ICLR 2024

On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes

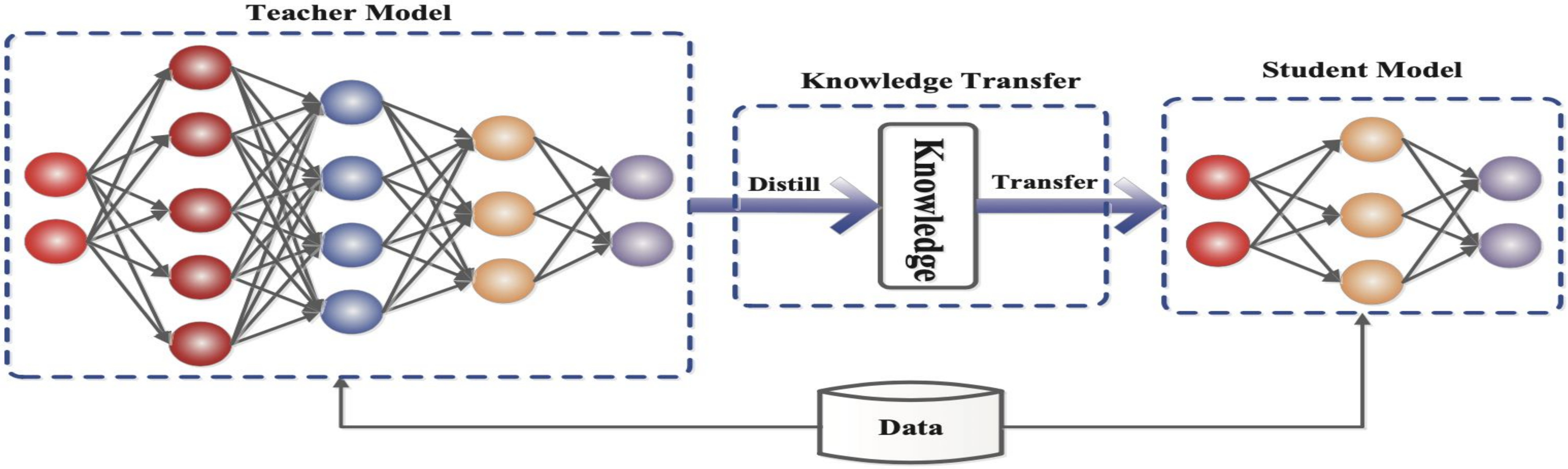
Rishabh Agarwal*, **Nino Vieillard*,** Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, Olivier Bachem

Google DeepMind

Knowledge Distillation (KD)

Goal: Transfer knowledge from a **teacher** model into a **smaller student** model.

Why: Deployment of "large" models often limited by their **inference cost** or **memory footprint**.

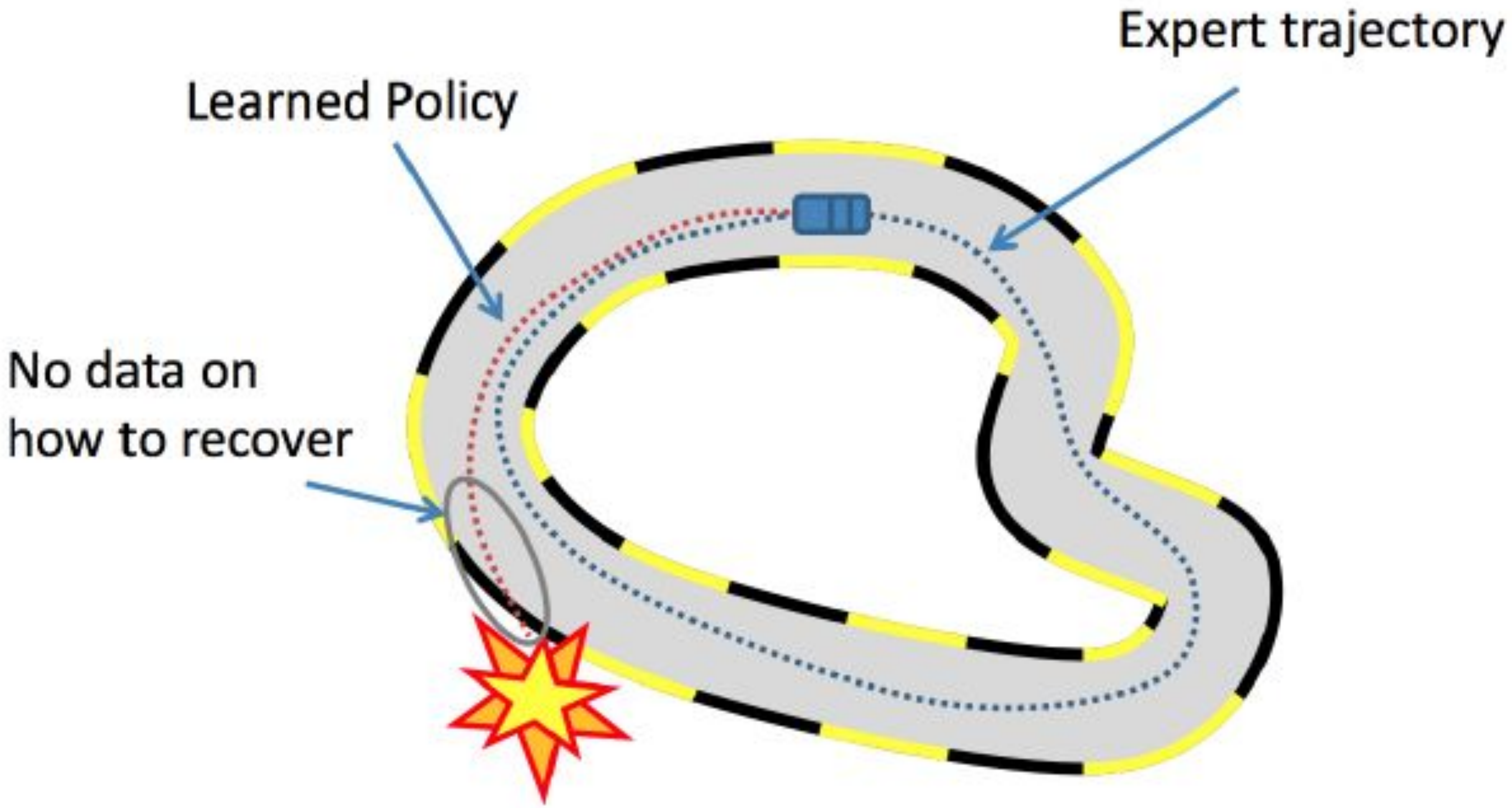


The generic framework of teacher-student knowledge distillation training. (Image source: Gou et al. 2020)

KD for LLMs: Distribution Mismatch (Exposure Bias)

Existing KD methods typically train on a **fixed** dataset of output sequences.

This results in a train-inference mismatch.

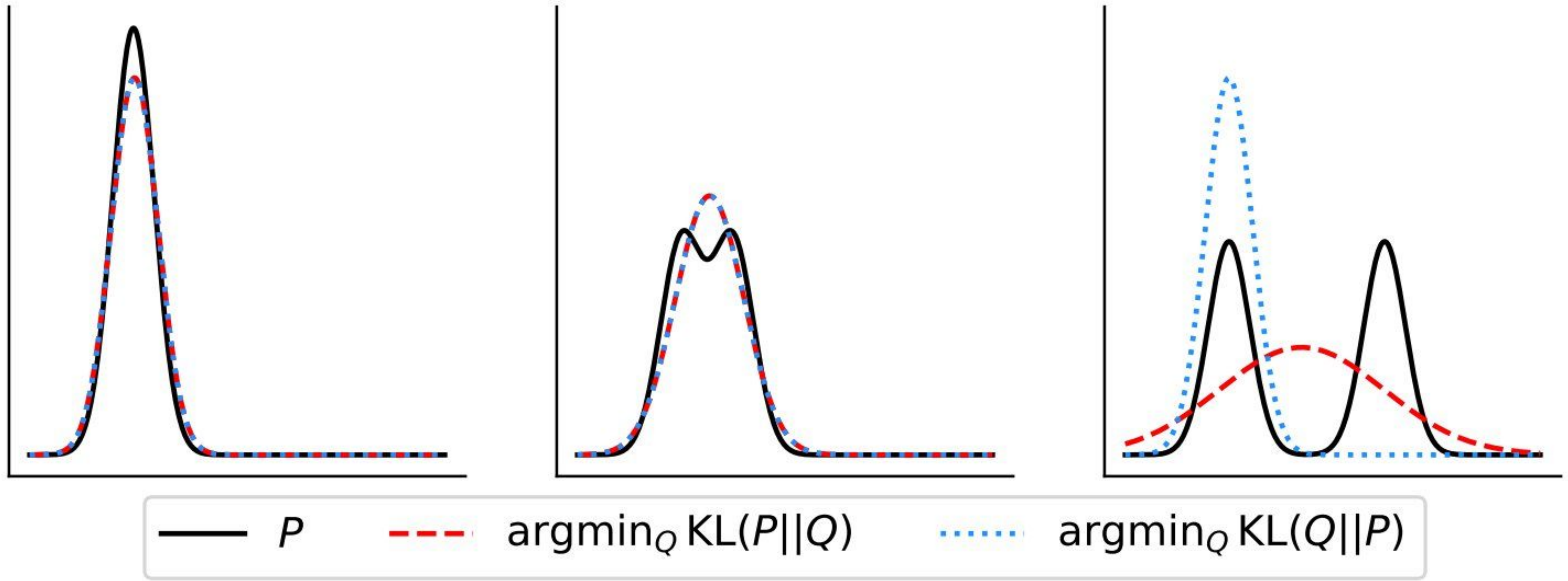


See DaggER. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning.

KD for LLMs: Model Capacity Mismatch

If student is often not expressive enough to fit the teacher's distribution, standard KD objective can lead to *unnatural* student-generated samples.

Supervised KD = KL(Teacher || Student), which is *mode-covering*.



RL-Inspired Soft-Distillation: On-Policy GKD

RL-Inspired Soft-Distillation: On-Policy GKD

1) **On-policy Data:** Sample self-generated output sequences from the student model.

RL-Inspired Soft-Distillation: On-Policy GKD

1) **On-policy Data:** Sample self-generated output sequences from the student model.

2) **Feedback:** Run inference on the teacher to get logits on student generated sequences (what teacher would predict given some text).

RL-Inspired Soft-Distillation: On-Policy GKD

1) **On-policy Data:** Sample self-generated output sequences from the student model.

2) **Feedback:** Run inference on the teacher to get logits on student generated sequences (what teacher would predict given some text).

3) **Supervised Training**: Minimize mismatch (e.g., KL-divergence) between student and teacher *token-level logits*.

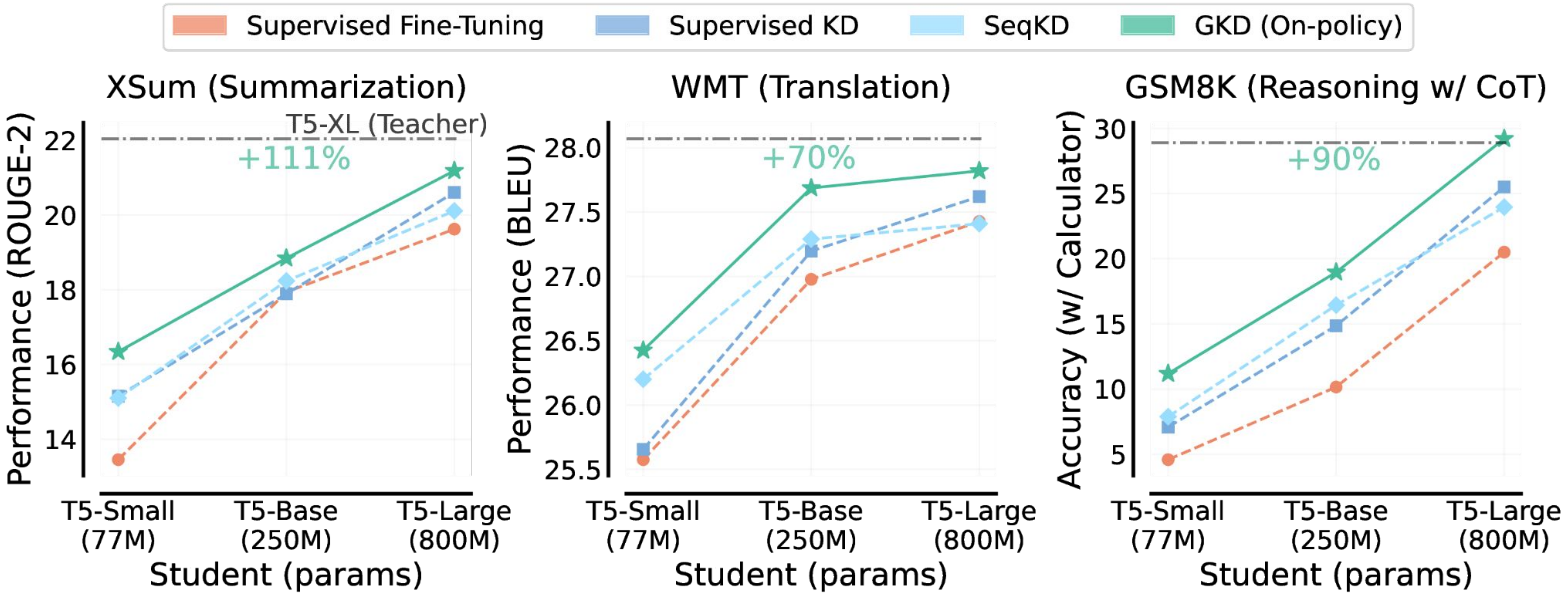
SFT Results with GKD

Setup:

- One SFT task (eg summarization, translation)
- Teacher = **big LM fine tuned on the task**
- Student = **small LM fine tuned on the task**
- Goal = close the performance gap between the two

SFT Results with GKD

On-Policy GKD **consistently improves** over common distillation approaches (SFT, SeqKD, Supervised KD) on different tasks.



On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. Agarwal*, Veillard* et al. ICLR 2024.

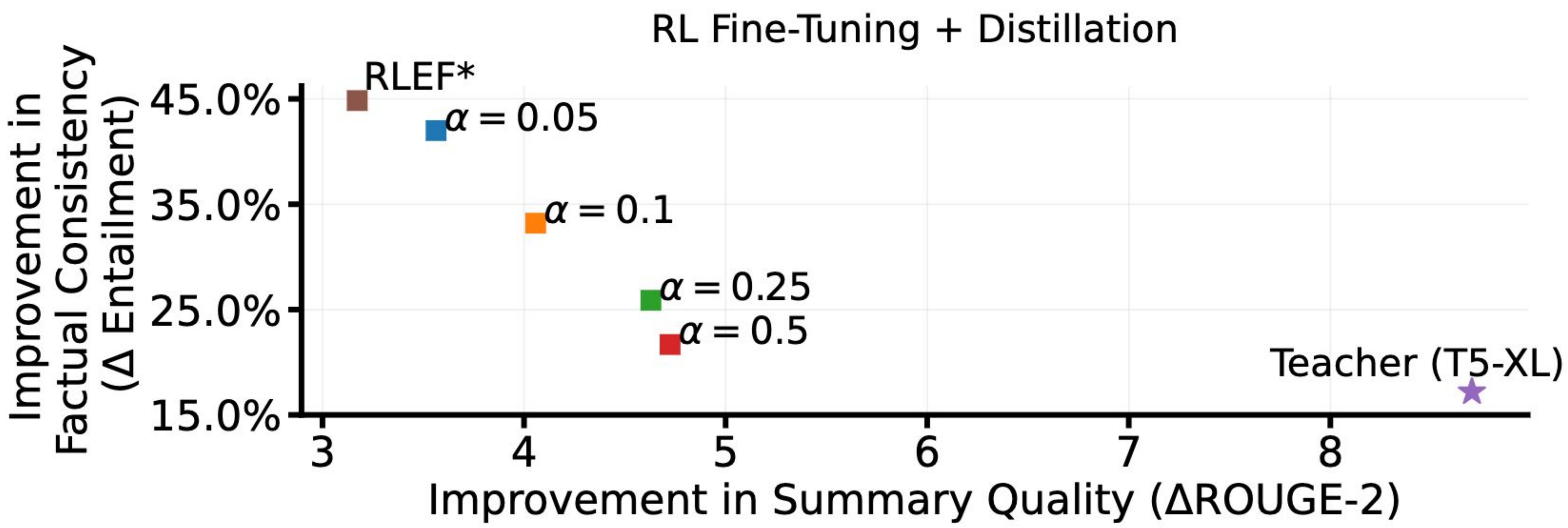
GKD + RLHF: A Natural Combination

GKD + RLHF: A Natural Combination

RL for LLMs is **regularized towards the initial policy**.

Instead, **regularize towards the teacher**: combine the two objectives !

$$\mathbb{E}_{x \sim X}\left[\underbrace{(1-\alpha)\, E_{y \sim p_S^\theta(\cdot|x)}\left[r(y)\right]}_{\text{RL objective}} - \underbrace{\alpha\, \mathbb{E}_{y \sim p_S(\cdot|x)}\left[\mathcal{D}_{KL}\left(\boxed{p_S^\theta(y|x) \| p_T(y|x)}\right)\right]}_{\text{Generalized On-Policy KD}} \right],$$



Conclusion and thanks!

- Takeaway message:
  - If you distill LLMs for an SFT task, **do it on the student distribution**

- Come visit us at the poster !
  - Friday, 4:30 pm, Halle B