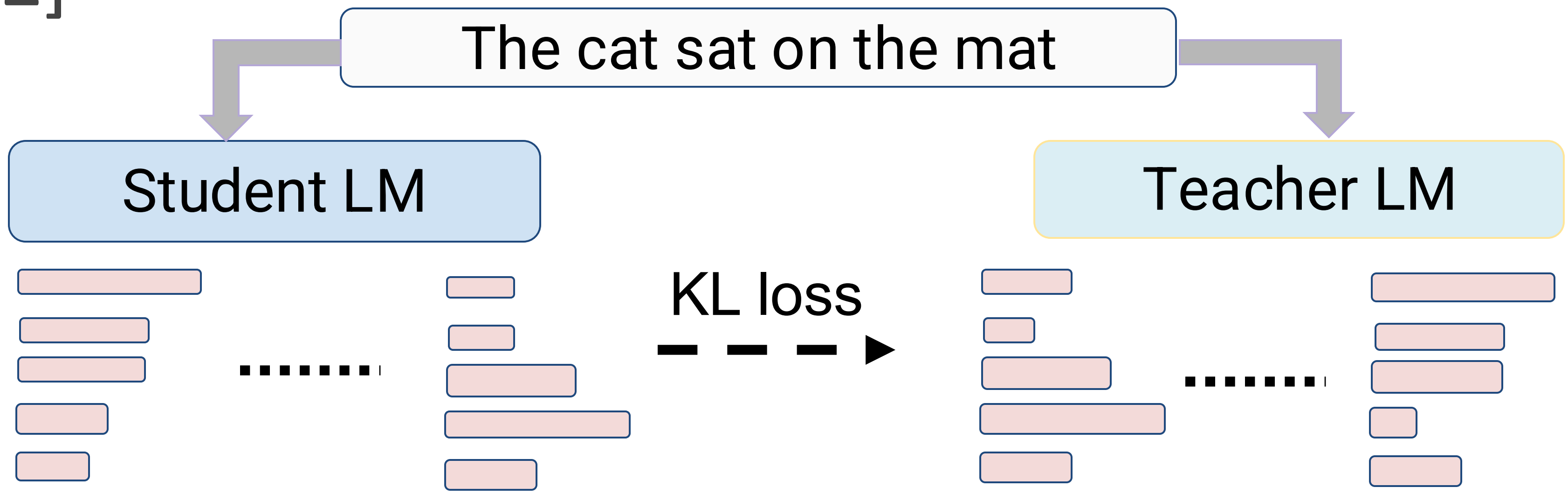


LLM Distillation Approaches

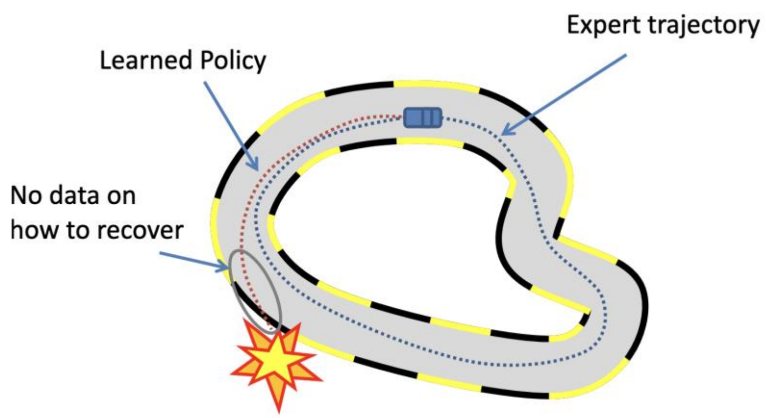
1) Sequence-Level KD [1]



2) Supervised KD [2]



Drawbacks of supervised KD



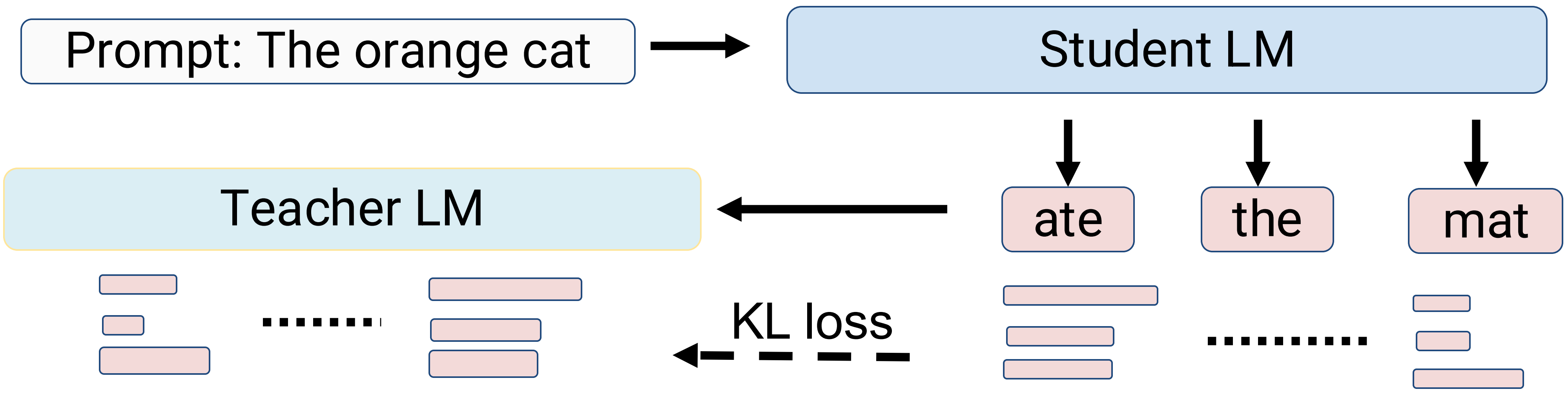
Prompt: The cat | sat | on | the | mat

Prompt: The orange cat → Student LM → ate | the | mat

Supervised KD
- Fixed dataset -> training and inference mismatch
- Student model never learns to correct previous mistakes

Efficient reductions for imitation learning
On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes

On-policy KD can help but not perfect!



On-policy KD

- Low quality student samples makes convergence harder
- Inaccurate assessment from teacher

Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

On-policy KD generates low-quality samples

Prompt : Translate Assamese sentence into English.

Reference : It was to last for the next 40 years and would be fought for real, by proxy armies, on battlefields from Africa to Asia, in Afghanistan, Cuba and many other places.

On-policy (COMET: 36) : This has been a long war, and has been fought in the past, in the past, in the past, in the past, in the past, in the past, in the past......

SKD (COMET: 72): This has been going on for 40 years and is still ongoing, with the possibility of war, a civil war, Afghans in the Balkans, Ethiopia, Kenya, Somalia, and possibly even the United States

Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

On-policy samples are OOD to teacher

Student Gemma-2b-it	On-policy	SKD (Interleaved)	Supervised KD
Teacher PPL	46.8	10.8	1.57

Teacher samples are OOD to student

Student Gemma-2b-it	On-policy	SKD (Interleaved)	Supervised KD
Student PPL	44.0	358	5606.3

Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

Speculative KD: Interleaved text sampling

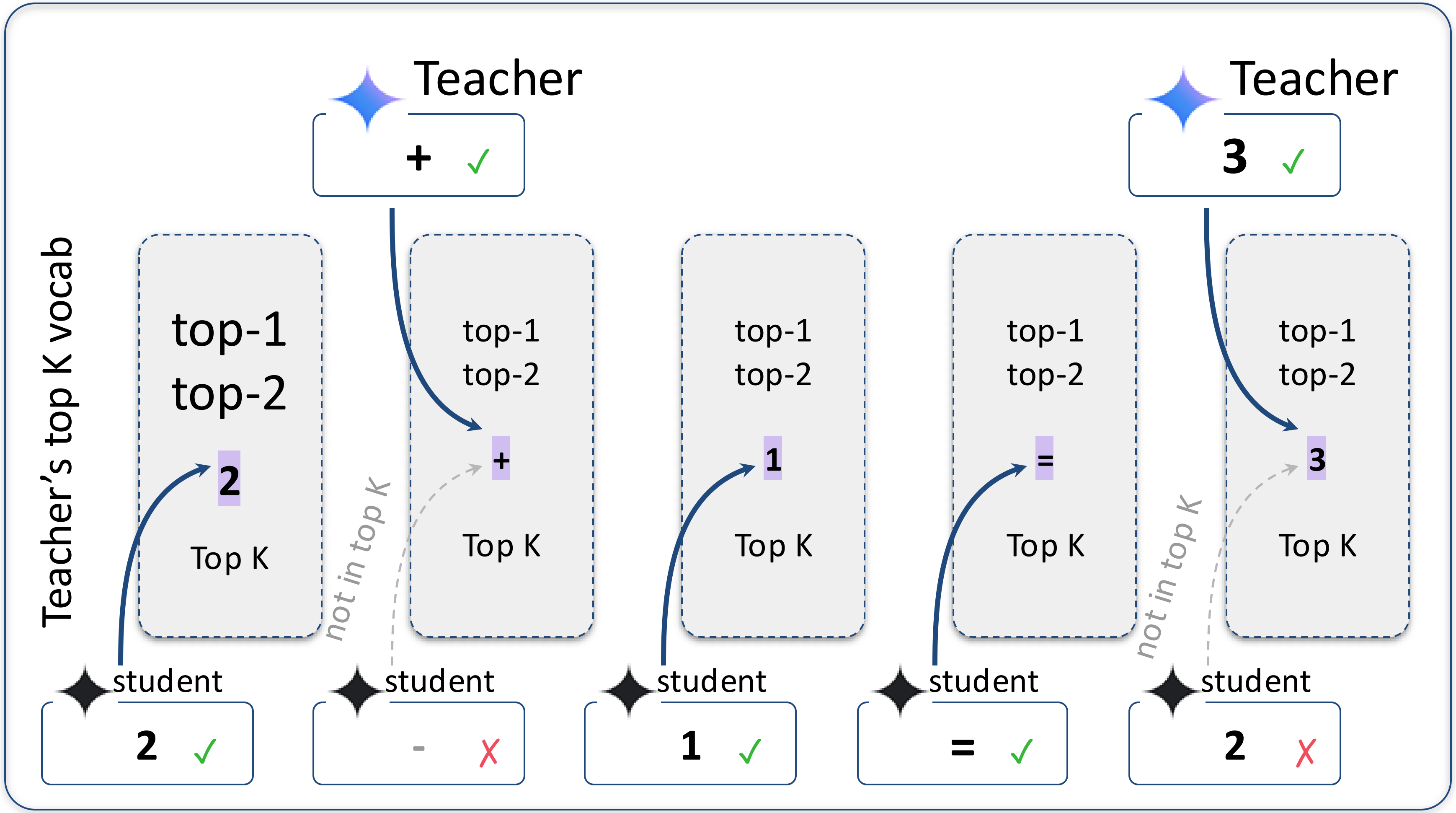
Input

Bob ate two apples yesterday and ate one apple today.
How many apples did he eat? Think step by step.

Output

On-Policy KD: 2-1=2

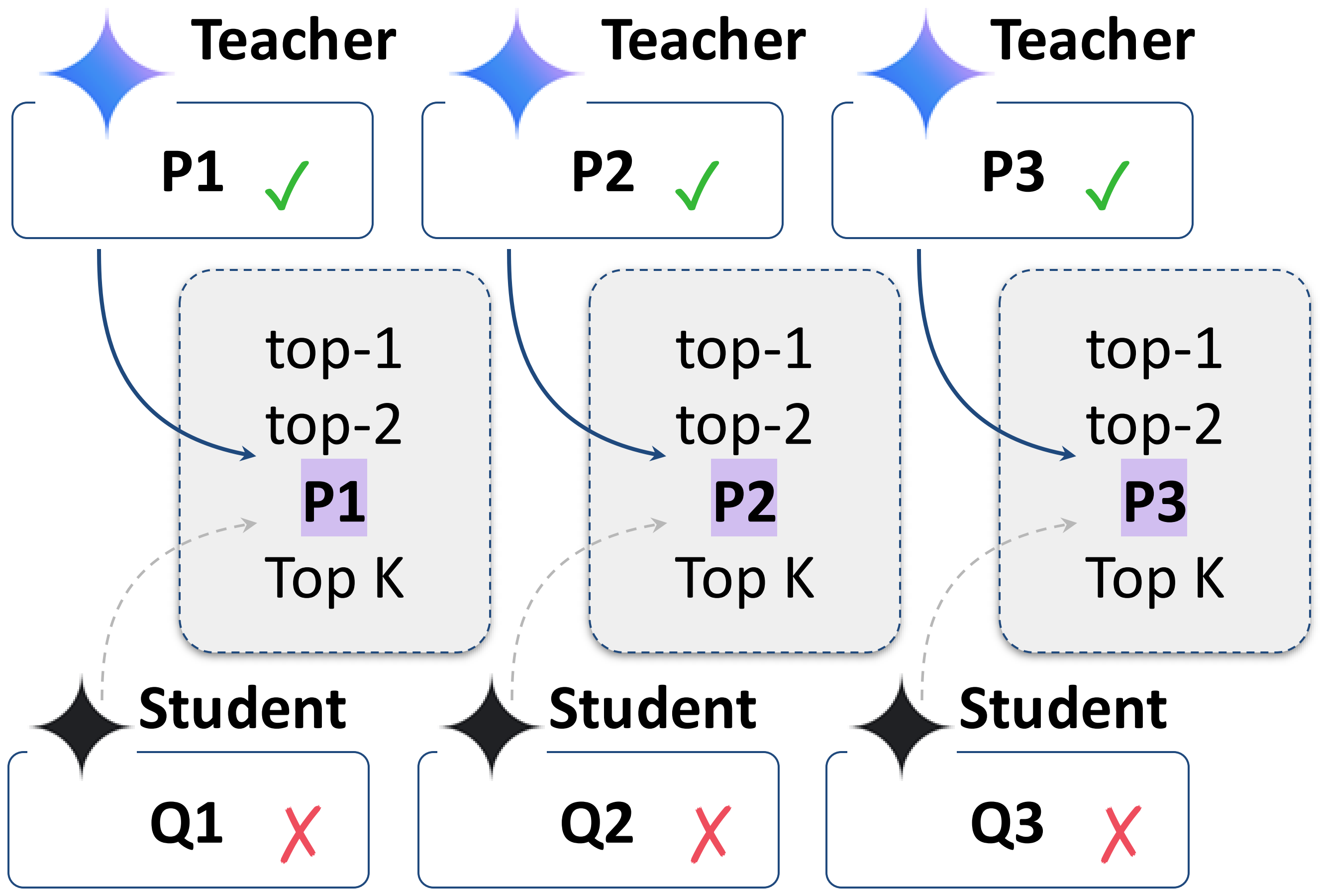
SKD (ours): 2+1=3



Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

SKD degenerates to Supervised KD

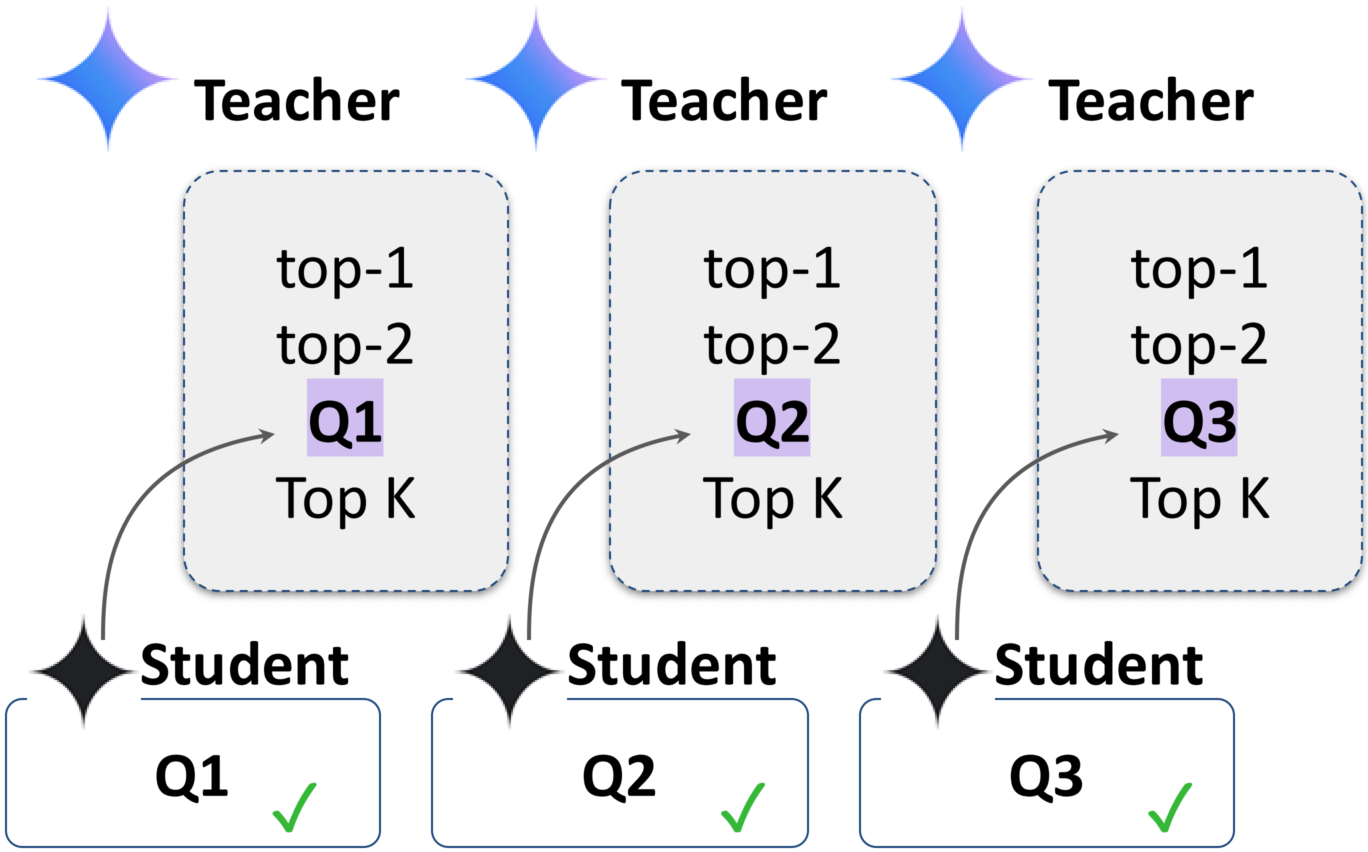
SKD degenerates to **Supervised KD** when the teacher rejects all tokens proposed by the student



Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

SKD degenerates to On-policy KD

SKD degenerates to **On-policy KD** when the teacher accepts all tokens proposed by the student



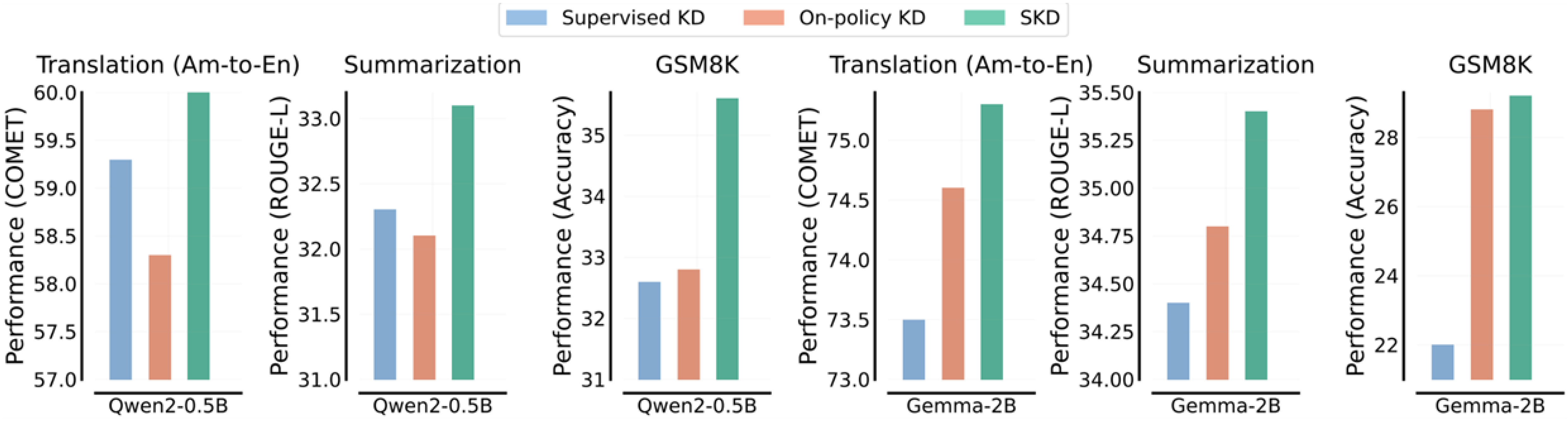
Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

Evaluation Setup

- Tasks:
  - Translation (Am-En), Dialogue summarization,
  - Arithmetic reasoning (GSM-8k) and Math instruction following task
- Baselines:
  - SFT, Supervised KD, on-policy KD and ImitKD
- Base Model:
  - Teacher: Gemma-7B and Qwen2-7B
  - Student: Gemma-2B and Qwen2-0.5B

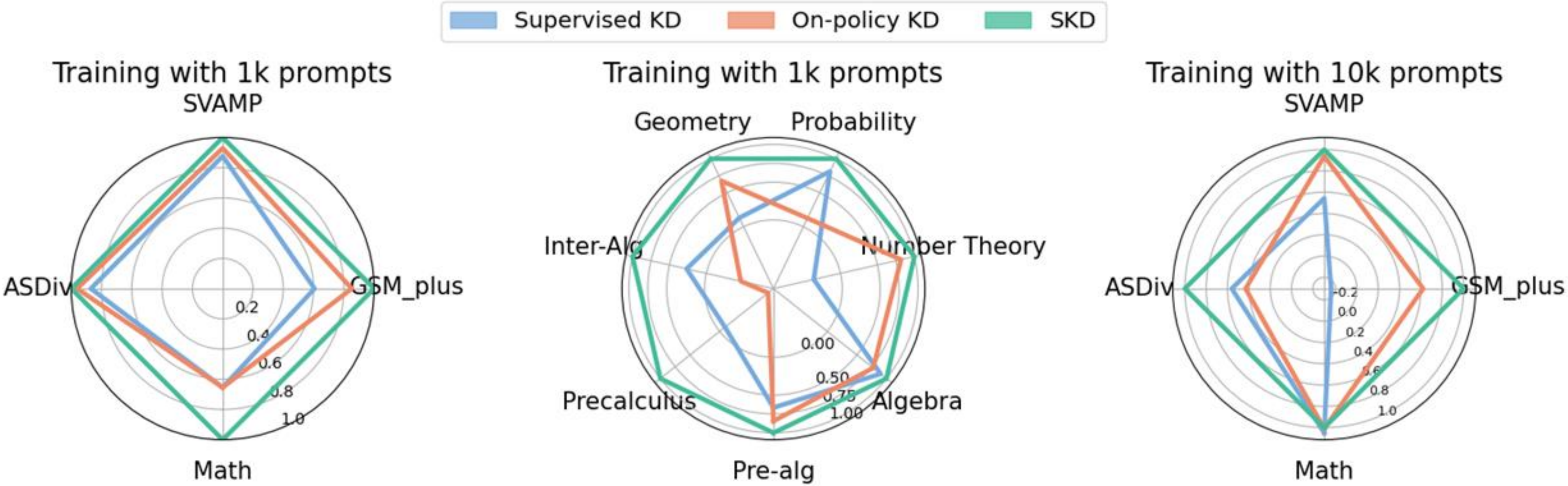
Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

SKD outperform all baselines in task-specific distillation



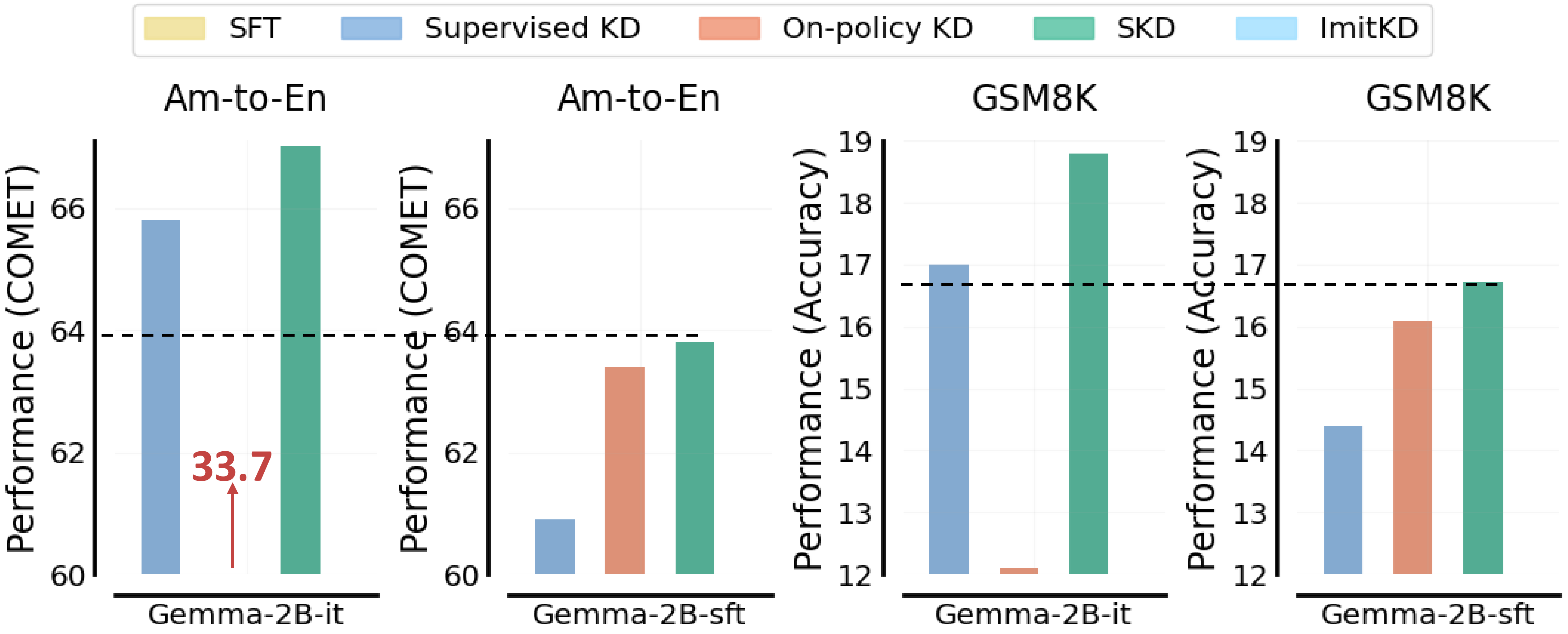
Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

SKD outperforms all baselines in task-agnostic distillation



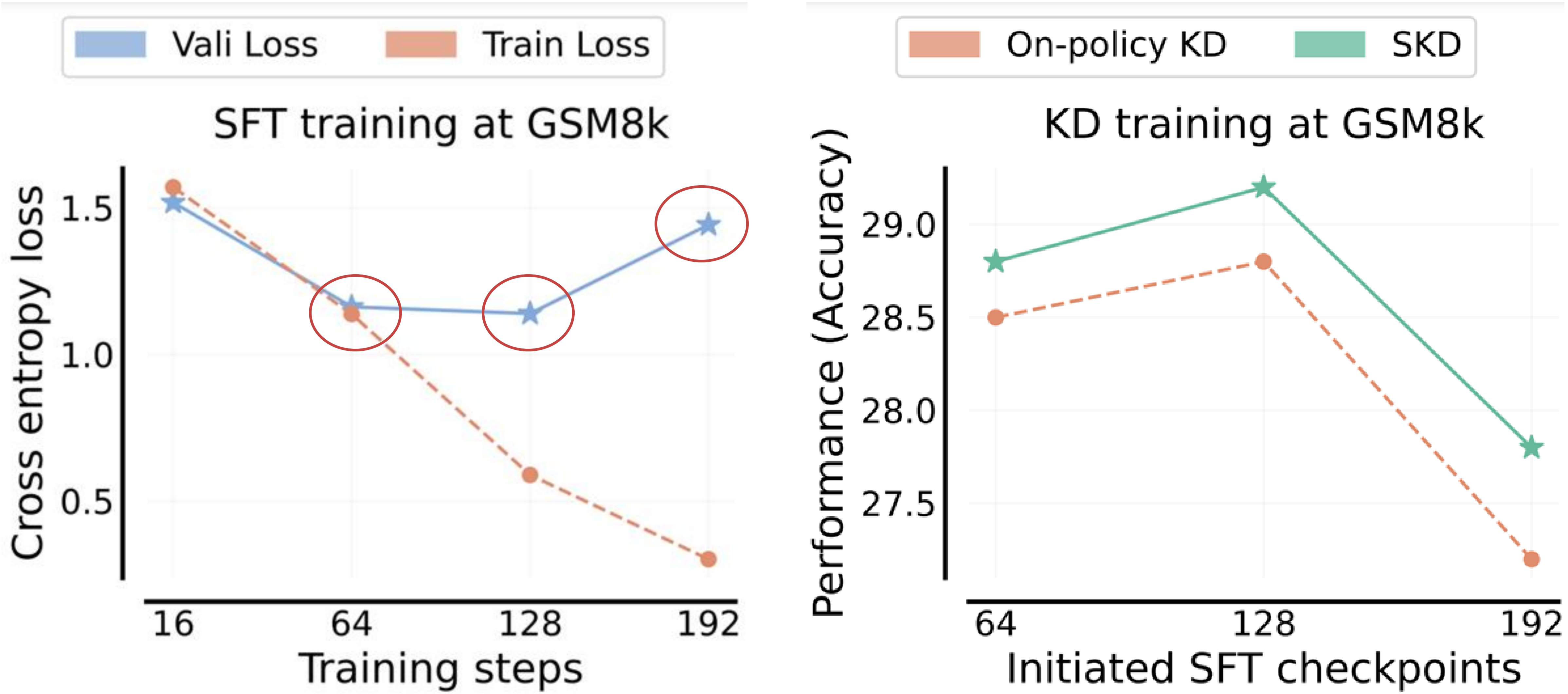
Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

SKD outperforms all baselines under low data regime (100 data points)



Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

Overfitting at SFT stage is bad for KD



Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

Highlights of SKD

- SKD mitigate training-inference mismatch
- SKD reduces low-quality samples from poor student
- Superior performance than supervised KD and on-policy KD across various text generation tasks in both task-specific and task-agnostic scenarios.
- More robust to different initialization.

Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.

Summary and Takeaway

- 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
- Speculative Knowledge Distillation for LLM
  - an adaptive version of supervised seq and on-policy distillation

Reference

- Xu, Wang, Pan, Song, Freitag, Wang, Li. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.
- 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.
- Wenda Xu et al. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling. ICLR 2025.