Knowledge Distillation in LLMs

Tanmoy Chakraborty
Associate Professor, IIT Delhi
https://tanmoychak.com/

Advances in Large Language Models

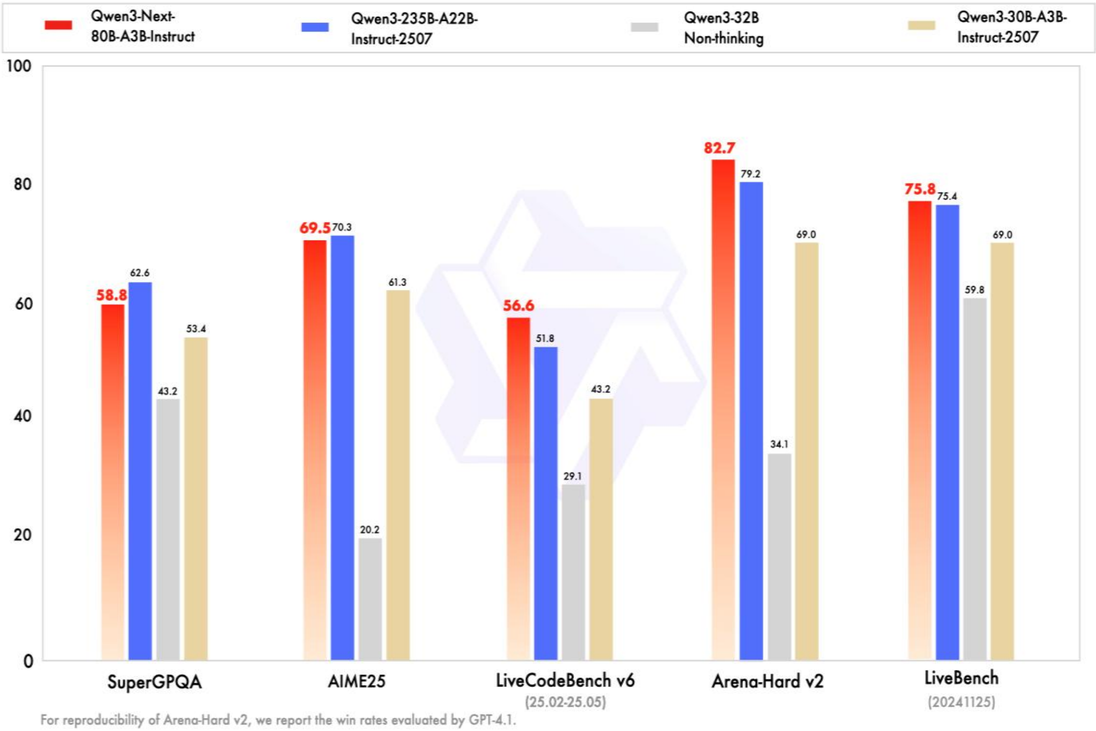
Qwen-3-Next-80B-A3B

The first in the series of next-generation foundation models that are optimized for extreme context length and large-scale parameter efficiency

Announced on September 12, 2025

[Qwen-3-Next-Blog]

Qwen-3-Next-80B-A3B introduces several architectural innovations to maximize performance while minimizing computational cost. It uses a combination of Gated DeltaNet and Gated Attention, enabling efficient context modeling for ultra-long sequences.

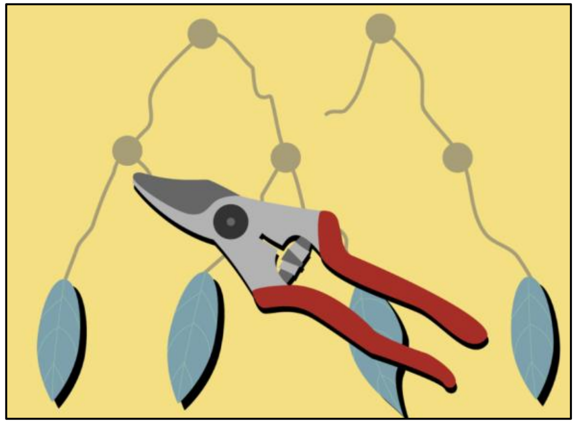


Qwen3-Next-80B-A3B uses a highly sparse MoE design, having a total of 80 billion parameters with only 3 billion activated, making it highly efficient. A thinking version is also released along with the base model.

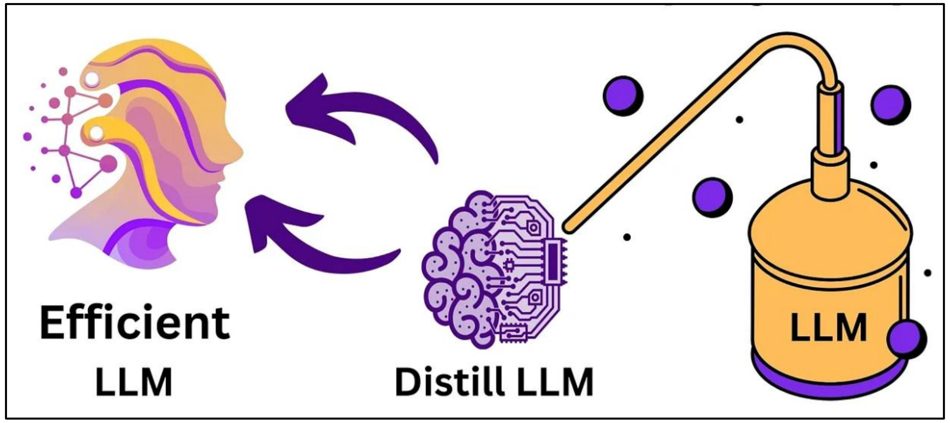
Model Compression

175 B
Tech
KD
7 B
Student
Teacher- student
Paradigm

Model Pruning



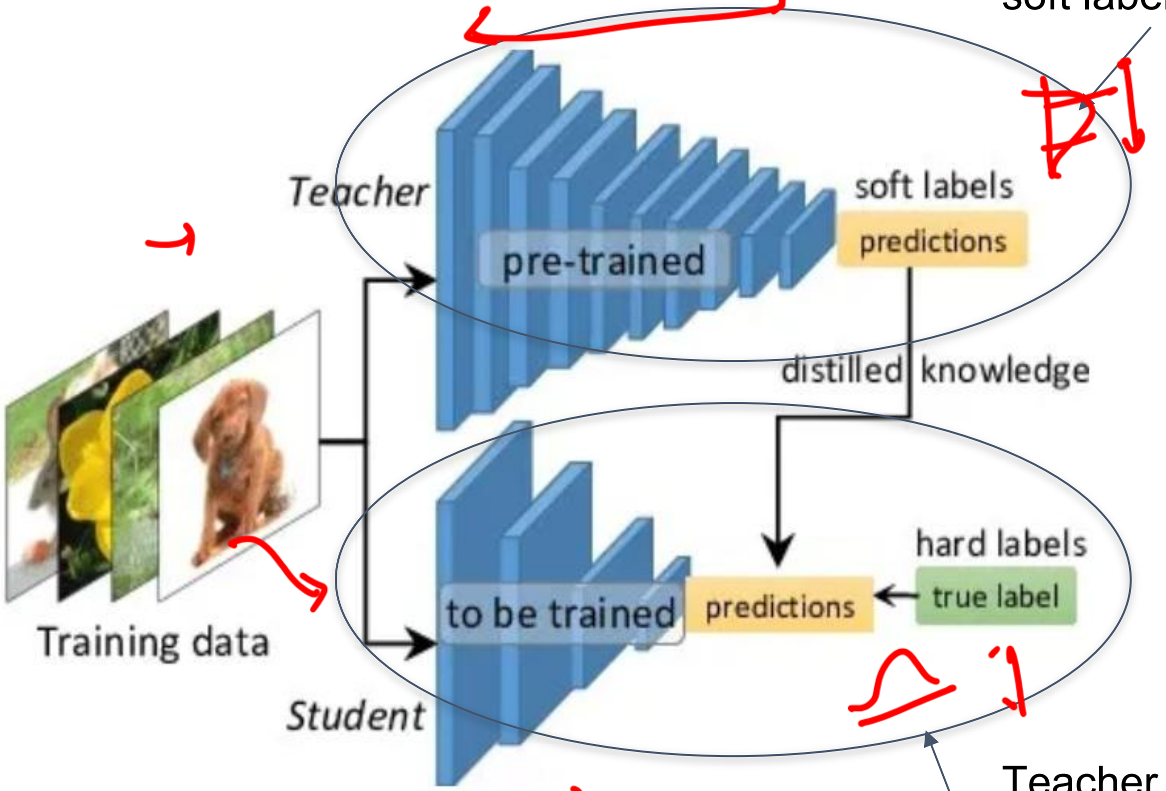
Knowledge Distillation



Knowledge Distillation (KD): Types

Hinton et al., 2015

Teacher model generate soft labels (logits)



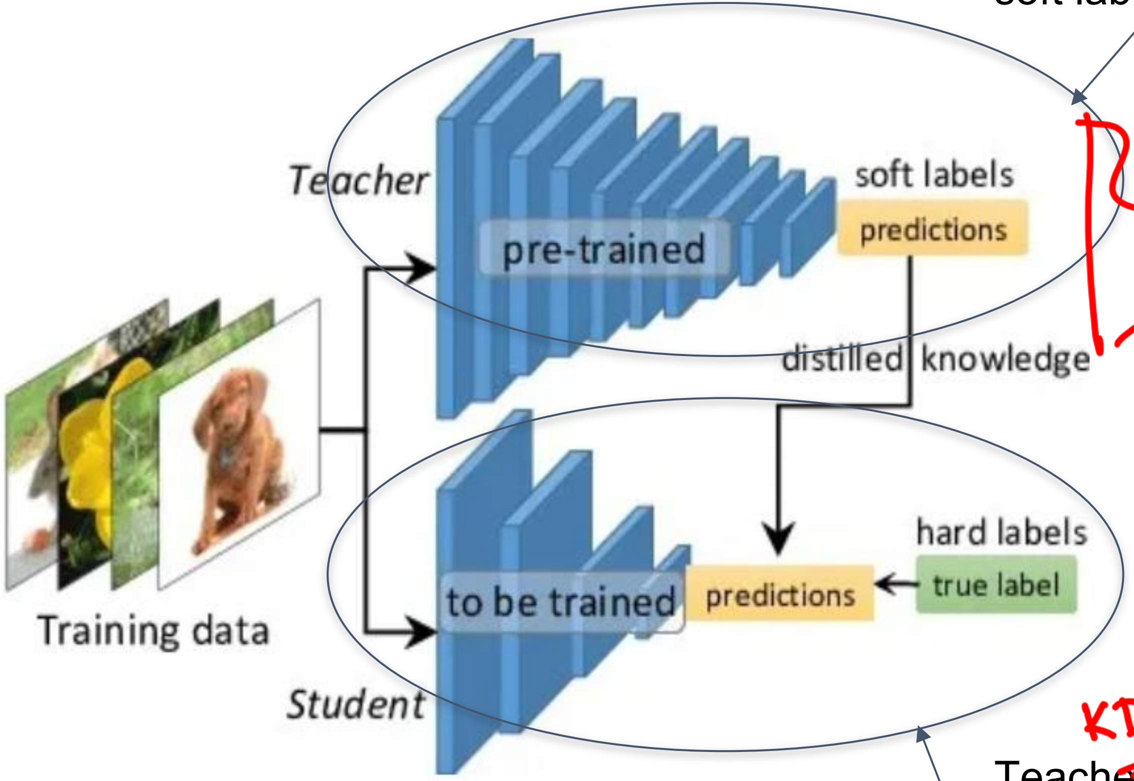
Teacher generated logits used to fine-tune the student

Loss

$$\left\| \Phi\left(\frac{\hat{y}_{(.,T)}}{\tau}\right) - \Phi\left(\frac{\hat{y}_{(.,S)}}{\tau}\right)\right\|_2$$

Knowledge Distillation (KD): Types

Hinton et al., 2015



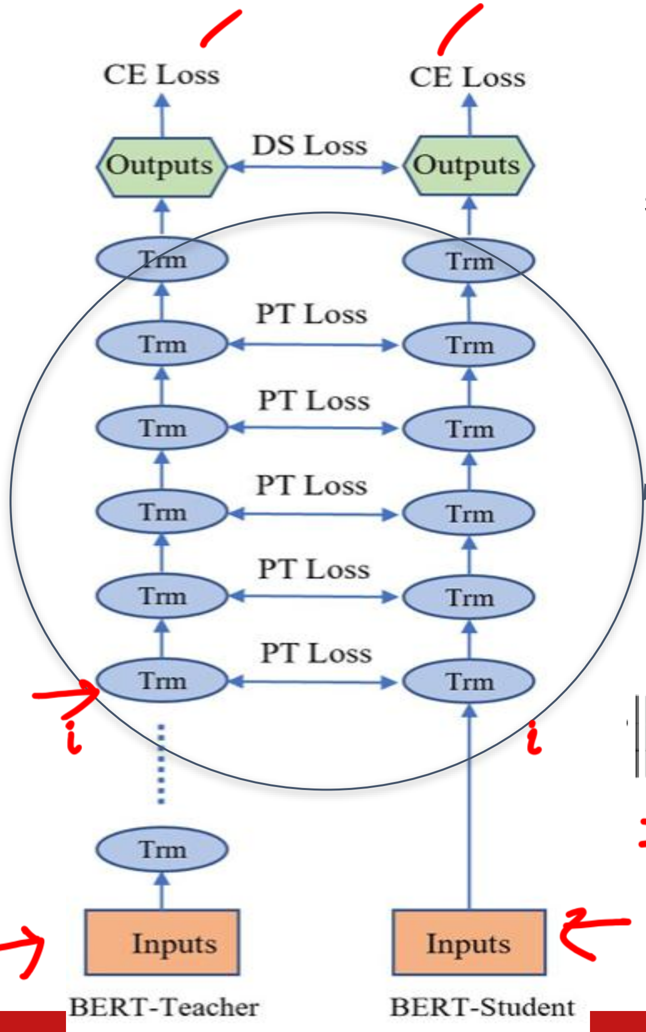
Teacher model generate soft labels (logits)

I am hungry.

KD Loss

Teacher generated logits used to fine-tune the student

$$\left\|\Phi\left(\frac{\hat{y}_{(.,T)}}{\tau}\right)-\Phi\left(\frac{\hat{y}_{(.,S)}}{\tau}\right)\right\|_2$$



Distance between layer-wise representations used to enforce student model to imitate the teacher

Layer-wise

$$\left\|h_{(.,T)}-h_{(.,S)}\right\|_2$$

Sun et al., 2019

Divergence and Similarity Functions

$p(t)$ T $q(t)$ S

Divergence Type	$D(p, q)$ Function
Forward KLD	$\sum p(t) \log \frac{p(t)}{q(t)}$
Reverse KLD	$\sum q(t) \log \frac{q(t)}{p(t)}$
JS Divergence	$\frac{1}{2}\left(\sum p(t) \log \frac{2p(t)}{p(t)+q(t)} + \sum q(t) \log \frac{2q(t)}{p(t)+q(t)}\right)$

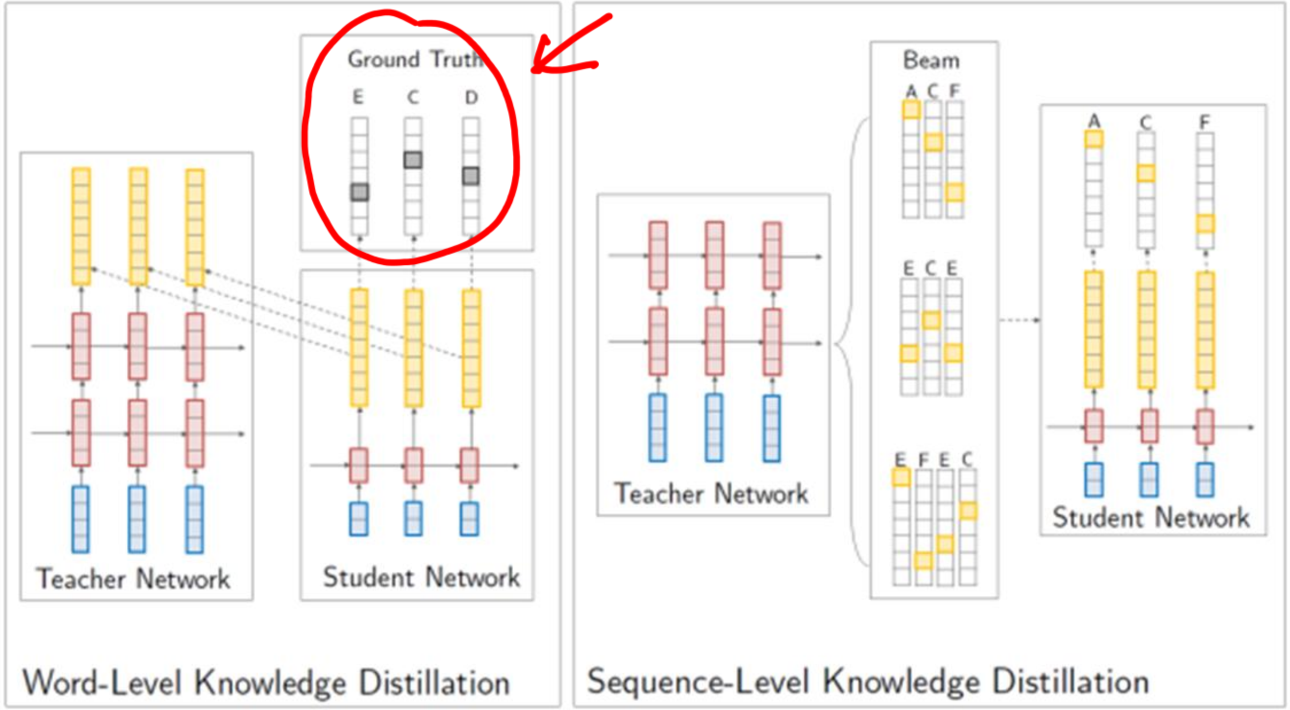
Similarity Function $\mathcal{L}_F$	Expression
L2-Norm Distance	$\lVert \Phi_T(f_T(x, y)) - \Phi_S(f_S(x, y)) \rVert_2$
L1-Norm Distance	$\lVert \Phi_T(f_T(x, y)) - \Phi_S(f_S(x, y)) \rVert_1$
Cross-Entropy Loss	$-\sum \Phi_T(f_T(x, y)) \log(\Phi_S(f_S(x, y)))$
Maximum Mean Discrepancy	$\mathrm{MMD}(\Phi_T(f_T(x, y)), \Phi_S(f_S(x, y)))$

Categories of KD

- **White-box KD**: Full access to the teacher's internal components (logits, hidden states, attention maps)
- **Meta KD**: Teacher helps guide student training strategies (e.g., data selection, curriculum)
- **Black-box KD**: Only the final output of the teacher is available, e.g., via API

KD for Language Models

Kim and Rush 2016 extended the idea to word-level and sequence-level KD for language models, which aligns the student model with the teacher's output distributions



KD for Language Models

- Applied in sequence generation tasks (e.g., machine translation)
- Student model is trained using the teacher's best decoded sequence (e.g., via beam search)

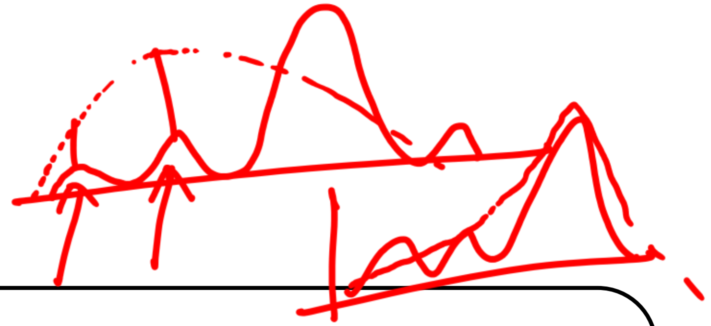
Advantages:

- Instead of label sequences, student mimics the teacher's generation process
- Better for long-form tasks like summarization or machine translation

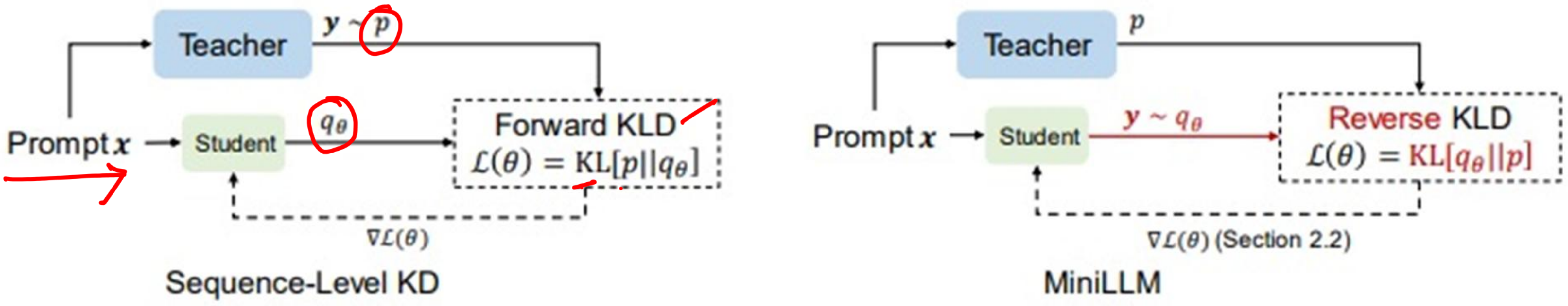
Disadvantages:

- Beam search is computationally expensive
- Generated sequences may propagate teacher's errors

KD for LLMs – MiniLLM (Gu et al. 2023)

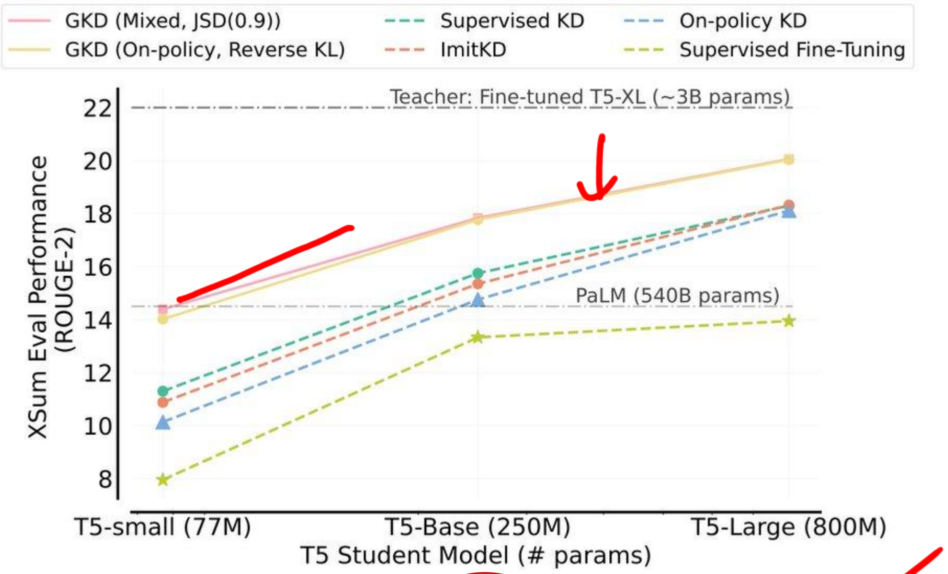


- MiniLLM (Gu et al. 2023) replace the forward Kullback-Leibler divergence (KLD) objective in the standard KD approaches with reverse KLD, which is more suitable for KD on generative language models.
- This prevents the student model from overestimating the low-probability regions of the teacher distribution.



KD for LLMs – GKD (Agarwal et al. 2024)

- Current KD methods for auto-regressive sequence models suffer from distribution mismatch between output sequences seen during training and those generated by the student during inference.
- Instead of solely relying on a fixed set of output sequences, GKD trains the student on its **self-generated output (SGO) sequences** by leveraging feedback from the teacher on such sequences.



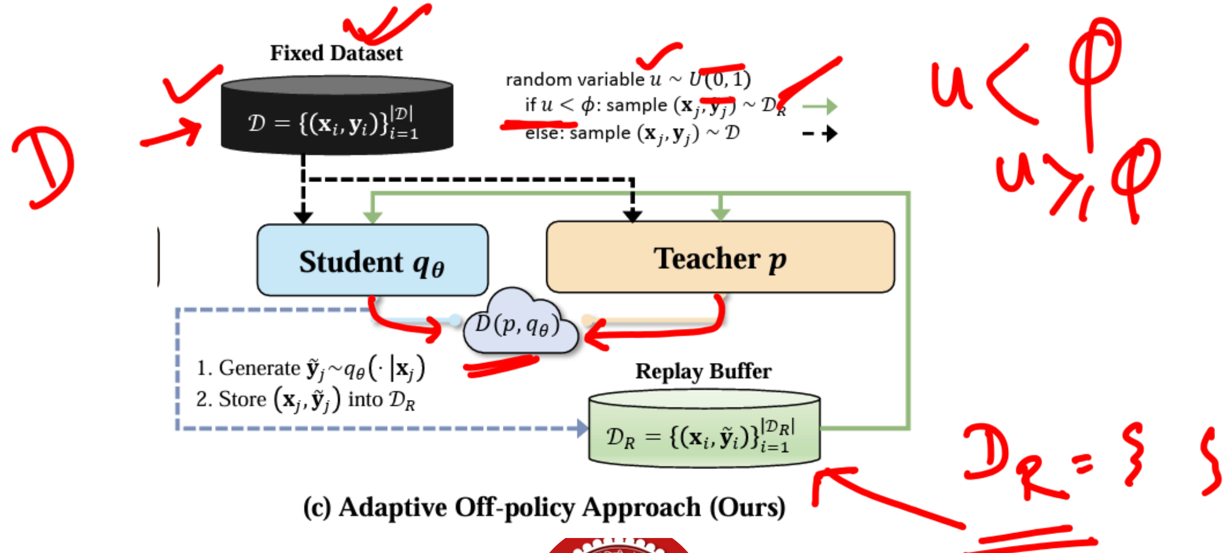
I am high

.9 .8 .3

.8 .9 .95
I am mid.

Adaptive SGO for KD – DistiLLM (Ko et al. 2024)

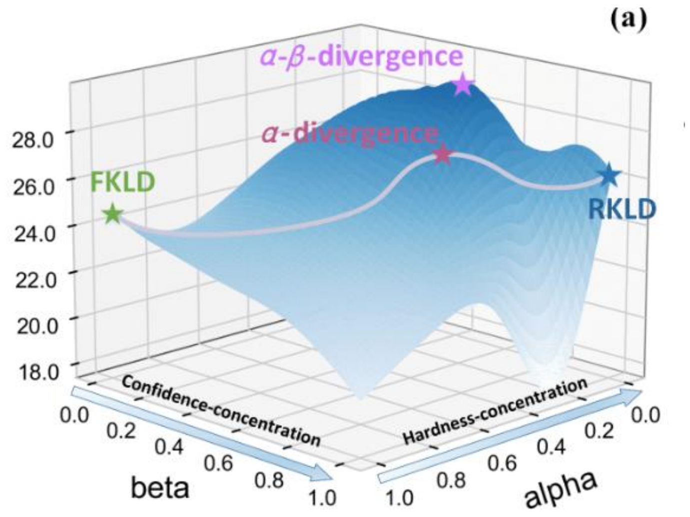
- Generating SGO for each step can increase distillation time significantly. Ko et al., suggested an adaptive method with replay buffer to adaptively determine when to generate SGO vs. when to use original ground truth texts for distilling knowledge.
- KD optimization stability depends on the smoothness of the distillation loss objective. Ko et al., suggested a skewed divergence loss, where a mixture probability of teacher and student logits is used.

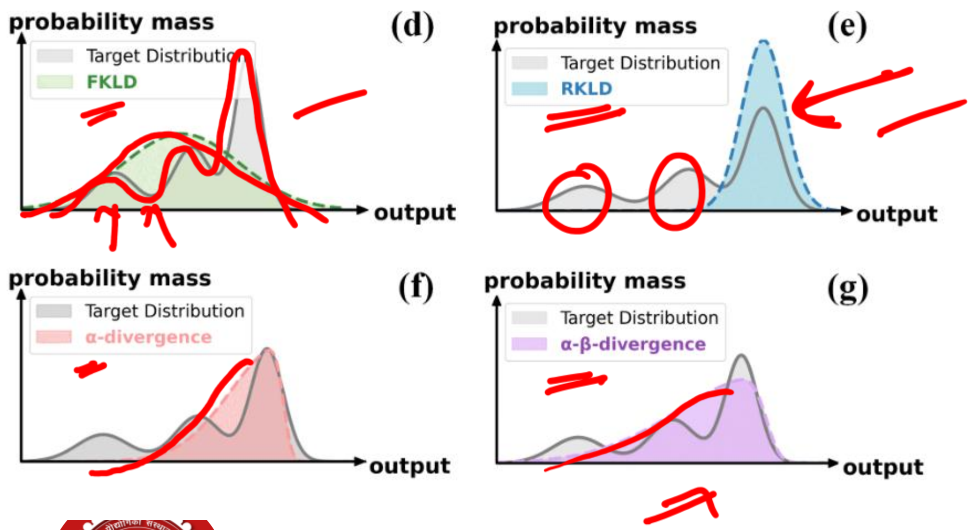


(c) Adaptive Off-policy Approach (Ours)

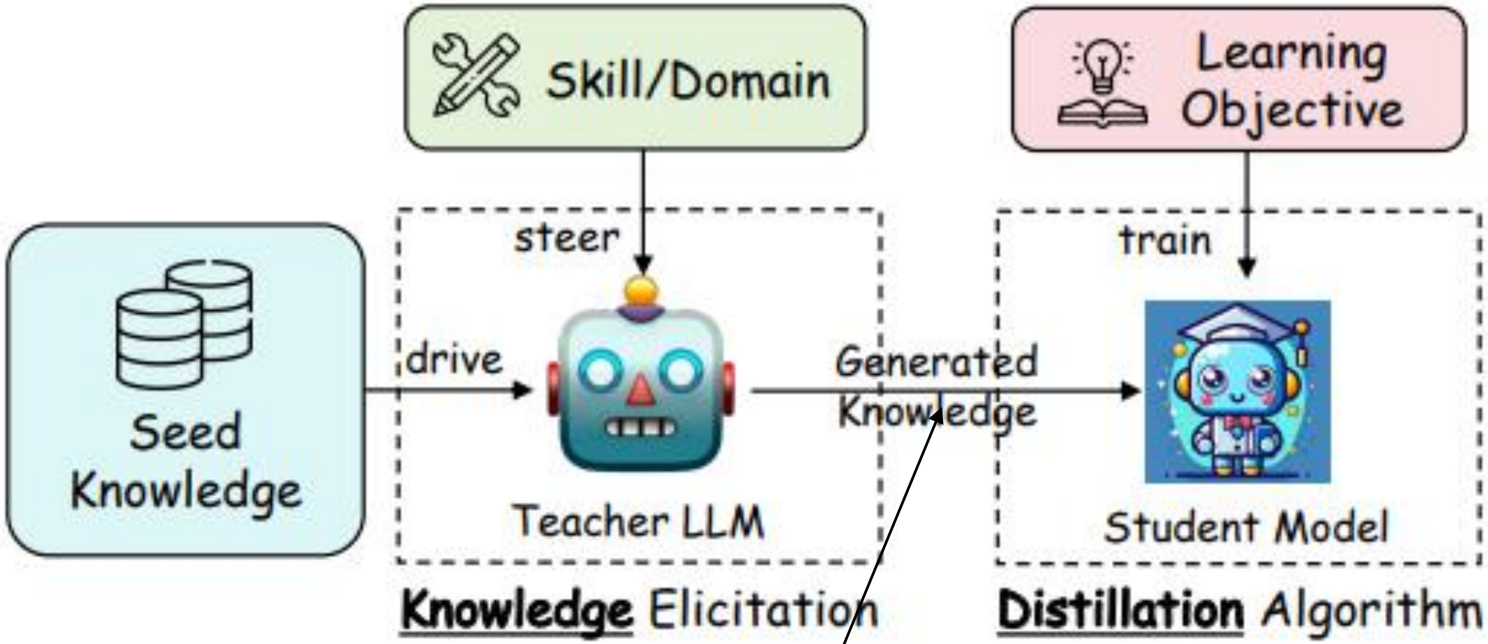
Confidence-Concentrated Loss for KD – ABKD (Wang et al. 2025)

- Traditional distillation loss functions – forward KLD and reverse KLD tackles two different properties – while FKLD makes student distribution overly smoothened (higher recall), RKLD captures prominent modes of the teacher (higher precision).
- Wang et al., proposed a weighted scheme between FKLD and RKLD, capturing the confidence and hardness of teacher-student output probabilities. ABKD is a generalized variation of the popularly used divergence-based loss functions used in KD.



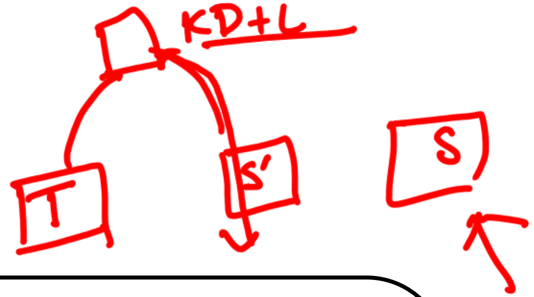


Limitations of Vanilla KD

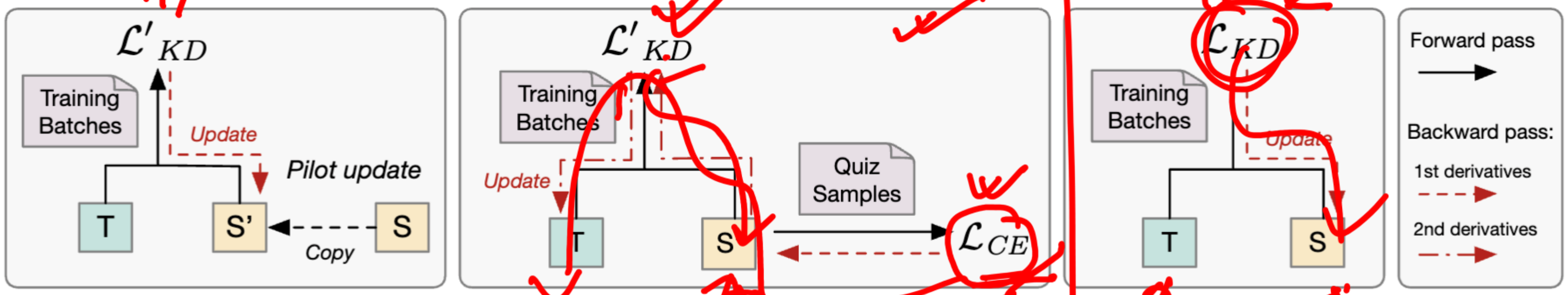


Knowledge sharing is **unidirectional**, *i.e.*, teacher is not aware of student's capacity

KD with Meta Learning – Zhou et al., 2022



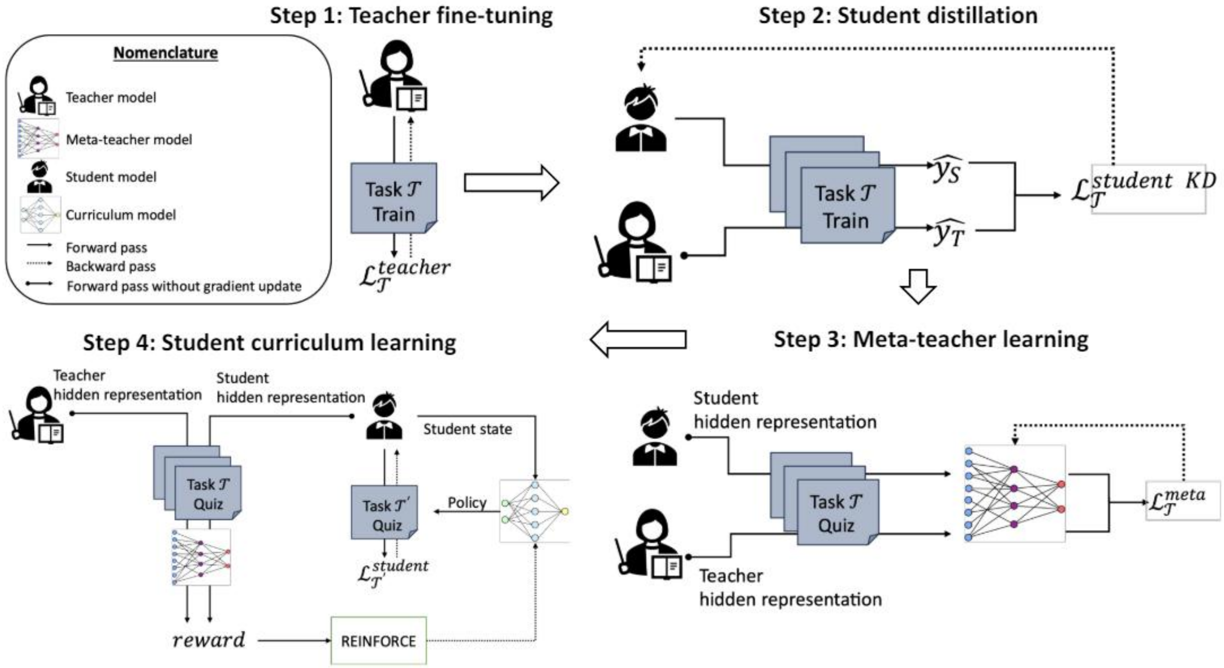
- Traditional KD approaches are uni-directional, *i.e.*, teacher is mostly trained prior to the KD process; therefore, teacher is unaware of the student's capacity.
- The teacher pre-training procedure is not optimized for distillation purposes; good model may not be always a good teacher
- To address these challenges, Zhou et al., proposed a meta-KD method where the teacher model is also trained in a meta loop, enabling better knowledge dissipation in the subsequent KD step.



(1) Teaching experiment (2) Quiz & Meta update (3) Knowledge distillation

MPDistil: Student-Aware Meta Distillation: *Learning to teach*

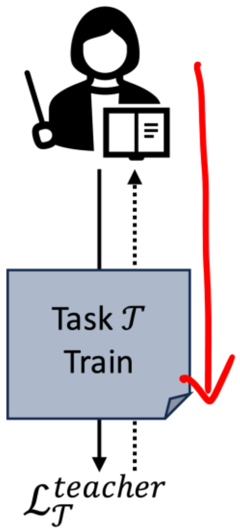
- A **healthy competition** between the teacher and student can encourage both the models to perform better.
- A **better teacher can set a higher benchmark for the student**, enhancing student's performance.
- The student can devise **better learning strategy** (curriculum) to perform better than the teacher.



Sengupta, Dixit, Akhtar, **Chakraborty**. A Good Learner Can Teach Better: Teacher-Student Collaborative Knowledge Distillation. **ICLR** 2024.

MPDistil: Step 1 -- Teacher Fine-tuning

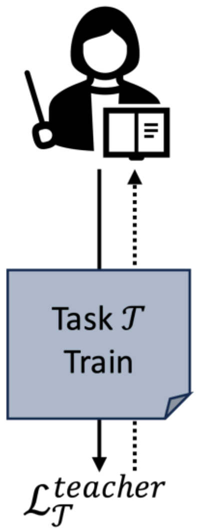
1. Teacher Fine-tuning



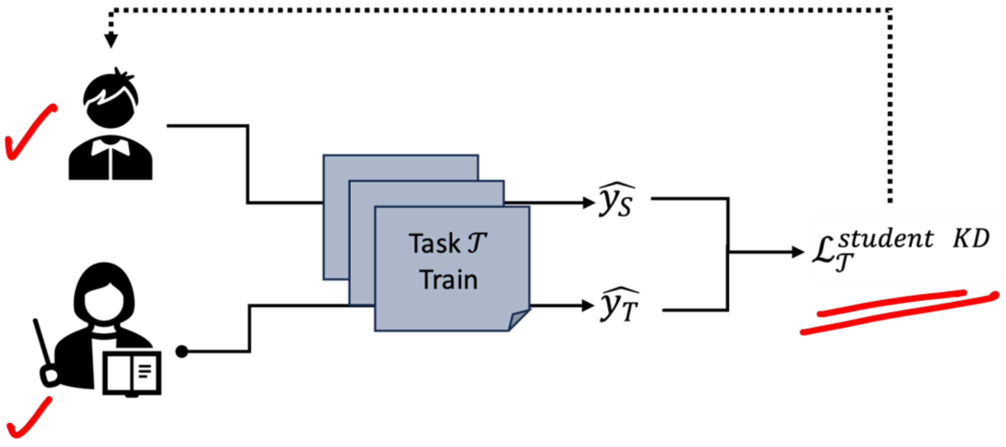
$$\mathcal{L}_{\mathcal{T}}{}^{teacher} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathcal{T}}(y_i, \hat{y}_{(i,T)}), \text{ with } \hat{y}_{(i,T)} = T(x_i; \theta_T)$$

MPDistil: Step 2 -- Student Distillation

1. Teacher Fine-tuning



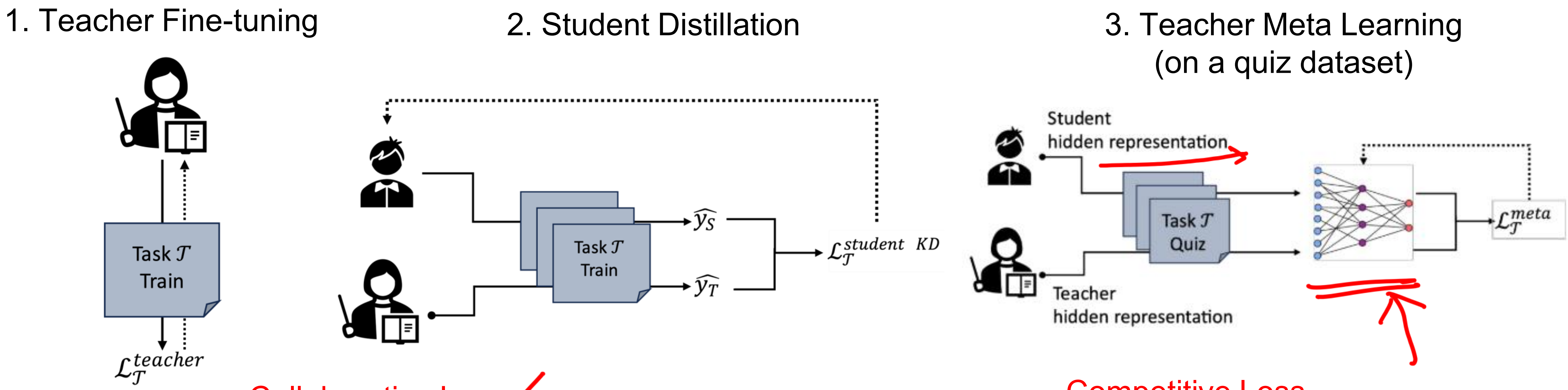
2. Student Distillation



KD

$$\mathcal{L}_{\mathcal{T}}{}^{\text{student KD}} = \frac{1}{N}\sum_{i=1}^{N} \alpha \mathcal{L}_{\mathcal{T}}(y_i, \hat{y}_{(i,S)}) + (1-\alpha)\left\|\Phi\left(\frac{\hat{y}_{(.,T)}}{\tau}\right) - \Phi\left(\frac{\hat{y}_{(.,S)}}{\tau}\right)\right\|_2 + \beta\left\|h_{(.,T)} - h_{(.,S)}\right\|_2$$

MPDistil: Step 3 -- Meta-teacher Learning



Collaborative Loss

$$\mathcal{L}_{\mathcal{T}}{}^{\text{meta col}} = \begin{cases} -\dfrac{1}{2N}\displaystyle\sum_{i=1}^{N}\left[\log \bar{y}_{(i,\mathcal{T})} + \log \bar{y}_{(i,S)}\right], & \text{if } \mathcal{T} \text{ is a classification task} \\ \dfrac{1}{2}\left\|y-\hat{\hat{y}}_{(.,\mathcal{T})}\right\|_2 + \dfrac{1}{2}\left\|y-\hat{\hat{y}}_{(.,S)}\right\|_2, & \text{if } \mathcal{T} \text{ is a regression task} \end{cases}$$

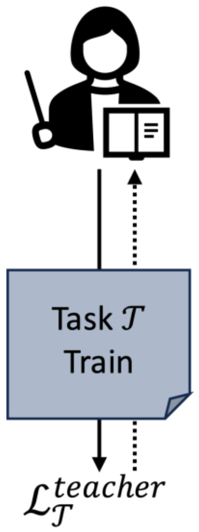
Competitive Loss

$$\mathcal{L}_{\mathcal{T}}{}^{\text{meta com}} = \begin{cases} -\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left[2\log \bar{y}_{(i,\mathcal{T})} - \log \bar{y}_{(i,S)}\right], & \text{if } \mathcal{T} \text{ is a classification task} \\ \left\|y-\hat{\hat{y}}_{(.,\mathcal{T})}\right\|_2 - \dfrac{1}{2}\left\|y-\hat{\hat{y}}_{(.,S)}\right\|_2, & \text{if } \mathcal{T} \text{ is a regression task} \end{cases}$$

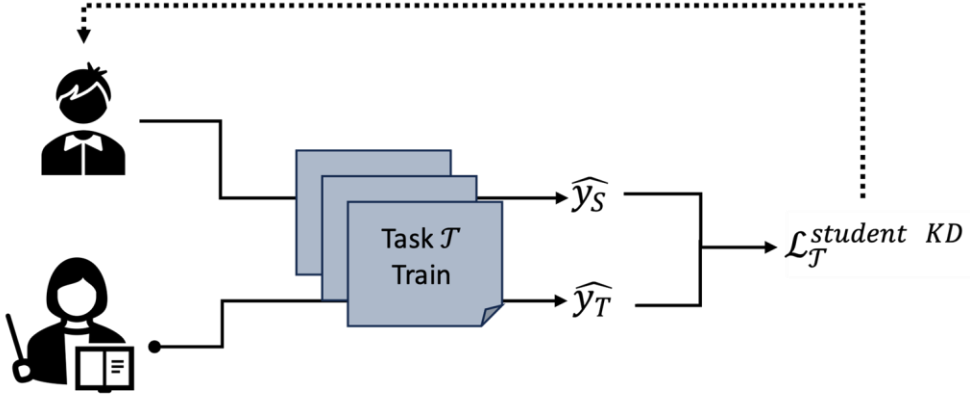
Intuition: The meta-teacher obtains the hidden states from both teacher and student and creates a healthy competition between the models.

MPDistil: Step 4 -- Student Curriculum Learning

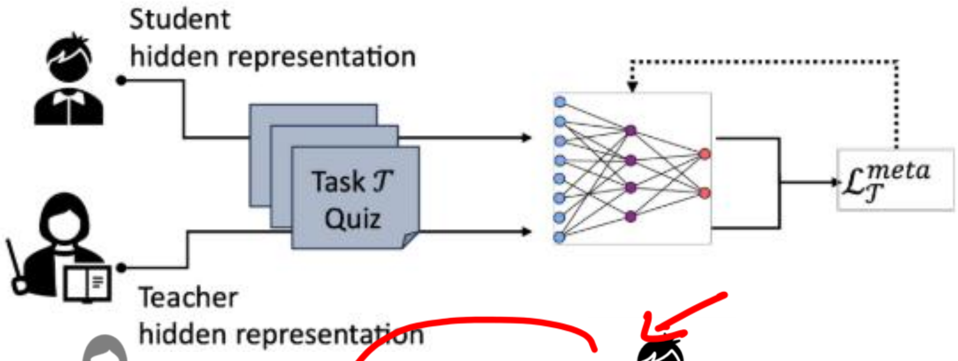
1. Teacher Fine-tuning



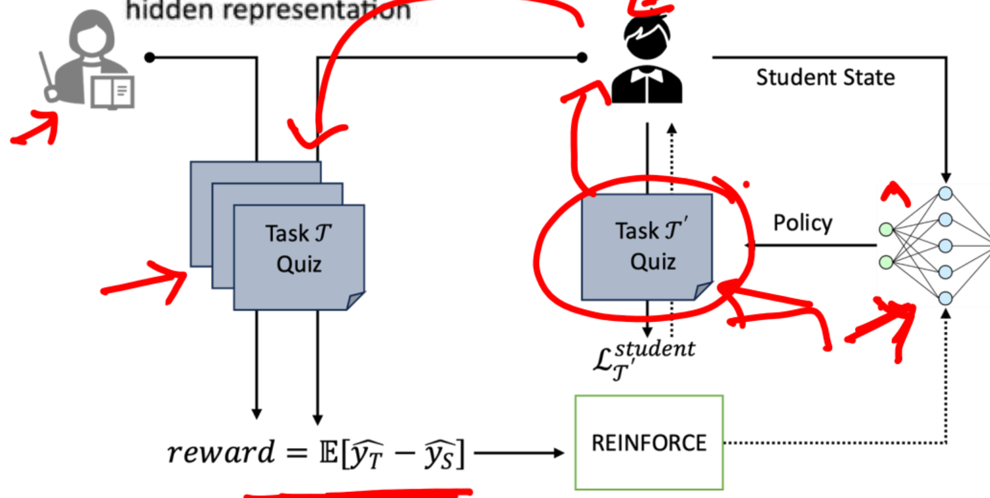
2. Student Distillation



3. Teacher Meta Learning
(on a quiz dataset)



4. Student Curriculum
Learning

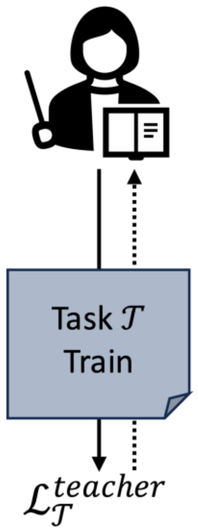


Why Curriculum Learning in KD?

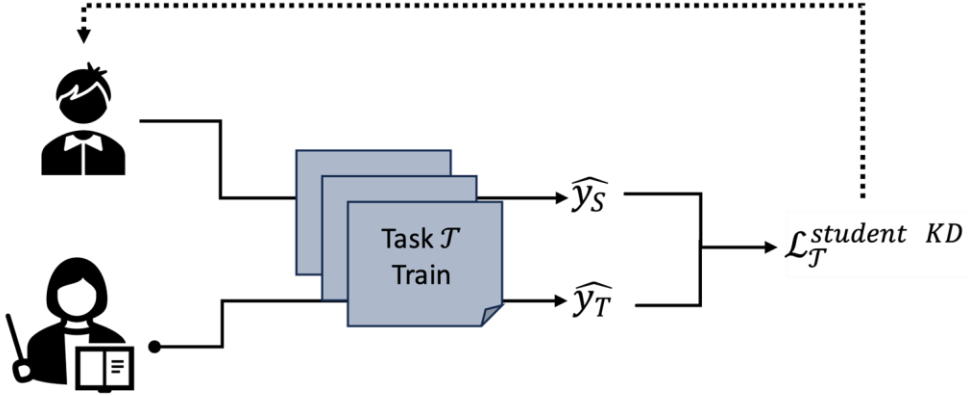
In real world, a student might aim to improve her understanding of Physics by studying selected concepts from Mathematics.

MPDistil: Step 4 -- Student Curriculum Learning

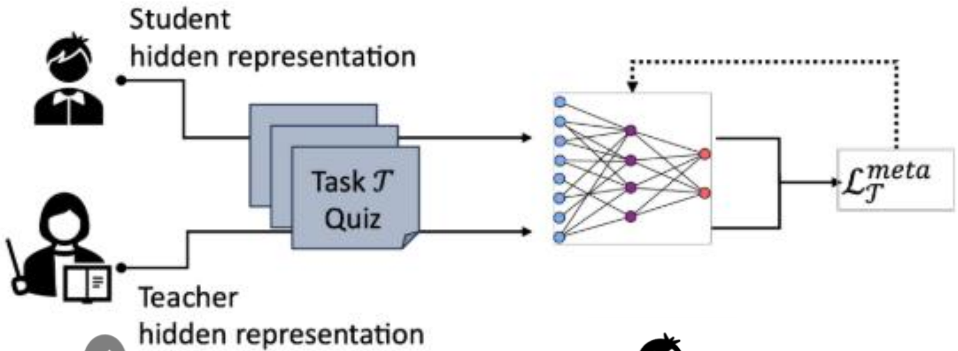
1. Teacher Fine-tuning



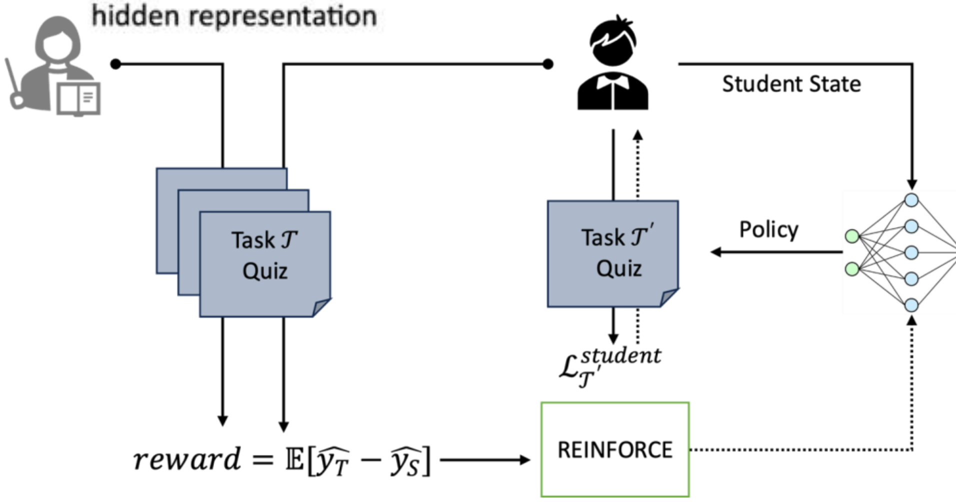
2. Student Distillation



3. Teacher Meta Learning
(on a quiz dataset)



4. Student Curriculum Learning



Competing student tries to beat the teacher

A policy network selects optimal curriculum to fine-tune the student by maximizing the reward

$$R^{\text{binary}} = \begin{cases} \mathbb{I}_{\hat{y}_{(i,S)} > \hat{y}_{(i,T')}}, & \text{if } \mathcal{T} \text{ is a classification task} \\ \mathbb{I}_{\left\|y_i - \hat{y}_{(i,T')}\right\|_2 > \left\|y_i - \hat{y}_{(i,S)}\right\|_2}, & \text{if } \mathcal{T} \text{ is a regression task} \end{cases}$$

$$R^{\text{real}} = \begin{cases} \hat{y}_{(i,S)} - \hat{y}_{(i,T')}, & \text{if } \mathcal{T} \text{ is a classification task} \\ \left\|y_i - \hat{y}_{(i,T')}\right\|_2 - \left\|y_i - \hat{y}_{(i,S)}\right\|_2, & \text{if } \mathcal{T} \text{ is a regression task} \end{cases}$$

ICLR’24

A “smart” student can beat a teach!!

Methods	BoolQ	CB	COPA	RTE	WiC	WSC
KD Hinton et al. (2015)	-13.3	-19.1	-4.3	-3.7	-9.1	-14.4
PD Turc et al. (2019) †	-9.6	-9.5	-0.3	-13.5	-6.9	-11.2
PKD Sun et al. (2019)	-1.7	-5.9	-6.0	-3.8	-0.4	-12.5
DistilBERT Sanh et al. (2019) †	-6.0	-7.7	-1.0	-12.0	-5.8	-9.3
Theseus Xu et al. (2020) †	-1.6	-3.6	-4.3	-4.8	-1.8	-11.5
TinyBERT Jiao et al. (2019)	**-1.4**	-1.2	4.3	-3.7	1.7	-2.9
MobileBERT Sun et al. (2020) †	-4.8	-2.4	-0.7	-14.0	-2.3	-9.3
SID Aguilar et al. (2020) †	-10.1	-17.3	-1.0	-14.8	-9.0	-12.8
MiniLM Wang et al. (2020b) †	-3.5	-11.9	-4.0	-5.3	-1.2	-14.4
MiniLMv2 Wang et al. (2020a) †	-2.7	-14.3	-4.0	-6.3	-2.5	-15.1
ALP-KD Passban et al. (2021) †	-2.2	-11.3	-5.3	-4.8	-1.3	-13.1
LRC-BERT Fu et al. (2021) †	-4.5	-9.5	-0.3	-16.4	-8.5	-11.2
Annealing-KD Jafari et al. (2021) †	-8.8	-5.9	3.3	-14.0	-6.3	-11.2
CKD Park et al. (2021) †	-7.8	-6.6	-1.0	-11.7	-7.3	-11.2
Universal-KD Wu et al. (2021a) †	-1.8	-5.4	-7.3	-2.8	-0.6	-11.2
DIITO Wu et al. (2021b) †	-3.9	-5.9	6.0	-7.5	-5.4	-8.6
Continuation-KD Jafari et al. (2022) †	-8.0	-7.1	2.7	-14.2	-7.9	-13.1
RAIL-KD Haidar et al. (2021) †	-10.4	-7.7	0.7	-12.4	-5.8	-7.7
MGSKD Liu et al. (2022a)) †	-6.1	-6.6	-1.0	-7.0	-3.0	-12.8
MetaDistil Zhou et al. (2021)	-2.7	-1.8	1.0	-2.0	-1.6	0.9
MPDistil (Ours)	-1.9	**0.0**	**7.0**	**0.4**	**2.5**	**1.0**
(-) Curriculum learning	-2.8	-5.3	-4.0	-1.8	1.2	0.0

S – T

Positive value indicates the **student model** is **better than the teacher model**

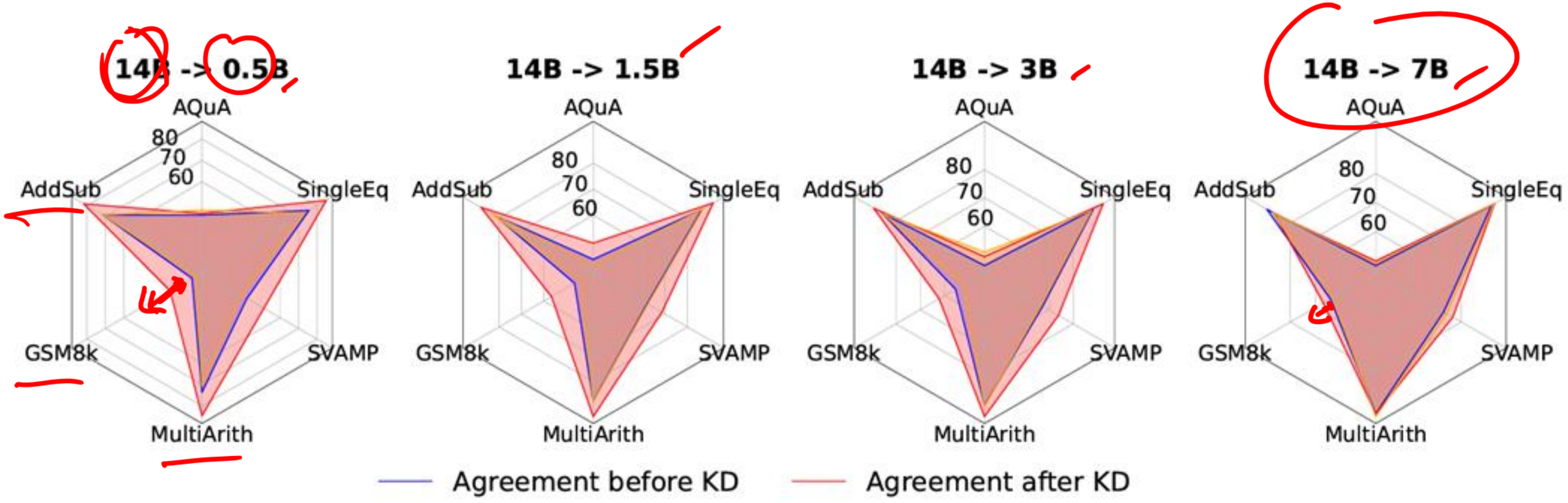
Explaining Knowledge Distillation

Known: KD improves generalization abilities of student models.

Questions

(i) Post-KD, does student perfectly *imitate* a teacher?
(ii) What are the *key drivers* influencing the effectiveness of KD methods?

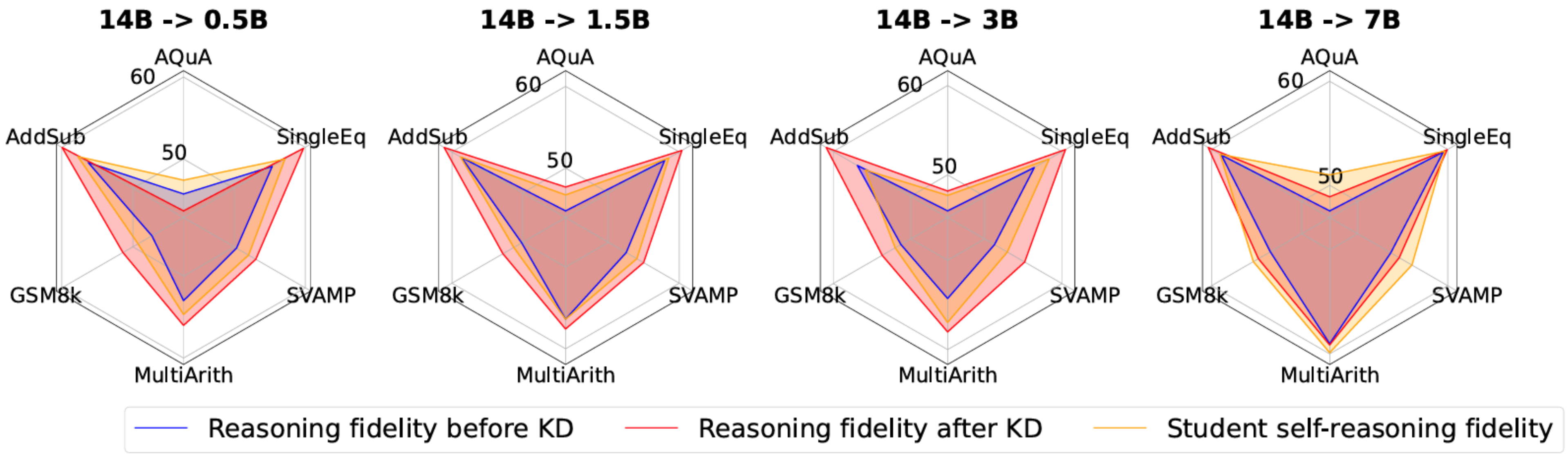
Agreement b/w Teacher-Student Post-KD



Agreement: Overlap between the final output generated by teacher and students.

Teacher-student agreement improves post KD, mostly for smaller LMs (<7B).

Fidelity b/w Teacher-Student Post-KD



•**Fidelity:** Ability of the student to imitate the teacher's reasoning behaviors.

- Smaller LMs tend to have better fidelity post-KD.
- However, statistical tests show that fidelity does not necessarily improve the generalization abilities of student models!!

Fidelity vs Generalization Paradox of KD

Instruction	Student KD Output	Teacher SFT Output	Fidelity
Robin has 28 packages of gum and 13 packages of candy. There are 4 pieces in each package. How many pieces of gum does Robin have?	To solve this problem, we need to use multiplication. 1. Identify the given information: Robin has 28 packages of gum and 13 packages of candy. Each package has 4 pieces of gum and candy. 2. Determine the operation needed to find the total number of pieces of gum: We need to multiply the number of packages of gum by the number of pieces of gum in each package. 3. Multiply: 28 by 4 to find the total number of pieces of gum. 56. The answer is 56. Therefore, Robin has 56 pieces of gum.<|endoftext|>	To solve this problem, we need to use multiplication. 1. Multiply the number of packages of gum by the number of pieces in each package: 28 packages x 4 pieces = 112 pieces of gum 2. Therefore, Robin has 112 pieces of gum. The answer in Arabic numerals is: 112.<|endoftext|>	0.79

- High teacher-student fidelity, but wrong answer predicted by student (poor generalization)

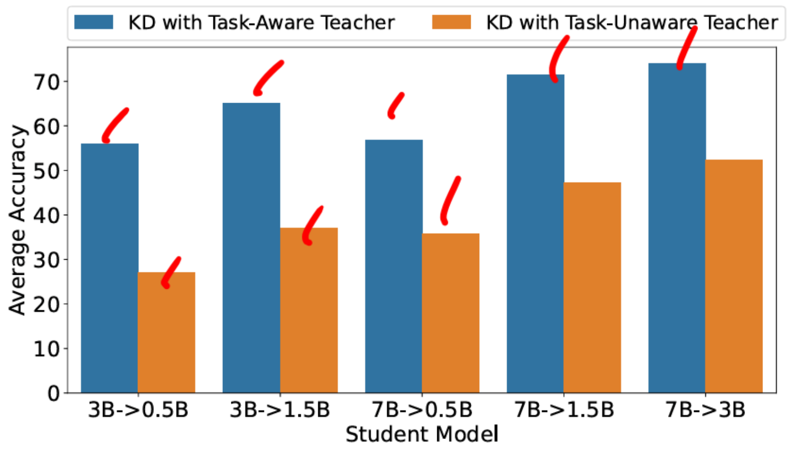
Fidelity vs Generalization Paradox of KD

Instruction	Student KD Output	Teacher SFT Output	Fidelity
Because of the decision Sofia asked 310 students to suggest specific types of food. 185 students suggested adding mashed potatoes while others suggested adding bacon to the menu. How many students suggested bacon?	To solve the problem, we can use subtraction:number of students = 310of students who suggested mashed potatoes = 185find the number of students who suggested bacon, we subtract the number of students who suggested mashed potatoes from the total number of students: 310 - 185 = 125 Therefore, 125 students suggested adding bacon to the menu.<|endoftext|>	We can start by using the concept of complementary counting. Let x be the number of students who suggested bacon. We know that the total number of students is 310, so we can set up an equation: 185 + x = 310 Subtracting 185 from both sides: x = 310 - 185 Simplifying: x = 125 Therefore, 125 students suggested adding bacon to the menu. Answer: 125.<|endoftext|>	0.13

- Low teacher-student fidelity, but good generalization

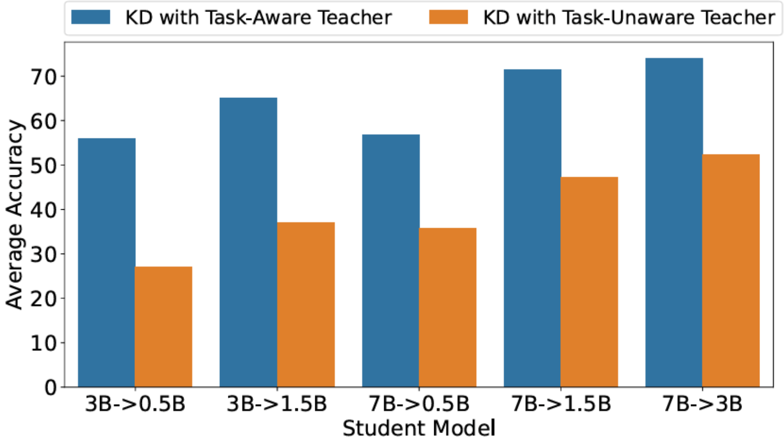
Therefore, the tradeoff between generalization vs fidelity-agreement remains prominent.

Drivers behind Successful KD

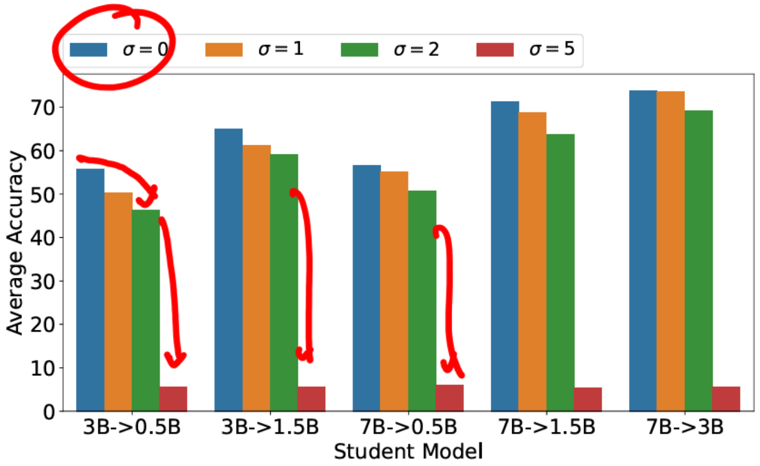


1. Teacher model should be task-aware

Drivers behind Successful KD



1. Teacher model should be task-aware

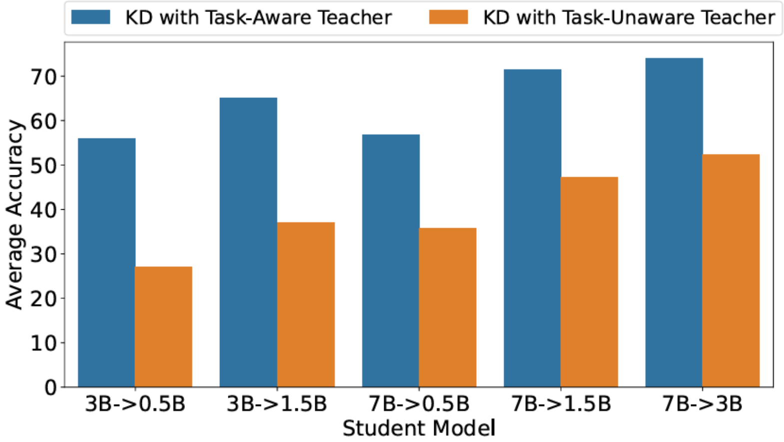


2. Teacher signals to student should be noise-free.

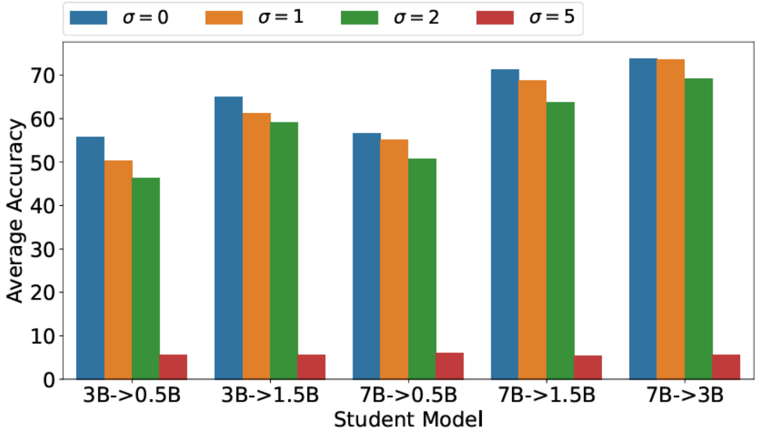
Here σ is the amount of Gaussian noise added to the teacher logits before distilling to student. For σ, student performance drops drastically.

Teacher model performance ***minimally affects*** student outcomes; however, the teacher's ***task-specific expertise is crucial***

Drivers behind Successful KD

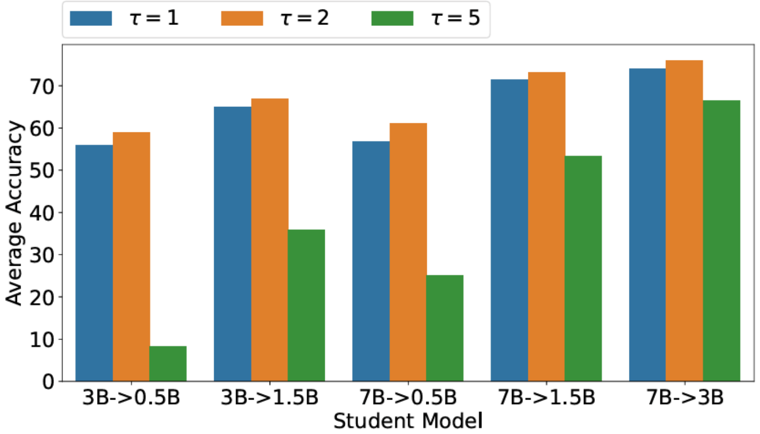


1. Teacher model should be task-aware



2. Teacher signals to student should be noise-free.

Here σ is the amount of Gaussian noise added to the teacher logits before distilling to student. For σ, student performance drops drastically.



3. Logit smoothing is important

Here τ is the temperature used to smoothen the teacher logits. Too much smoothing hurts student performance, but moderate smoothing shows benefit.

Temperature (τ) in KD balances precision ($\tau \downarrow$) and recall ($\tau \uparrow$) of the student model.