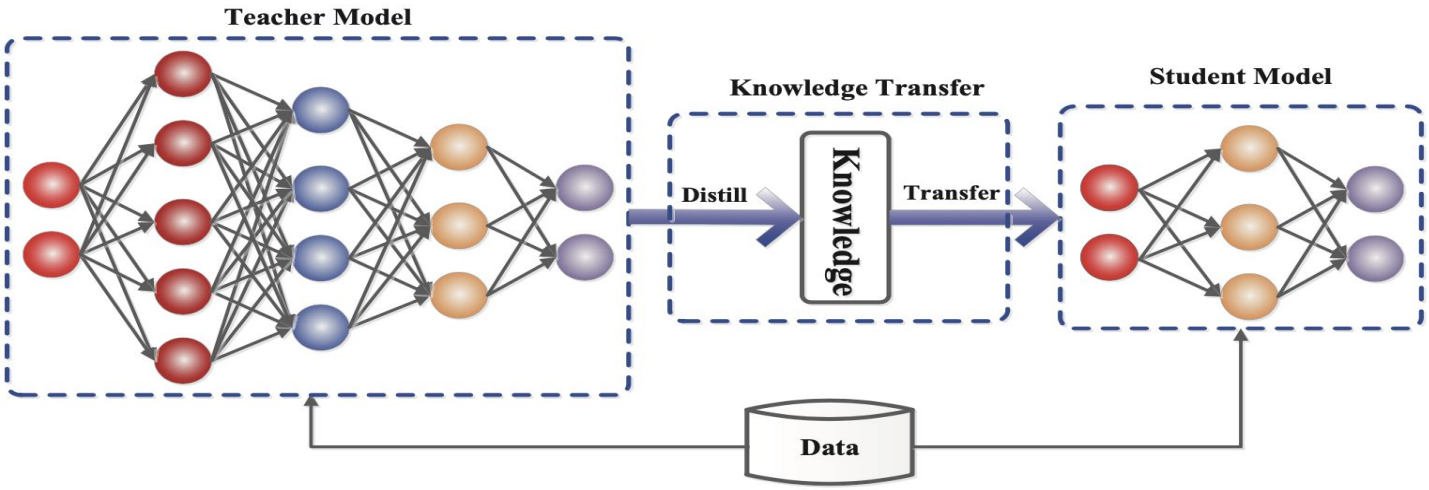
Post-Training Distillation for LLMs

Rishabh Agarwal

Knowledge Distillation (KD)

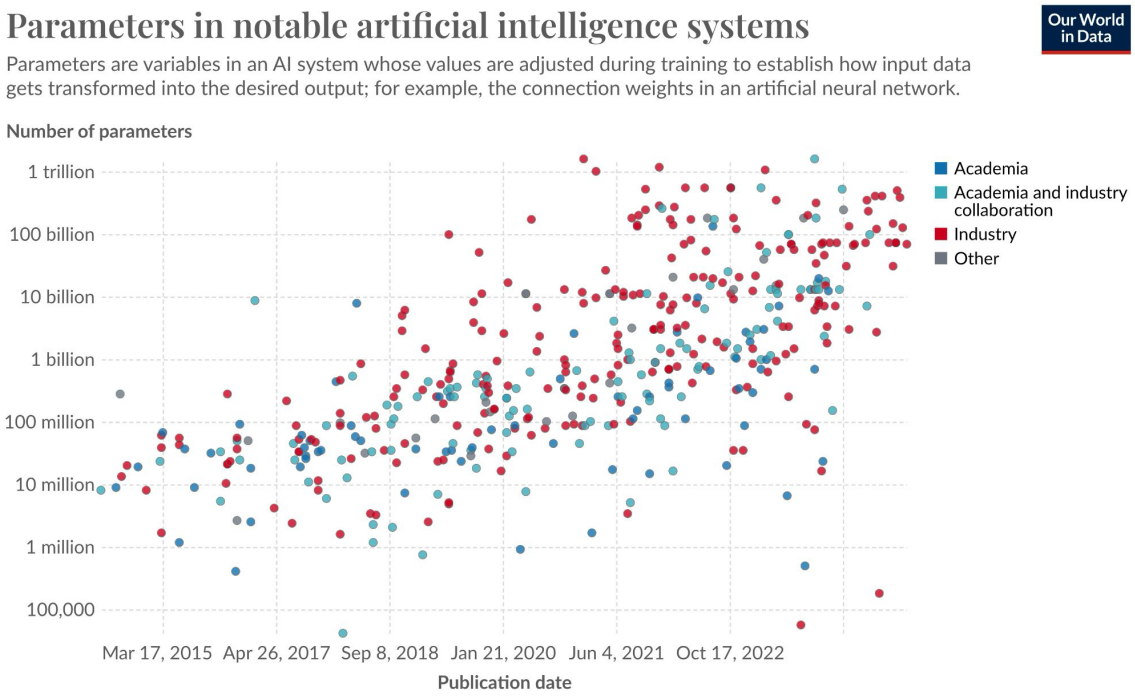


The generic framework of teacher-student knowledge distillation training. (Image source: Gou et al. 2020)

Goal: Transfer knowledge from **expensive teacher** model(s) into a **smaller* student** model, while retaining capabilities.

*Self-distillation uses the same-sized student and teacher, and can still result in better performance.

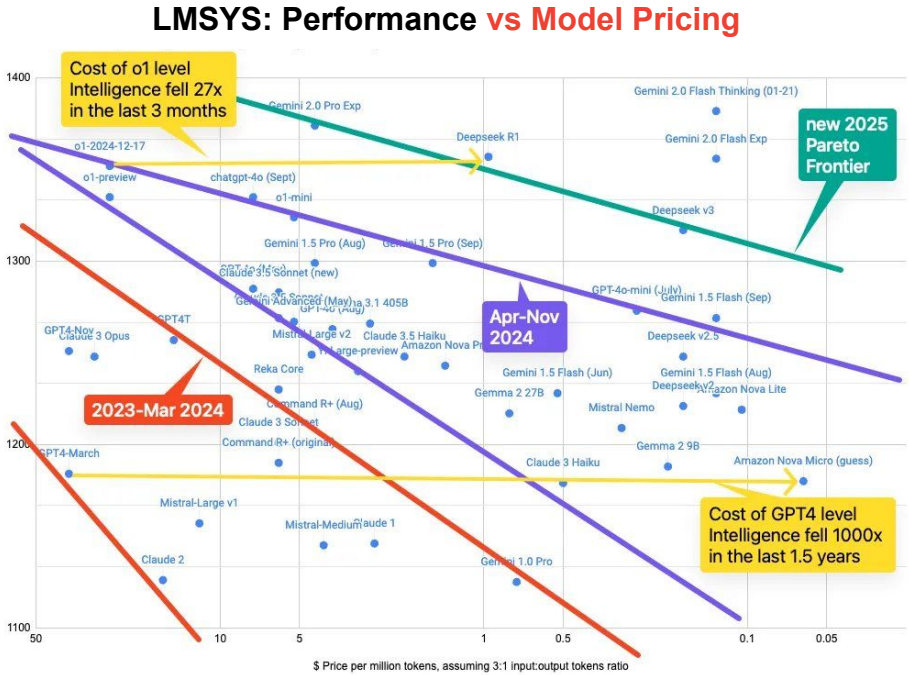
Distillation: Aren’t Bigger LLMs Always Better?



Distillation: Aren't Bigger LLMs Always Better? No

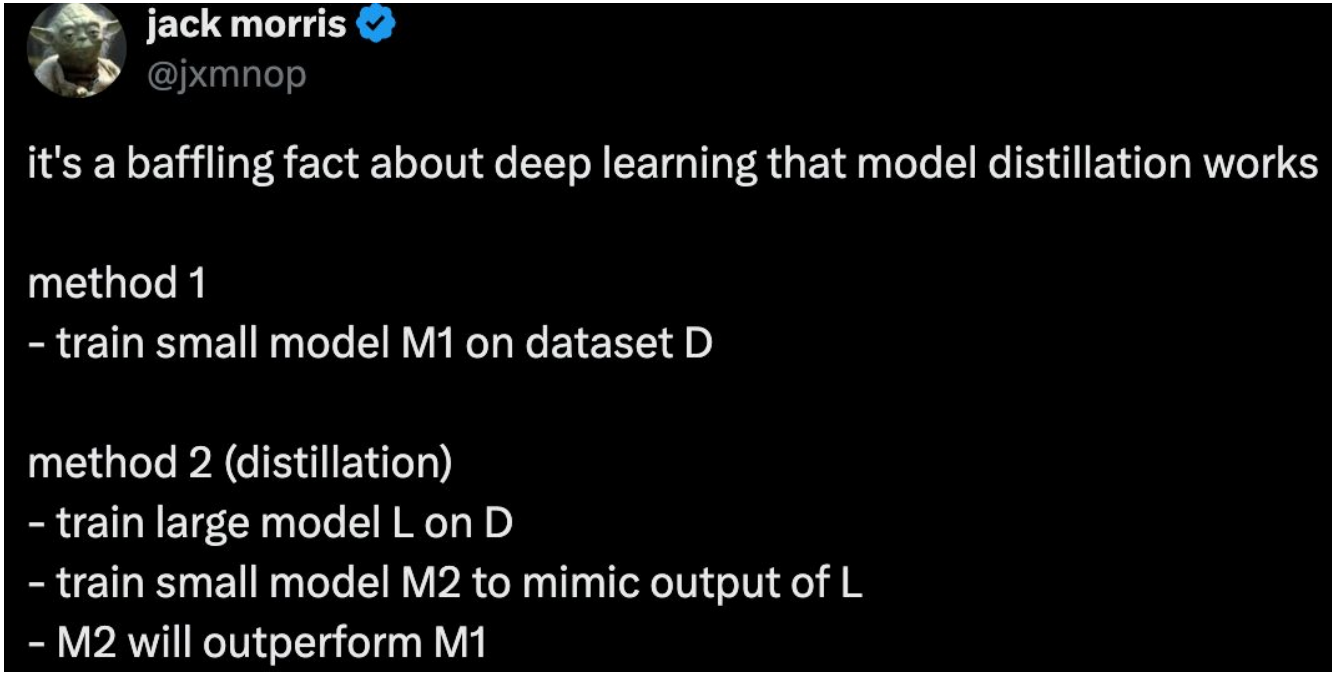
Deployment of LLMs often limited by their inference cost or memory footprint

- Putting 100B parameters on your smartphone needs a lot of memory.
- You typically don't want to wait several minutes for getting an output.



Credit: This post by @swyx

Distillation Recovers “Dark Knowledge”



Source: https://x.com/jxmnop/status/1877761437931581798

See Dark Knowledge (Slides from Geoff Hinton)

This Tutorial:
Covers Progress in Distillation, especially for LLMs (Biased towards some of my own work)

Distilling the Knowledge in a Neural Network

Geoffrey Hinton[*†]
Google Inc.
Mountain View
geoffhinton@google.com

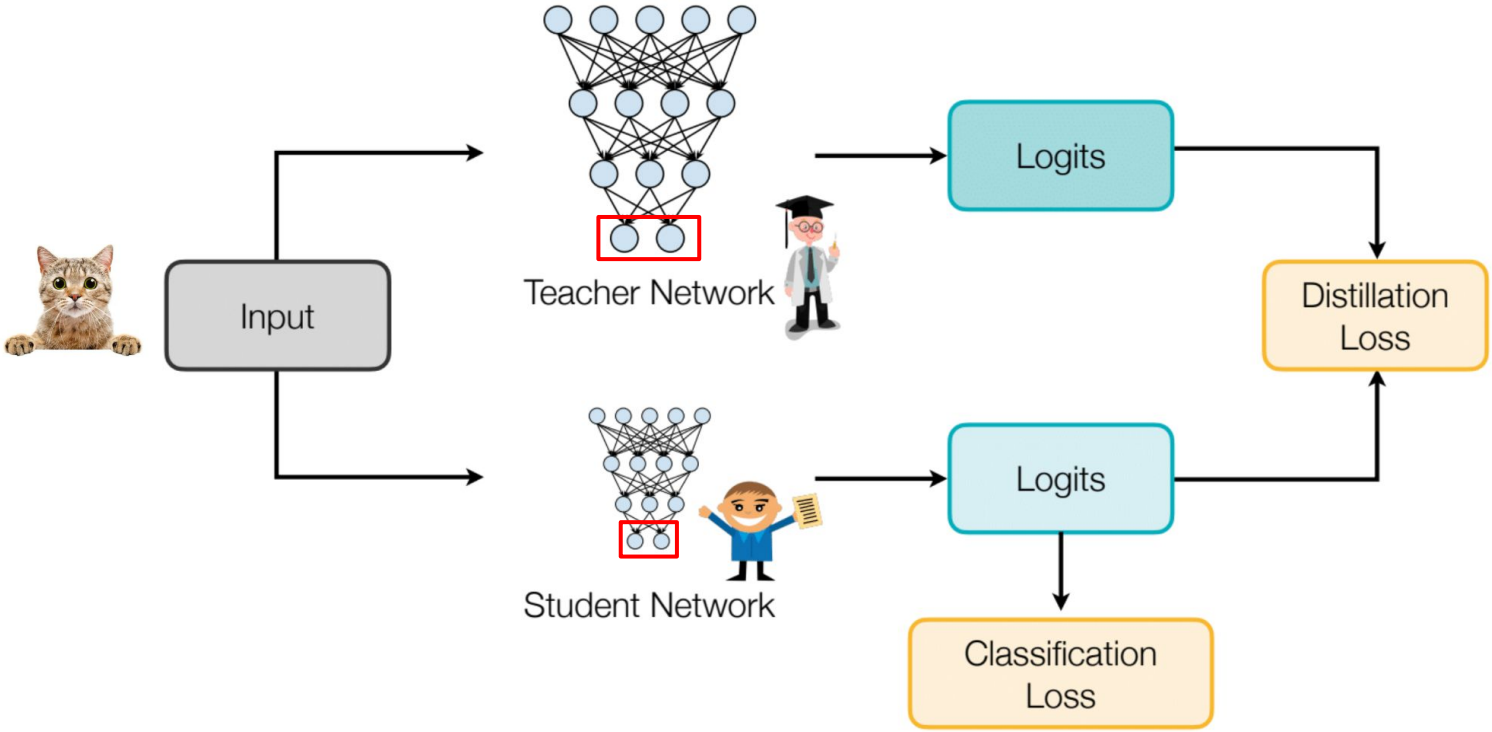
Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com

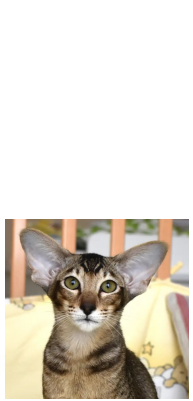
Abstract

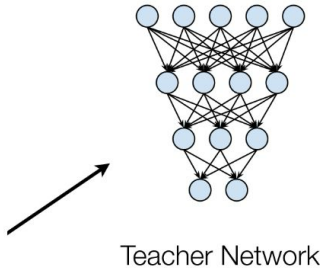
A very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions [3]. Unfortunately, making predictions using a whole ensemble of models is cumbersome and may be too computationally expensive to allow deployment to a large number of users, especially if the individual models are large neural nets. Caruana and his collaborators [1] have shown that it is possible to compress the knowledge in an ensemble into a single model which is much easier to deploy and we develop this approach further using a different compression technique. We achieve some surprising results on MNIST and we show that we can significantly improve the acoustic model of a heavily used commercial system by distilling the knowledge in an ensemble of models into a single model. We also introduce a new type of ensemble composed of one or more full models and many specialist models which learn to distinguish fine-grained classes that the full models confuse. Unlike a mixture of experts, these specialist models can be trained rapidly and in parallel.

Supervised Distillation (Hinton et al., 2015)



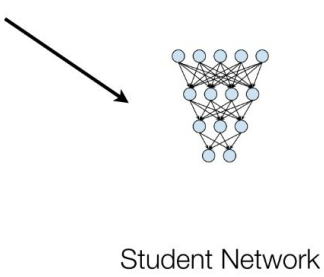
Supervised Distillation (Hinton et al., 2015)





	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

$$\frac{\exp(5)}{\exp(5)+\exp(1)} \qquad \frac{\exp(1)}{\exp(5)+\exp(1)}$$



	Logits	Probabilities
Cat	3	0.731
Dog	2	0.269

The student model is less confident

Distillation Loss: Match the prediction probabilities between the student and teacher.

Supervised Distillation (Hinton et al., 2015)

Teacher probabilities

Student probabilities over the labels “y”

$$\min_\theta D_{KL}\big(q(x)||p_\theta(x)\big)$$

$$= \min_\theta \sum_y q(y|x)\big(\log q(y|x) - \log p_\theta(y|x)\big)$$

$$= \min_\theta \underbrace{\sum_y -q(y|x)\log p_\theta(y|x)}_{\text{Cross-entropy loss}}$$

Teacher Network

	Logits	Probabilities
Cat	5	0.982
Dog	1	0.017

$\frac{\exp(5)}{\exp(5)+\exp(1)}$

$\frac{\exp(1)}{\exp(5)+\exp(1)}$

Student Network

	Logits	Probabilities
Cat	3	0.731
Dog	2	0.269

The student model is less confident

Cross Entropy Loss: Match the prediction probabilities between the teacher q and student p_θ

Supervised Distillation for Language Models

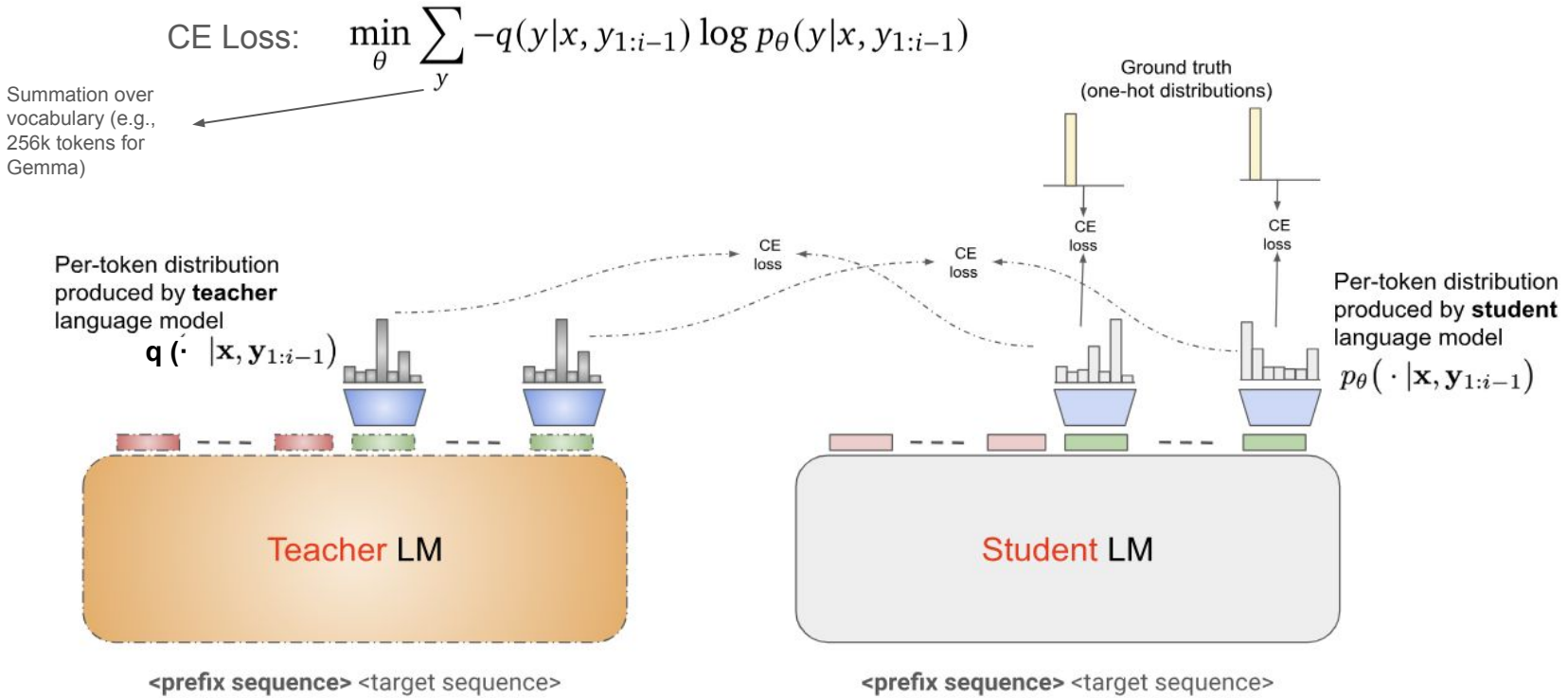


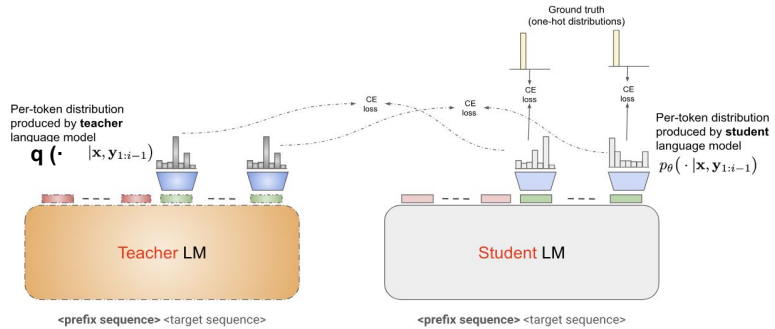
Figure Credit: Yongchao Zhou

Supervised KD Optimizes Forward KL

$$\min_\theta KL\big(\overbrace{q(x)}^{\text{Teacher}} \,||\, \overbrace{p_\theta(x)}^{\textit{Student}}\big)$$

$$= \min_\theta \mathbb{E}_{y\sim q(x)}\left[\sum_i \underbrace{KL\big(q(\cdot|x, y_{1:i-1})||p_\theta(\cdot|x, y_{1:i-1})\big)}_{\text{Per-Token KL}}\right]$$

$$= \min_\theta \mathbb{E}_{y\sim q(x)}\left[\underbrace{\sum_i \sum_{y_i\in\mathcal{V}} -\underbrace{q(y_i|x, y_{1:i-1})}_{\text{Soft Labels}} \log p_\theta(y_i|x, y_{1:i-1})}_{\text{Next token Prediction Loss}}\right]$$



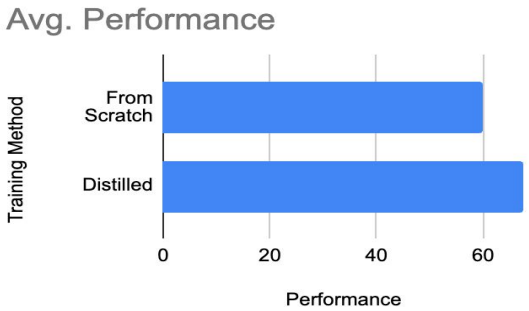
- For sample y from human data or teacher, generate teacher probabilities
  - Either using a live teacher during student training (“online”) or offline annotations
- Minimize next-token prediction loss using these “soft” probability labels

Supervised KD During LLM Pre-Training

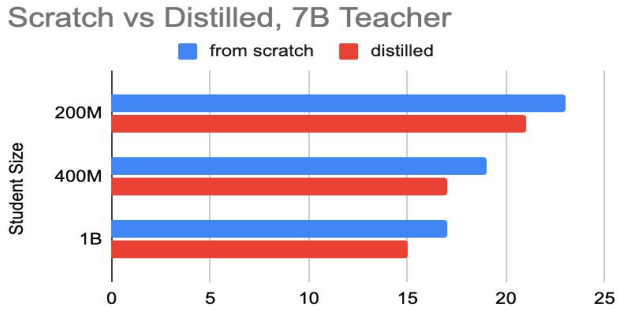
- For each next token in pre-training data, generate teacher probabilities q(. | context)
  - Either using a live teacher during student training (“online”) or offline annotations
- Minimize next-token prediction loss using these “soft” probability labels

Notable examples:
- DistillBert
- Gemma 2

Gemma 2: 7B Teacher → 2B Student



Gemma 2: Perplexity (Lower is Better)

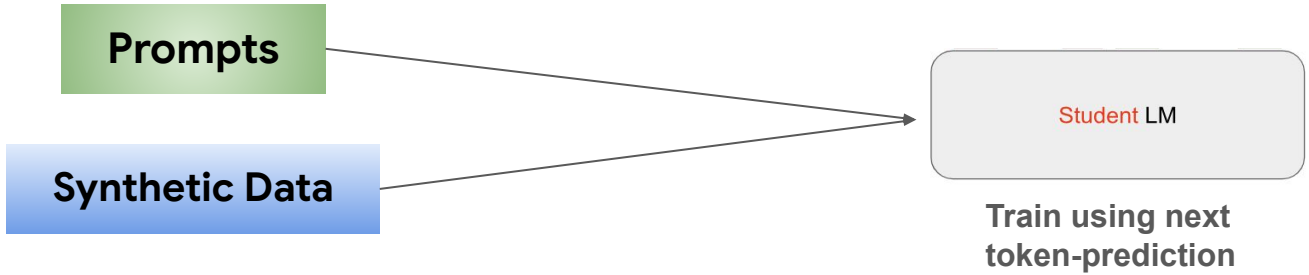


Distillation Using “Synthetic Data”

Synthetic Data Generation

e.g., What is knowledge distillation?

Supervised Fine-Tuning (SFT)



Initially popularized by Sequence-Level Knowledge Distillation (EMNLP, 2016).

“Synthetic Data” Distillation is Prevalent

MSN

DeepSeek and distillation: Why the AI race will never be the same

CNBC's Deirdre Bosa joins 'The Exchange' to discuss what DeepSeek's arrival means for the AI race.

WSJ WSJ

Why ‘Distillation’ Has Become the Scariest Word for AI Companies

DeepSeek's success learning from bigger AI models raises questions about the billions being spent on the most advanced technology.

Forbes

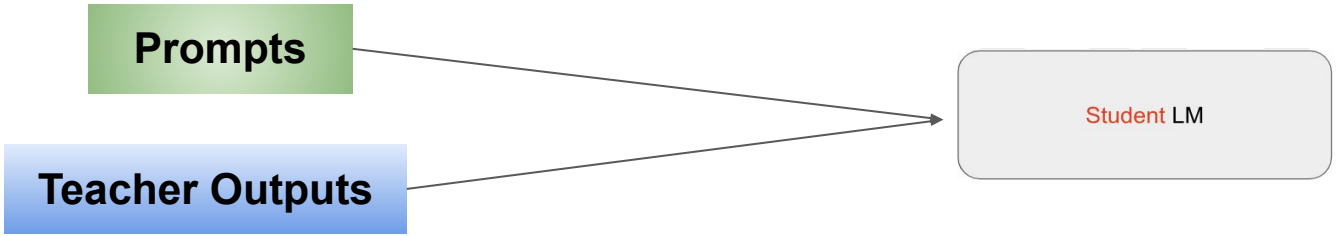
Here’s How Big LLMs Teach Smaller AI Models Via Leveraging Knowledge Distillation

AI-driven knowledge distillation is gaining attention. LLMs are teaching SLMs. Expect this trend to increase. Here's the insider scoop.

Black-box distillation: Only-need API access!

Synthetic Data Distillation Also Minimizes Forward KL!

$$\min_{\theta} KL\big(q(x)||p_\theta(x)\big) = \min_{\theta} \mathbb{E}_{\underbrace{y \sim q(y|x)}_{\text{Teacher Outputs}}} \overbrace{-\log p_\theta(y|x)}^{\text{Next token Prediction}}$$

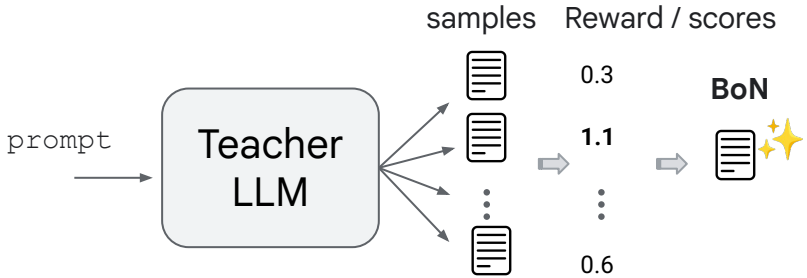


“Synthetic Data” Using Best-of-N (aka Rejection Sampling)

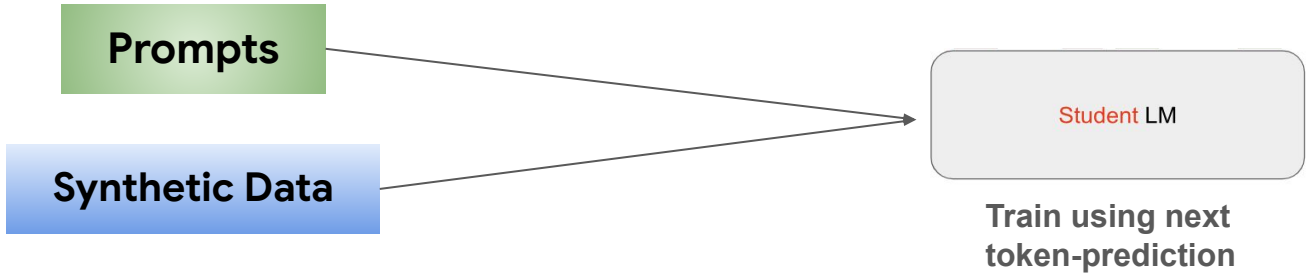
Synthetic Data Generation

Best-of-N sampling from the teacher

1. **Sample** N model generations,
2. **Score** them,
3. Keep the one with **highest score**

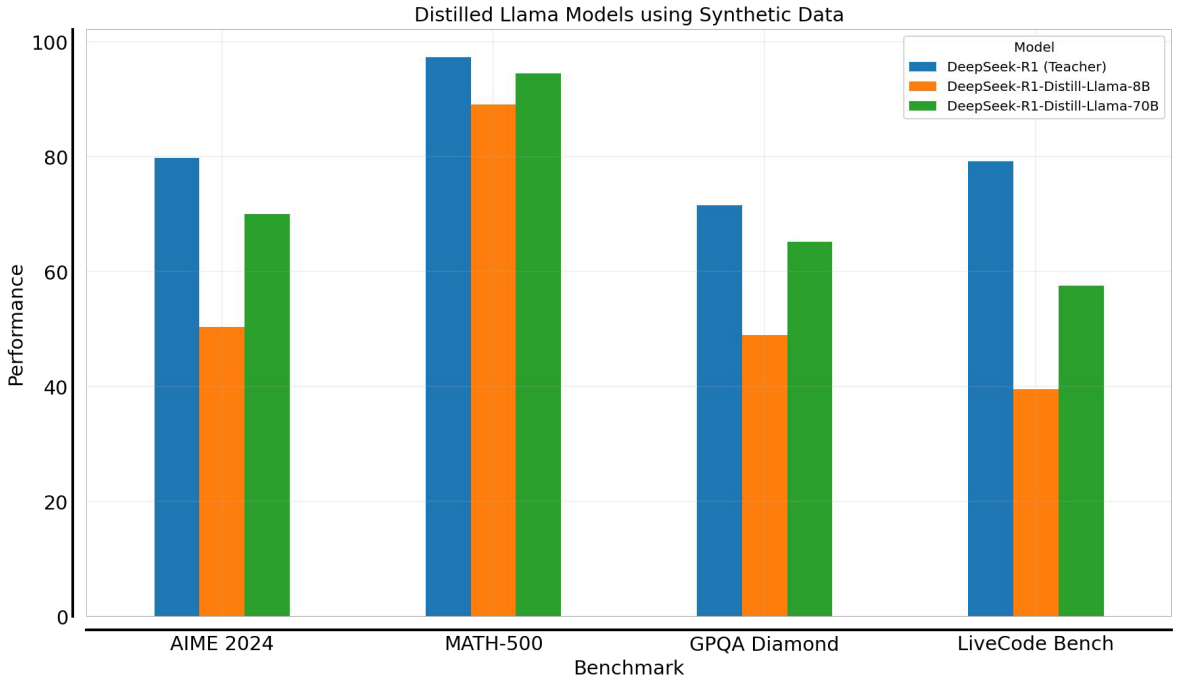


Supervised Fine-Tuning



See STaR / ReST^{EM} / Best-of-N-Distillation papers.

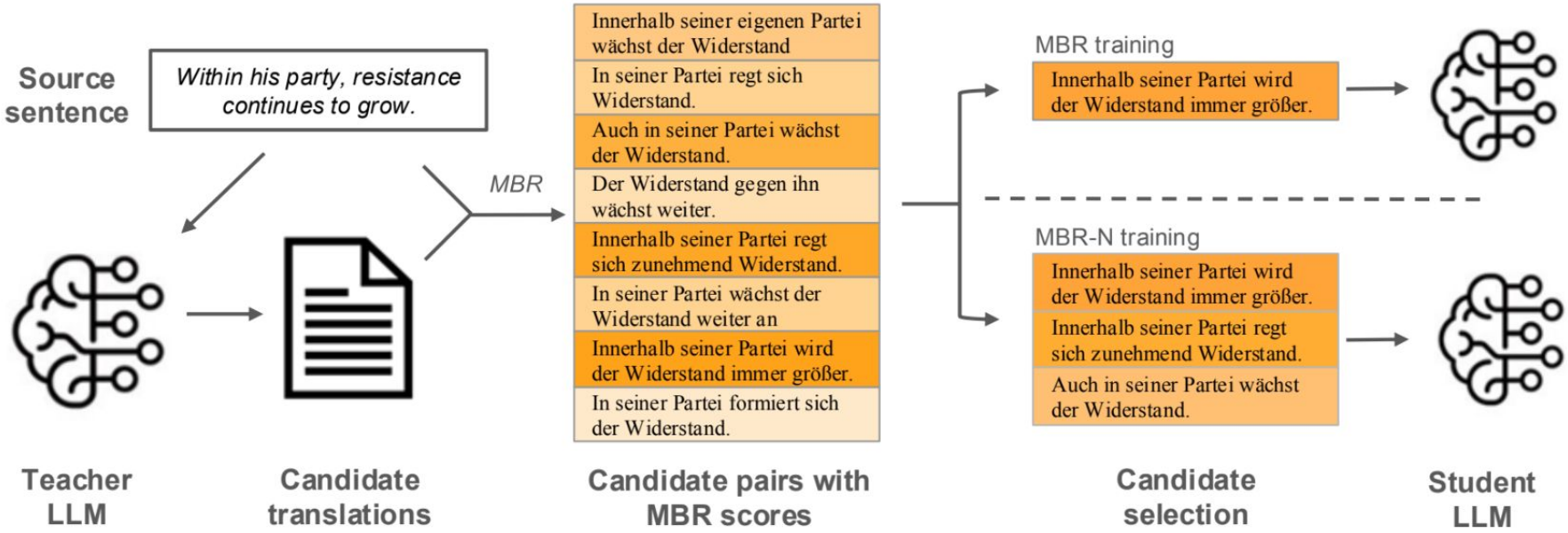
“Best-of-N” Distillation for Reasoning



Teacher and Student Vocabulary mismatch doesn’t matter!

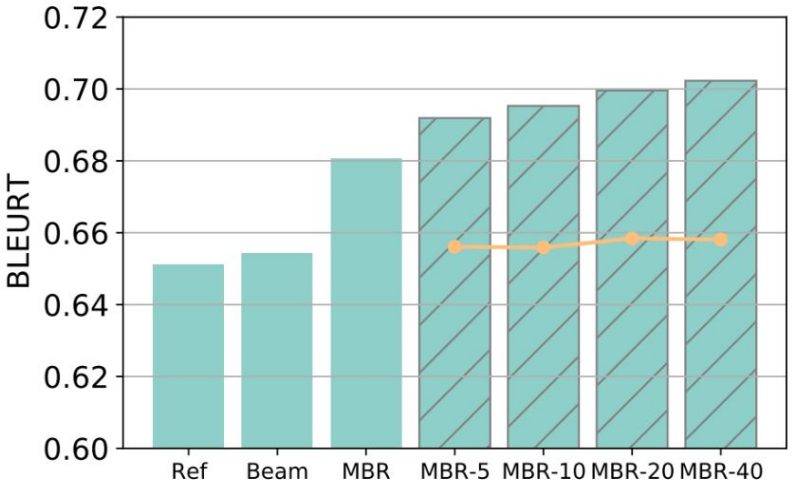
Results sourced from https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf

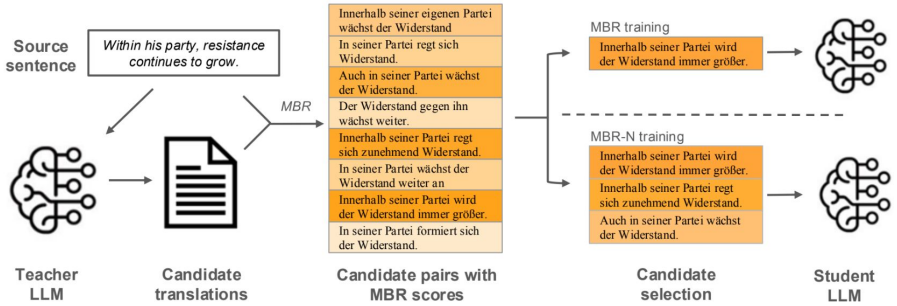
Revisiting “Best-of-N” Distillation



Don’t Throw Away Data: Better Sequence Knowledge Distillation, arxiv 2024

Revisiting “Best-of-N” Distillation





Don’t Throw Away Data: Better Sequence Knowledge Distillation, arxiv 2024

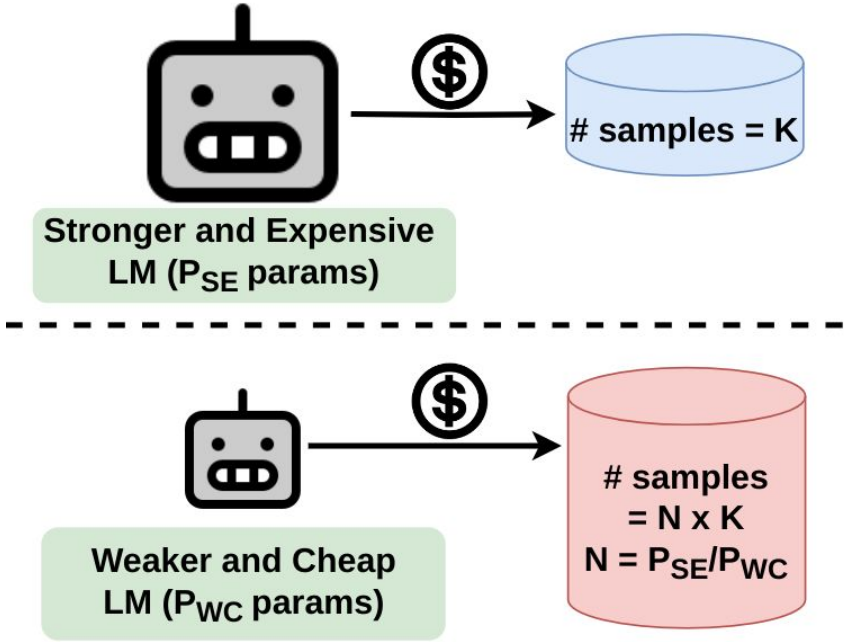
Compute-Matched Sampling for Synthetic Data Distillation

For autoregressive LLMs,

Sampling cost (FLOPs) $\propto$ ND

N is the number of model parameters

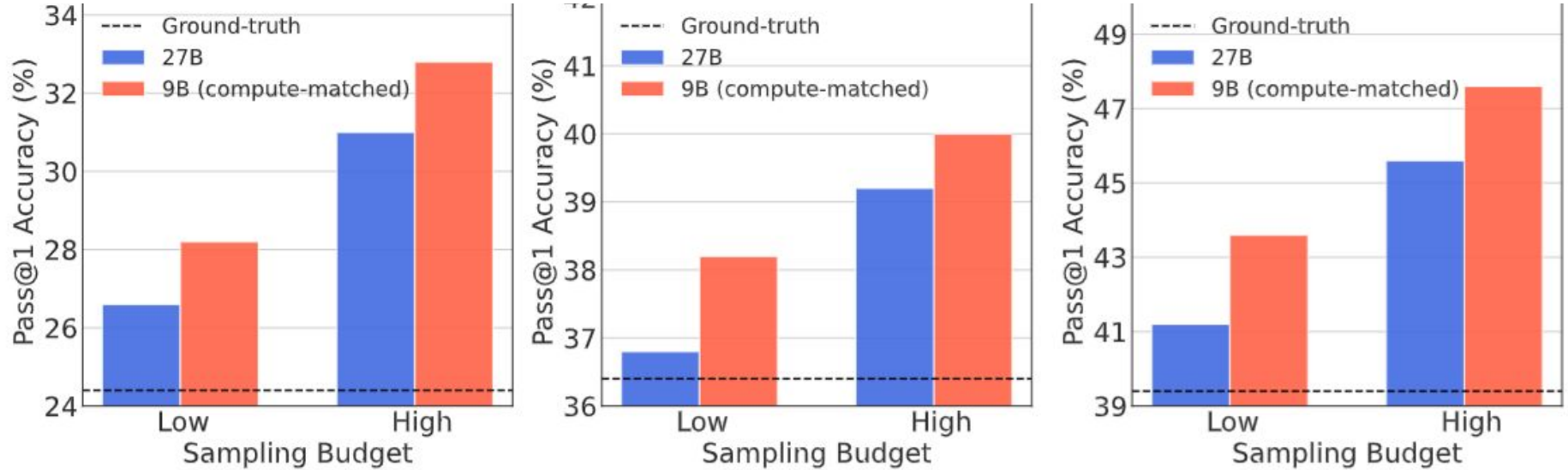
D is the number of inference tokens.



Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling. ICLR 2025.

Compute-Matched Sampling Can Be Better

Weak-to-Strong Distillation



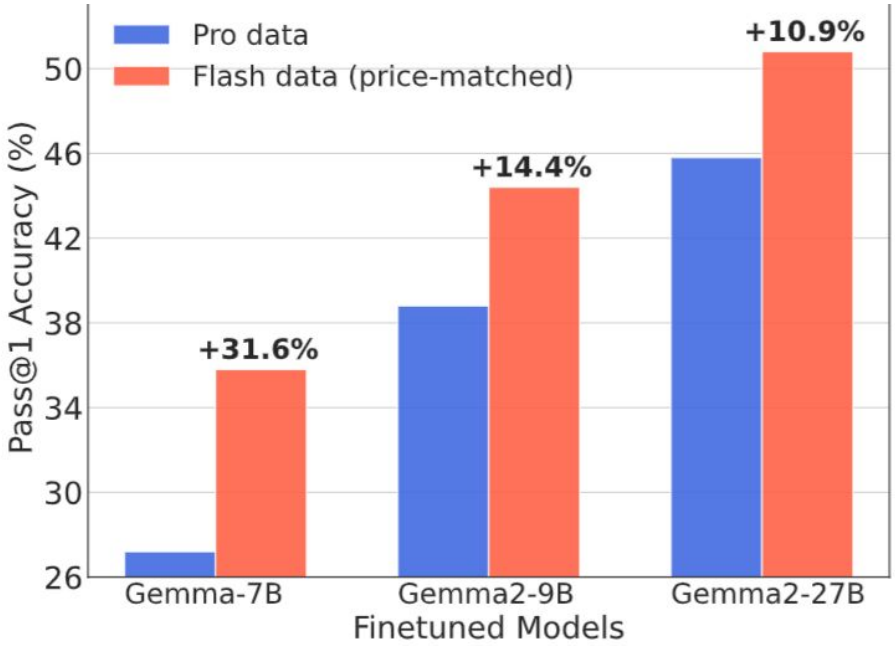
(a) Finetuning Gemma-7B. (b) Finetuning Gemma2-9B. (c) Finetuning Gemma2-27B.

Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling. ICLR 2025.

Cost-Matched Sampling Can Be Even Better!

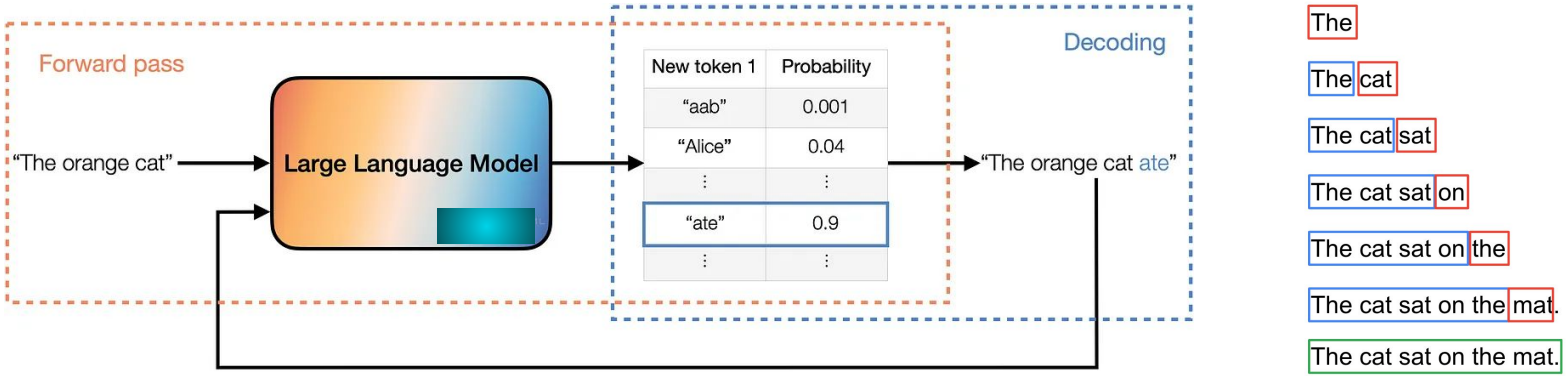
Price of Gemini 1.5 Flash =
35x Price of Gemini 1.5 Pro

Knowledge distillation:
Gemma-7B, 9B, and 27B



Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling. ICLR 2025.

Distillation for LLMs: Train-Inference Mismatch

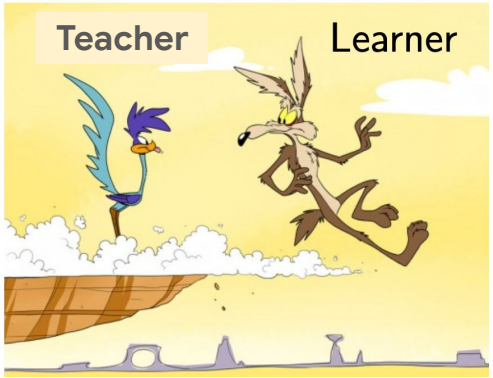


Supervised KD / SeqKD result in a train-inference mismatch[1] for LLMs. (Same motivation for SFT vs RLHF)

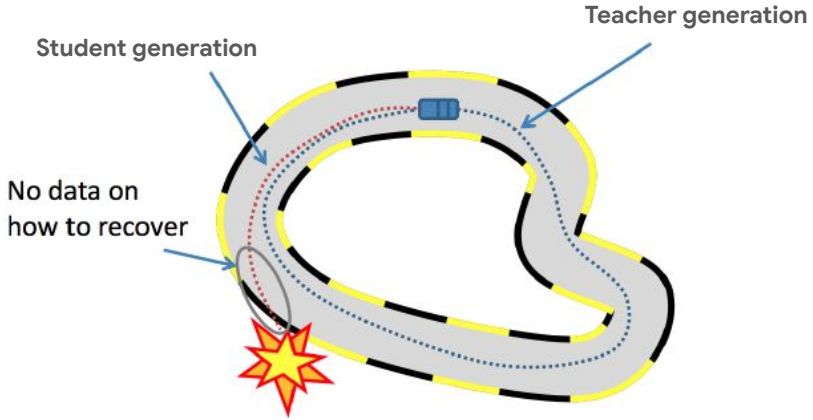
- The student is trained on a **fixed dataset** of ground-truth outputs (human or teacher samples).
- At inference time, the student generates sequences auto-regressively, where each new token depends on the previously generated tokens.

1. Also known as exposure bias.

Distillation for LLMs: Train-Inference Mismatch



Source: Cornell Lecture slides on DAgger



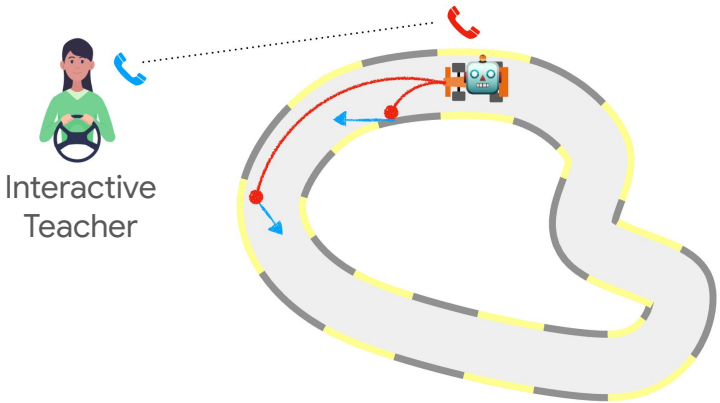
Source: CMU Lecture slides on Imitation Learning

Supervised KD or SeqKD result in a train-inference distribution mismatch for LLMs.

- The student might generate low-quality tokens that can affect future tokens, resulting in poor quality generations.

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

On-Policy Distillation: Learning From Self-Generated Mistakes



Train / Test distribution = Student's distribution

Used for Gemma 2 post-training.

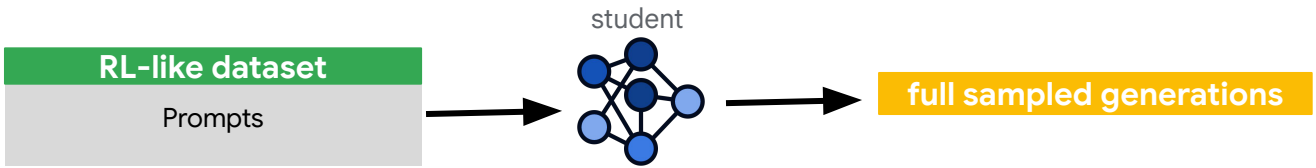
1. **On-policy Data:** Sample output sequences from the student
2. **Feedback:** Run inference using the teacher to get logits on student samples
3. **Supervised Training**: Minimize mismatch (e.g., KL-divergence) between student and teacher *token-level logits*.

On-Policy Distillation of Language Models, ICLR 2024. Also, ImitKD.

On-Policy Distillation (GKD)

1) **On-policy Data:** sample output sequences **from the student model.**

Slide From: Nino Veillard

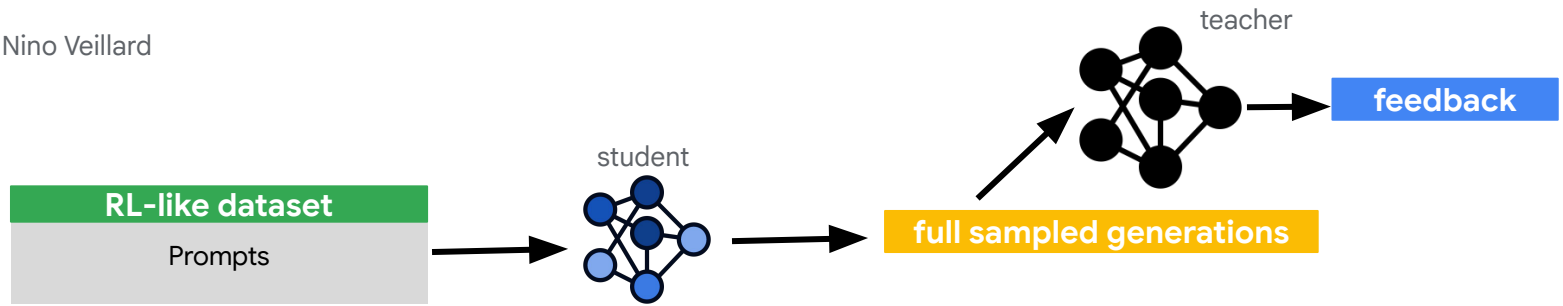


On-Policy Distillation (GKD)

1) **On-policy Data:** sample output sequences **from the student model.**

2) **Teacher feedback:** **get teacher logits on student generated sequences.**

Slide From: Nino Veillard



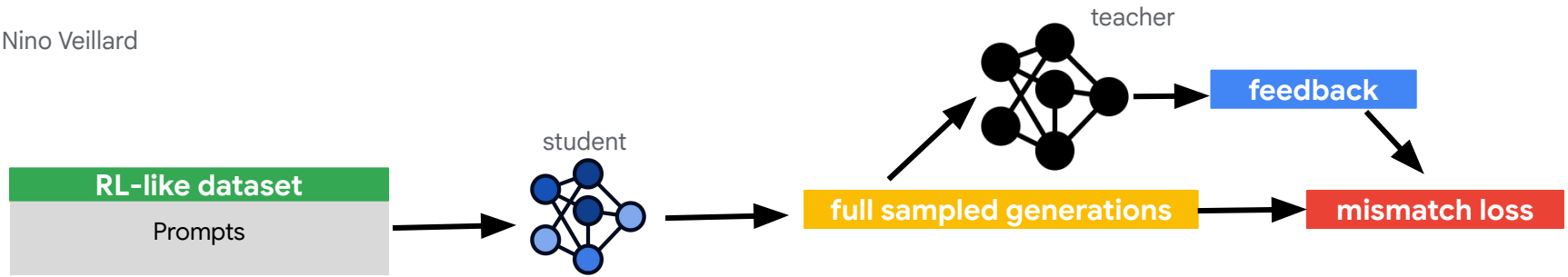
On-Policy Distillation (GKD)

1) **On-policy Data:** sample output sequences **from the student model.**

2) **Teacher feedback:** **get teacher logits on student generated sequences.**

3) **Distribution matching:** **minimize loss** between student and teacher *token-level logits*.

Slide From: Nino Veillard



Minimizing Reverse KL Leads To On-Policy Distillation

$$\min_\theta D_{KL}\big(\overbrace{p_\theta(x)}^{\textit{Student}} \,||\, \overbrace{q(x)}^{\text{Teacher}}\big)$$

$$= \min_\theta \mathbb{E}_{\underbrace{y \sim p_\theta(x)}_{\text{On-policy samples}}} \left[\sum_i \underbrace{D_{KL}\big(p_\theta(\cdot|x, y_{1:i-1}) \,||\, q(\cdot|x, y_{1:i-1})\big)}_{\text{Per-Token Reverse KL}} \right]$$

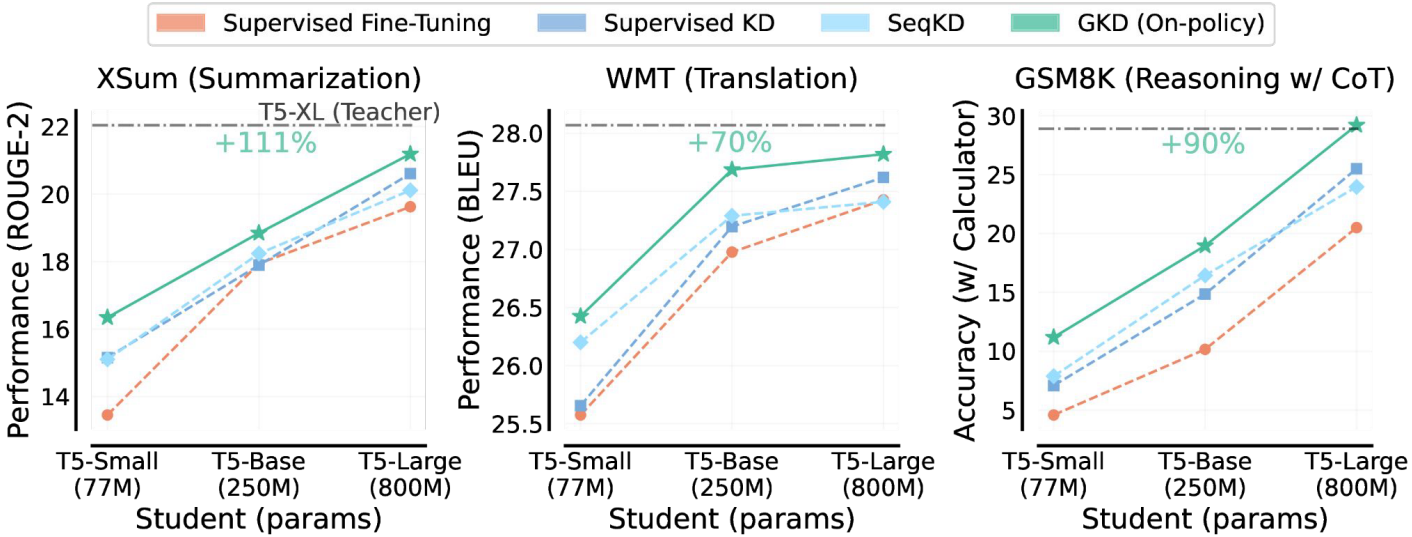
Note: We do not backpropagate through the student's sampling distribution, similar to DAgger.

GKD generalizes this to other f-divergences (JSD, Jeffreys)

1. **On-policy Data:** Sample output sequences from the student
2. **Feedback:** Run inference using the teacher to get logits on student samples
3. **Supervised Training**: Minimize mismatch (e.g., KL-divergence) between student and teacher *token-level logits*.

On-Policy Distillation of Language Models, ICLR 2024.

On-Policy Distillation Works Well in Practice



On-Policy GKD **consistently improves** over common distillation approaches (SFT, SeqKD, Supervised KD) on T5 models and "academic" benchmarks.

On-Policy Distillation of Language Models, ICLR 2024.

On-Policy Distillation for Long Reasoning Models

Table 21: Comparison of reinforcement learning and on-policy distillation on Qwen3-8B. Numbers in parentheses indicate pass@64 scores.

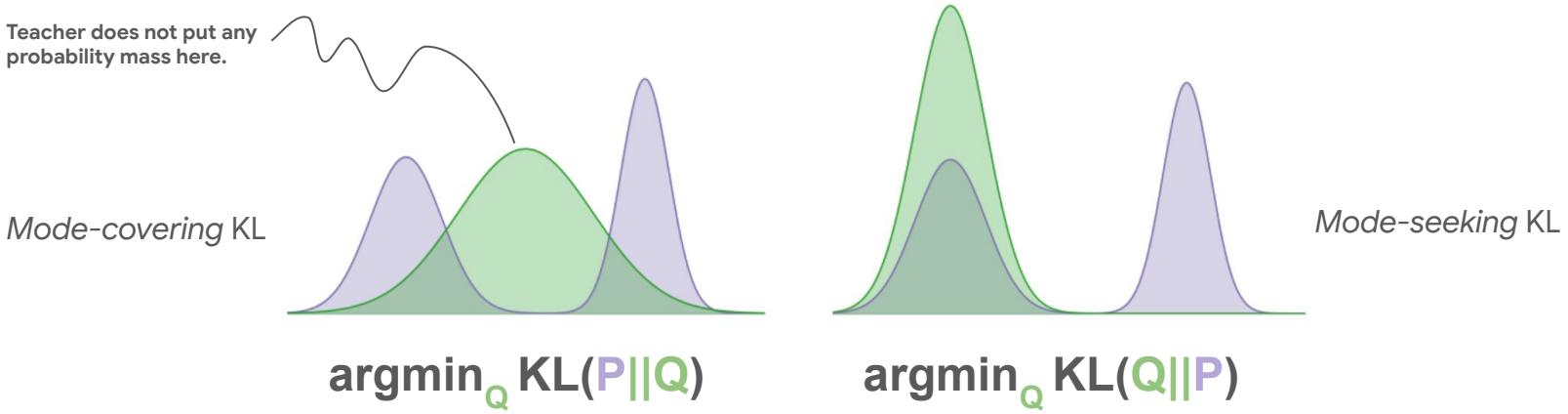
Method	AIME'24	AIME'25	MATH500	LiveCodeBench v5	MMLU -Redux	GPQA -Diamond	GPU Hours
Off-policy Distillation	55.0 (90.0)	42.8 (83.3)	92.4	42.0	86.4	55.6	-
+ Reinforcement Learning	67.6 (90.0)	55.5 (83.3)	94.8	52.9	86.9	61.3	17,920
+ On-policy Distillation	**74.4 (93.3)**	**65.5 (86.7)**	**97.0**	**60.3**	**88.3**	**63.3**	1,800

On-policy distillation is **10x more compute efficient** than reinforcement learning while being more performant.

Qwen3-8B performs comparably to Deepseek-R1.

Qwen3 Technical Report, Arxiv 2025.

Distillation for LLMs: Model Capacity Mismatch



Fitting a bimodal distribution using a unimodal Gaussian distribution. Source: This blog, which adapted images this blog.

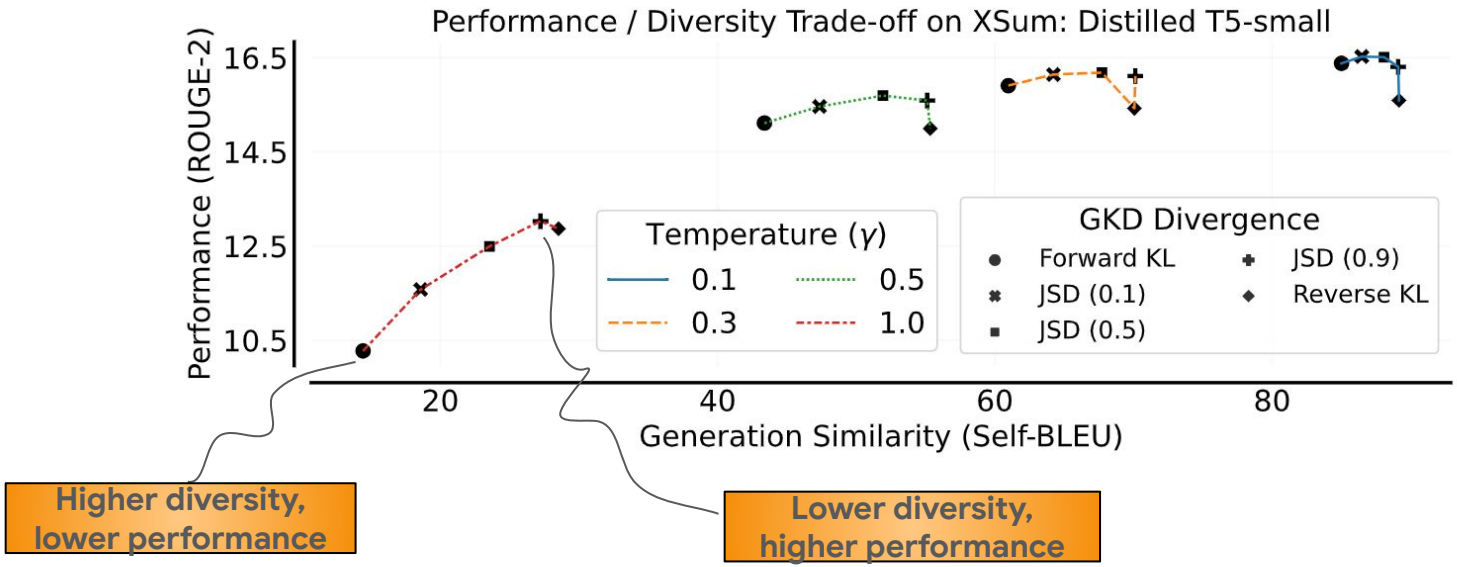
Supervised KD and SeqKD minimizes **KL**(Teacher || Student), which is *mode-covering*.

On-Policy Distillation minimizes **KL**(Student || Teacher), which is *mode-seeking*.

If student is not expressive enough to fit the teacher's distribution, mode covering distillation can lead to *unnatural* student samples.

On-Policy Distillation of Language Models, ICLR 2024.

What Divergence To Use For On-Policy GKD?

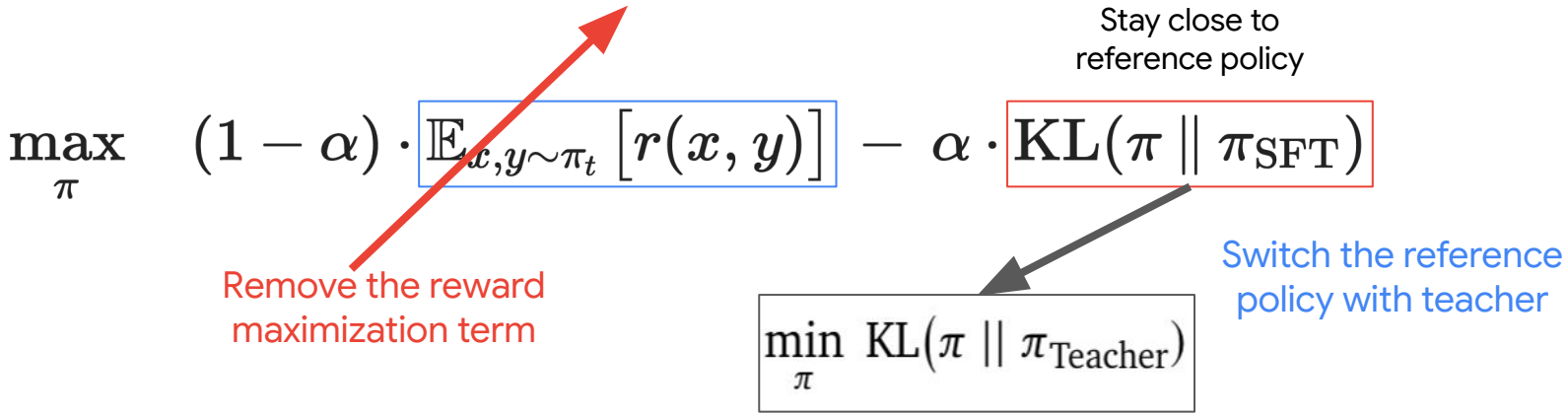


Combining mode-seeking and mode-covering divergences generally lead to best of the both worlds!

Also, see this paper which recommends Jeffreys' divergence: 0.5 * Forward KL + 0.5 * Backward KL

On-Policy Distillation of Language Models, ICLR 2024.

Implement On-Policy GKD in 3 easy steps



Step 1. *Go to your favorite RLxF framework*

Step 2. *Turn off the reward maximization term.*

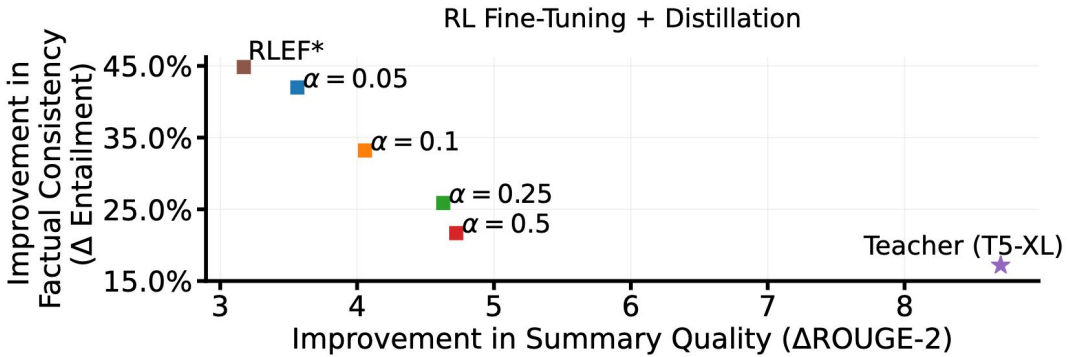
Step 3. *Switch the reference SFT policy with a teacher policy (*change partitioning).*

You have implemented on-policy GKD with reverse KL. Switch reverse KL with other ***token-level*** *f-divergences (e.g., JSD, Jeffreys) for best results.*

On-Policy GKD + RLHF: A Natural Combination

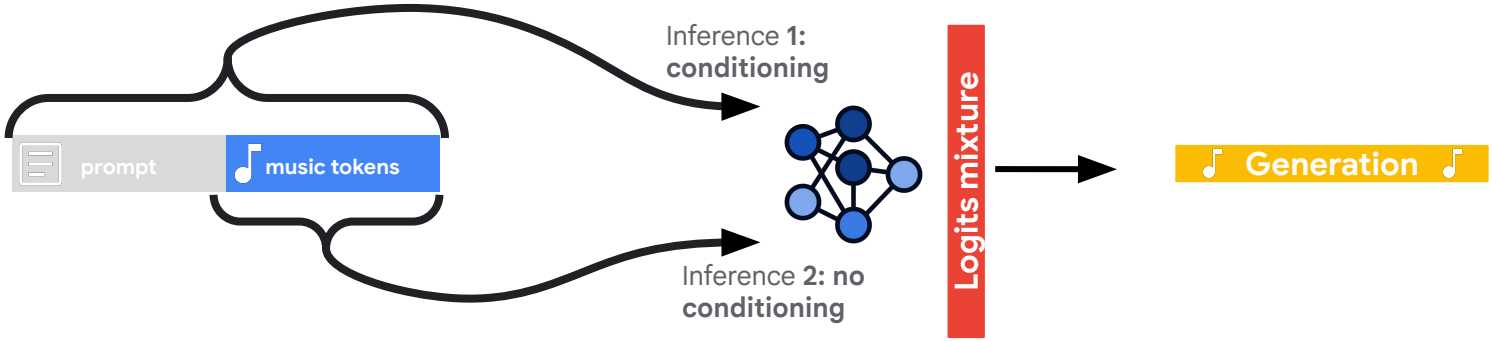
Here's a simple but powerful idea to improve RLHF / RLAIF by combining it with knowledge distillation: Simply regularize the LLM policy (student) to a more capable teacher model instead of the base student model for the KL regularization term.

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



GKD For Classifier-Free Guidance Distillation

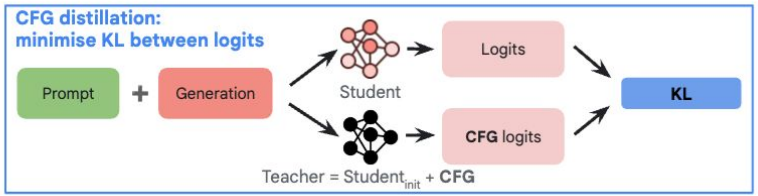
Music generation uses Classifier Free Guidance to improve text adherence



Costly, because of 2 inference runs.

GKD can be used to distill CFG in the original policy.

⟹ Same performance as CFG, without the inference cost !



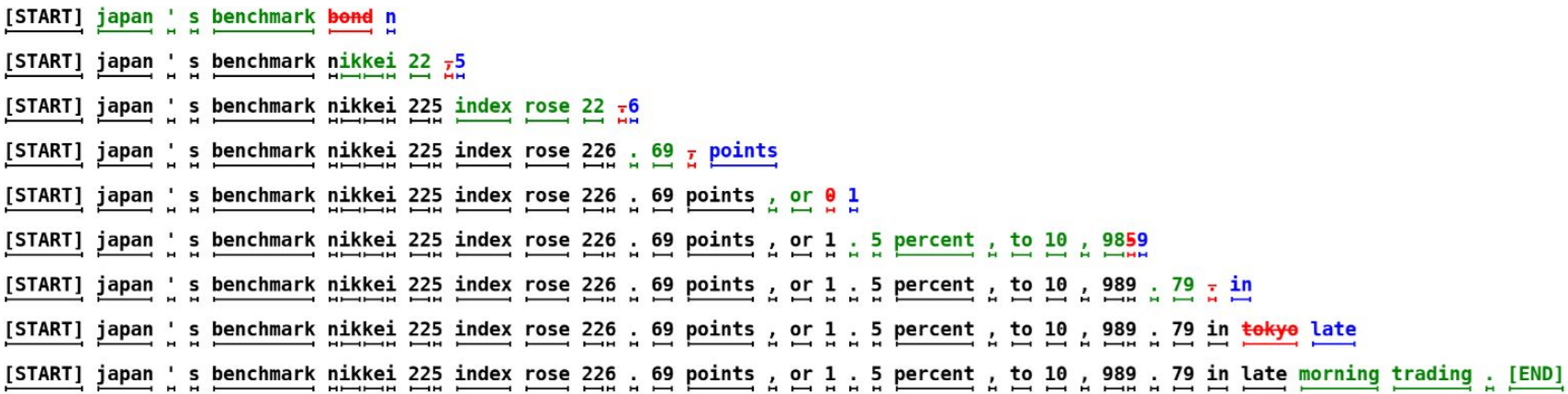
Diversity-Rewarded CFG Distillation, ICLR 2025

“Talk is cheap. Show me the code”: GKD In TRL

```python
# Apply temperature scaling
student_logits = student_logits / temperature
teacher_logits = teacher_logits / temperature

# Compute log probabilities for student and probabilities for teacher
student_log_probs = F.log_softmax(student_logits, dim=-1)
teacher_log_probs = F.log_softmax(teacher_logits, dim=-1)

# Compute the log of the mixture distribution
# log(a + b) = log(exp(log(a)) + exp(log(b))) -> for mixture
beta = torch.tensor(beta, dtype=student_log_probs.dtype)
mixture_log_probs = torch.logsumexp(
    torch.stack([student_log_probs + torch.log(beta), teacher_log_probs + torch.log(1 - beta)]),
    dim=0,
)

# Compute KL divergences using F.kl_div
# PyTorch differs from the standard mathematical definition, so the order of the probability dis
kl_teacher = F.kl_div(mixture_log_probs, teacher_log_probs, reduction="none", log_target=True)
kl_student = F.kl_div(mixture_log_probs, student_log_probs, reduction="none", log_target=True)

# Compute the Generalized Jensen-Shannon Divergence
jsd = beta * kl_teacher + (1 - beta) * kl_student
```

LLM Distillation Meets Speculative Decoding

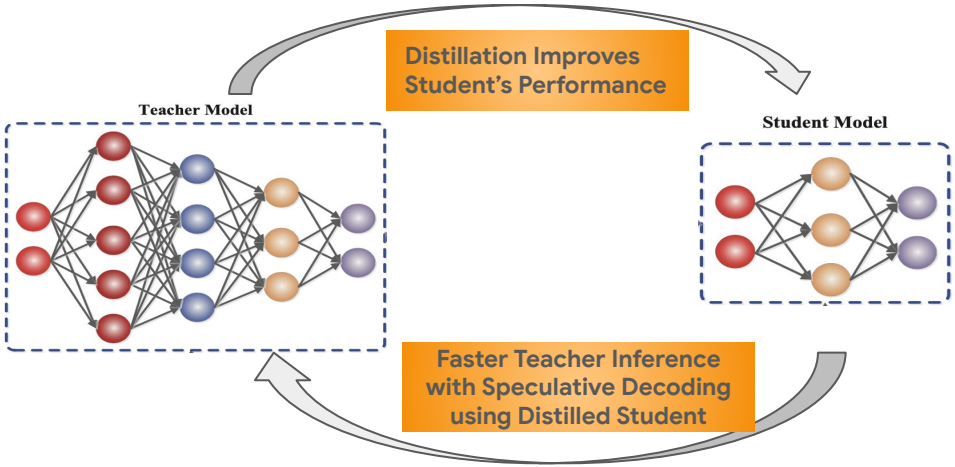
Background: Speculative Decoding



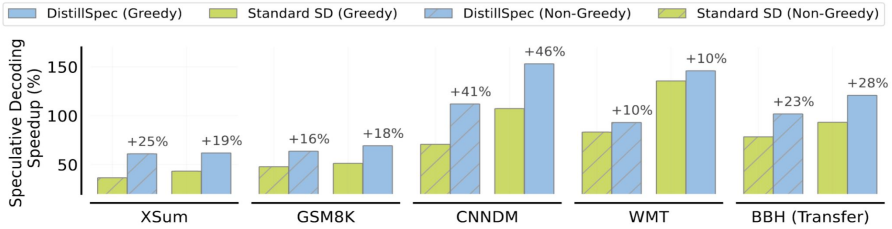
In speculative decoding, a smaller model is used as the guessing mechanism for the larger model. The accepted guesses are shown in green and the rejected suggestions in red. Source: This paper.

Fast Inference from Transformers via Speculative Decoding, ICML 2023.
Accelerating Large Language Model Decoding with Speculative Sampling, 2023

DistillSpec: GKD For Speculative Decoding



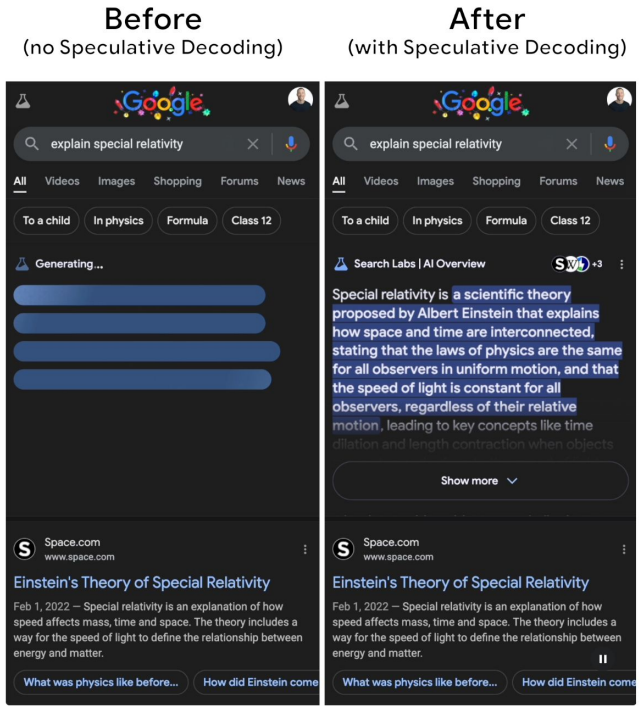
Effectiveness of Speculative Decoding depends on closeness of the student and teacher token-level distributions.



Distilled Student leads to 10–45% larger speedups for target model.

DistillSpec: Improving Speculative Decoding via Knowledge Distillation, ICLR 2024.

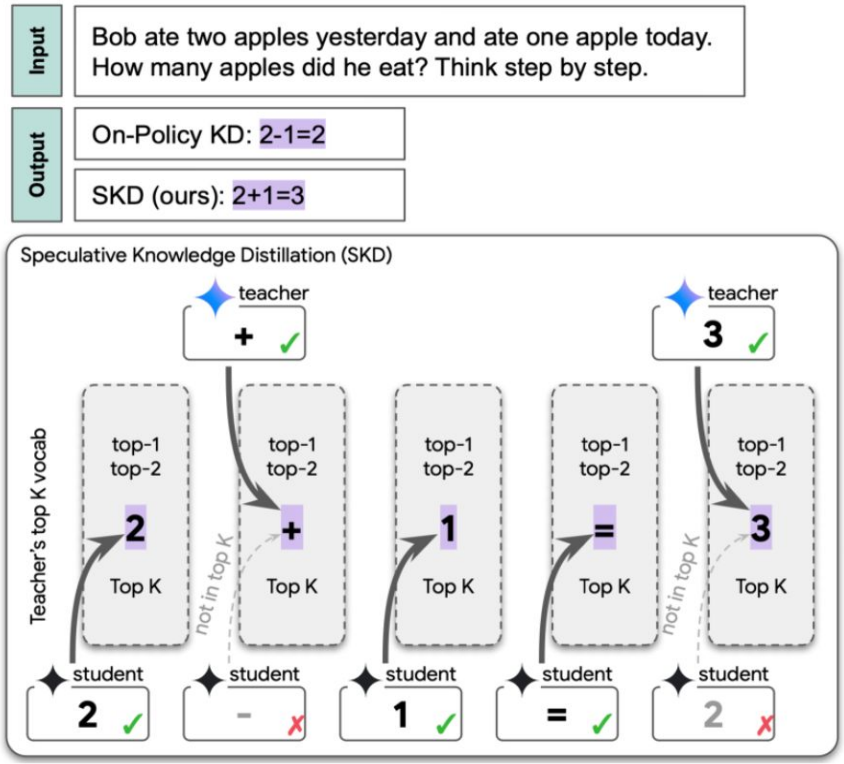
DistillSpec: Distillation For Speculative Decoding

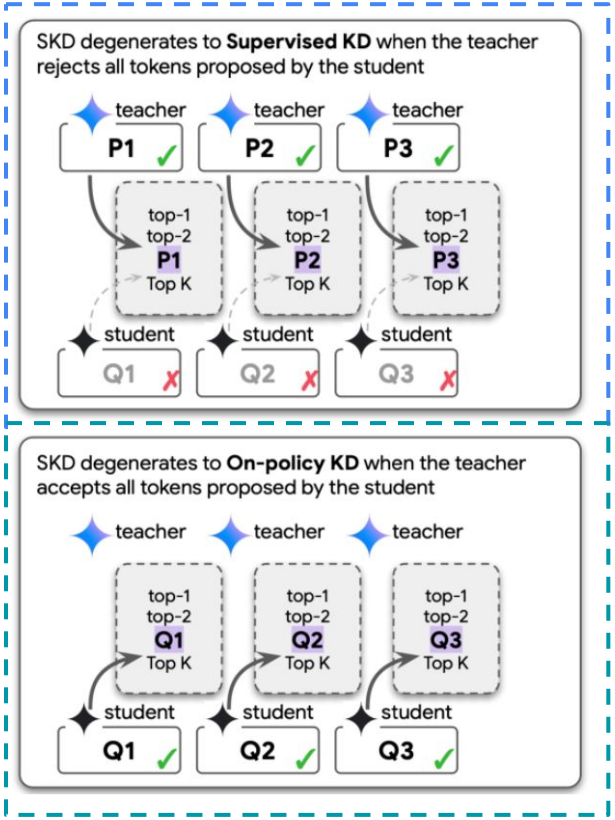


Source: [Looking back at speculative decoding]

[DistillSpec: Improving Speculative Decoding via Knowledge Distillation], ICLR 2024.

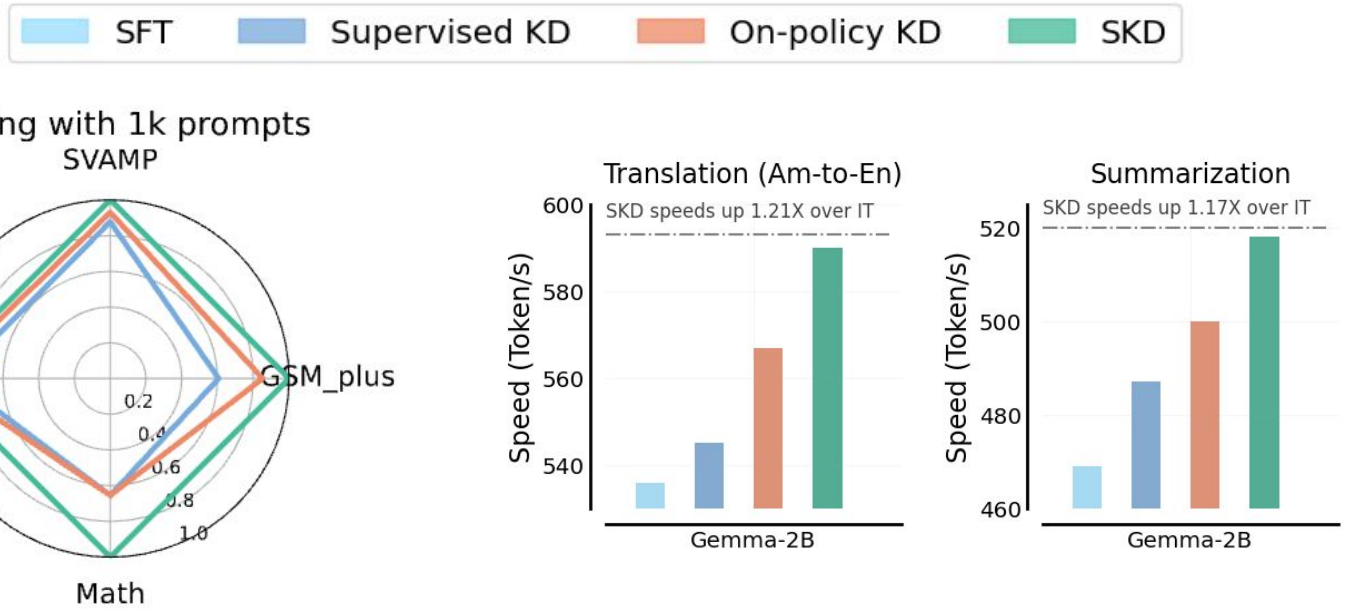
SKD: Speculative On-Policy Distillation





Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling, ICLR 2025.

SKD: Computationally Expensive But Better?



Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling, ICLR 2025.

On-Policy vs Offline “Soft” Distillation

	Online Distillation	Offline Distillation
Offline Logit Annotations	No	Yes
Compute Efficient / Sample Efficient	No / Yes	Yes / No
Live Teacher (Inference Only)	Yes	No
Student Sampling (Autoregressive)	Yes	No
Optimality on Agent / Long Horizon Tasks	Yes	Theoretically suboptimal
Train-Test Distribution Mismatch	No	Yes