On the Generalization vs Fidelity Paradox in Knowledge Distillation

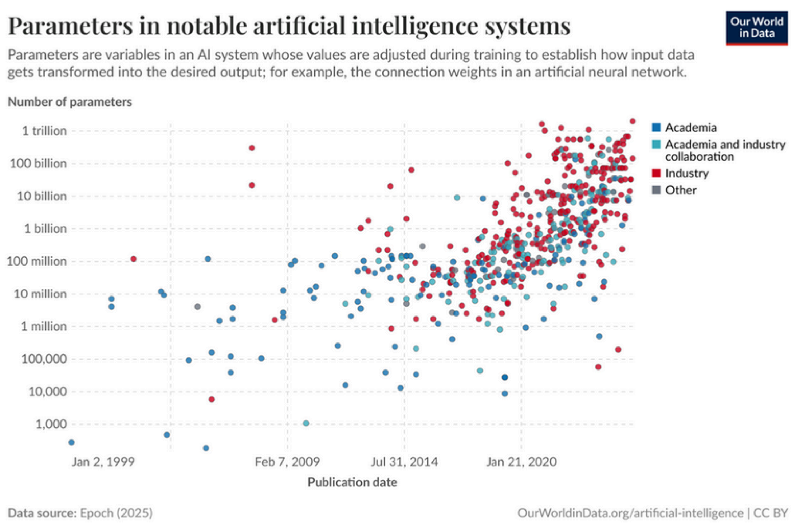
Suhas Kamasetty Ramesh*, Ayan Sengupta*, Tanmoy Chakraborty

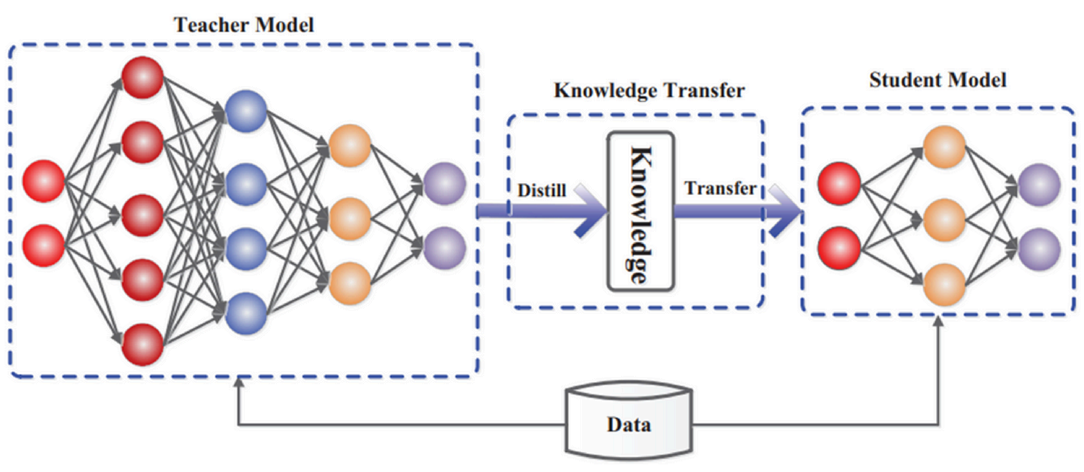
Indian Institute of Technology Delhi,

Laboratory for Computational Social Systems (LCS2)

Knowledge Distillation and Its Importance

- Rapid advancement of pre-trained language models(LMs) has led to the development of large-scale language models that achieve state-of-the-art performance across various NLP tasks.
- Deploying these large models presents significant challenges due to their high computational & memory requirements





- Knowledge Distillation is a model compression technique that aims to transfer knowledge from a large model(teacher model) to a smaller model(student model)
- Broadly, the goal is to train a small student network to match the predictions made by the larger teacher network

Analytical Understanding of KD

- Despite the recent traction of KD research, its effectiveness for smaller language models (LMs) and the mechanisms driving knowledge transfer remain underexplored.
- Existing studies overlook explainability of KD
- We present the first large-scale empirical and statistical analysis of KD across models ranging from 0.5B to 14B

Models:

Qwen 2.5 Llama 3

Models sizes: 0.5B to 14B

21 Teacher Student pairs

KD Methods:

Rev-KD

Seq-KD

GKD

Current KD methods for LLMs

Benchmarks:

Mathematical Reasoning

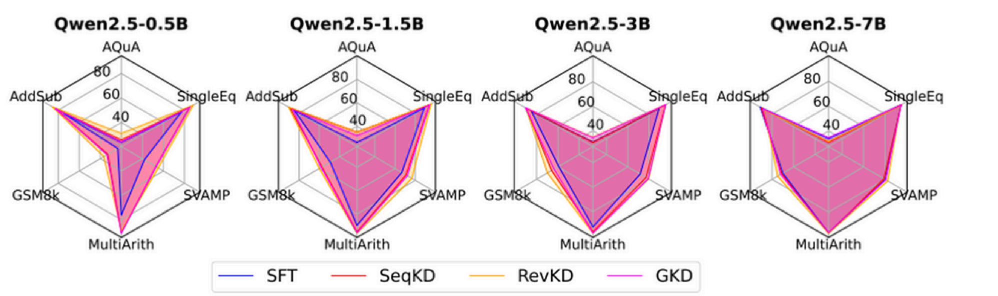
Instruction Following

Commonsense Reasoning

Results

Effectiveness of KD

- KD consistently outperforms SFT in all three benchmarks
- KD significantly improves model generalization
- Zero-shot performance of smaller LMs (<1B size) can be improved by up to 10% with peak task gains reaching 22% post distillation
- A one-way ANOVA test reveals no significant difference on math and commonsense reasoning benchmarks - suggesting that all KD methods perform similarly on these two benchmarks



(a) Performance on mathematical reasoning tasks

Does KD depend on student model size?

- We find a negative correlation (-0.66, p-value=0.0) between KD improvement and student size, indicating diminishing returns as model size increases.
- While smaller models (< 1B) improved by upto 10%, larger (7B) models exhibit only ~1.3% improvement after distillation, indicating that KD is most effective for smaller LMs.

Test Type	AQuA	AddSub	GSM8k	MultiArith	SVAMP	SingleEq	Average
Spearman rank	-0.23 (0.2)	**-0.64 (0.0)**	**-0.51 (0.0)**	**-0.83 (0.0)**	**-0.43 (0.0)**	**-0.59 (0.0)**	**-0.66 (0.0)**

(a) Mathematical reasoning

Test Type	ARC-c	ARC-e	BoolQ	Hellaswag	OBQA	PiQA	SiQA	Winogrande	Average
Spearman rank	**-0.63 (0.0)**	**-0.78 (0.0)**	-0.15 (0.4)	**-0.4 (0.0)**	**-0.55 (0.0)**	**-0.67 (0.0)**	**-0.49 (0.0)**	**-0.34 (0.0)**	**-0.54 (0.0)**

(b) Commonsense reasoning

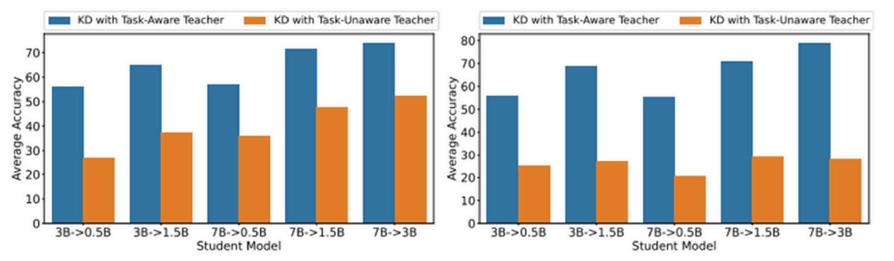
Test Type	Dolly	Self	SNI	UNI	Vicuna	Average
Spearman rank	0.08 (0.7)	0.3 (0.1)	-0.32 (0.1)	-0.02 (0.9)	0.12 (0.5)	-0.03 (0.9)

(c) Instruction following

Table 3: Spearman rank correlation and p-value between student performance and student model size.

Does KD depend on teacher performance?

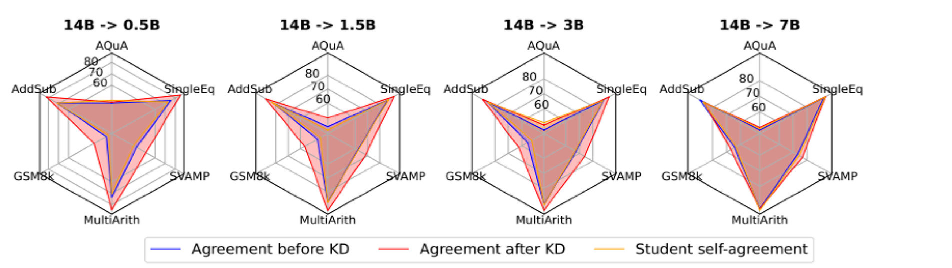
- We conducted Spearman rank correlation between student improvement after KD and teacher performance.
- In most tasks, correlations are generally weak or negative, indicating that teacher quality alone does not dictate KD success.
- But, a task-unaware teacher can significantly degrade student performance.
- Post-distillation performance of student model can drop up to 40%, if teacher is not fine-tuned on downstream domain.



(a) Mathematical reasoning (b) Commonsense reasoning

Teacher Student Agreement

- KD significantly improves teacher-student agreement
- Smaller models (0.5B & 1.5B) exhibit higher agreement with larger Qwen-14B model in structured mathematical tasks like AddSub (89.1%) and MultiArith (94.5%), indicating effective transfer of well-defined rules.
- However, a statistical test shows no significant corelation between student performance and teacher student agreement



(a) Mathematical reasoning tasks

Figure 8: Student agreement with Qwen-14B teacher.

New metrics: Agreement & Fidelity

- Apart from performance analysis of student model, we propose two metrics to understand alignment between teacher and student
- <u>Teacher-Student Agreement</u> - quantifies how often student replicates the teacher's outputs and is measured using top-1 agreement (fraction of matching predictions). (this metric only relies on the final answer and does not measure the quality of the intermediate reasoning steps)
- <u>Reasoning Fidelity</u> - captures how well the student mirrors the teacher's reasoning process rather than just final predictions. We use BLEU score between teacher and student reasoning outputs to compute fidelity

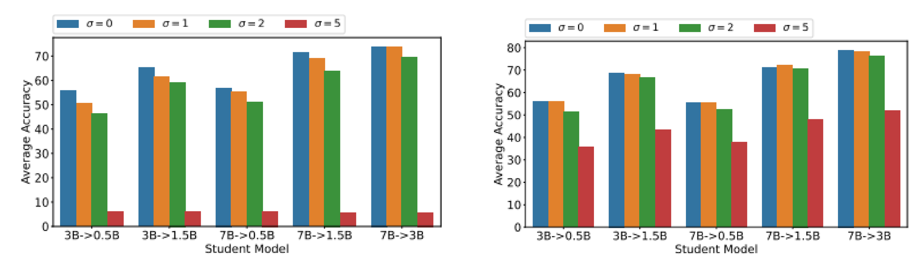
Reasoning Fidelity

- Although KD improves Fidelity over SFT, there is a weak correlation between fidelity and student performance, implying that strict imitation does not necessarily enhance performance
- The mismatch between student accuracy and poor fidelity suggests that while KD improves performance, it may fail to preserve the teacher's structured decision-making process, raising concerns about interpretability and reliability in critical applications.
- Error analysis on SVAMP test-set further highlights the feeble connection between teacher-student fidelity and student accuracy.

Instruction	Student KD Output	Teacher SFT Output	Fidelity
Randy has 37 blocks. He uses 33 blocks to build a tower and 13 blocks to build a house. How many more blocks did he use to build the tower than he did to build the house?	To find the difference between the number of blocks used to build the tower and the house, you can subtract the number of blocks used to build the house from the number of blocks used to build the tower. Here are the steps: 37 blocks - 13 blocks = 24 blocks Therefore, Randy used 24 more blocks to build the tower than he did to build the house.<|endoftext|>	Step 1: Find the total number of blocks used 33 blocks (for the tower) + 13 blocks (for the house) = 46 blocks Step 2: Find the difference between the number of blocks used for the tower and the house 33 blocks (for the tower) - 13 blocks (for the house) = 20 blocks Therefore, Randy used 20 more blocks to build the tower than he did to build the house. The answer in Arabic numerals is 20.<|endoftext|>	0.77

Noisy teacher signals

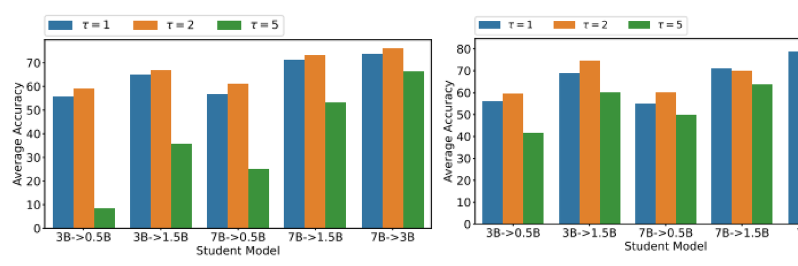
- We investigate the significance of teacher signals by injecting Gaussian noise($\epsilon \sim N(0, \sigma)$) into teacher logits before distillation
- Increasing σ from 0 (no noise) to 1 and 2 slightly reduces performance.
- At σ = 5, performance collapses



(a) Mathematical reasoning (b) Commonsense reasoning

Impact of temperature

- Temperature (τ) significantly impacts KD effectiveness.
- A moderate τ = 2 consistently yields the best results in most tasks.
- However, τ = 5 leads to severe performance drops, especially for smaller students.



(a) Mathematical reasoning (b) Commonsense reasoning

Conclusion

- KD significantly benefits smaller models and its effectiveness diminishes with increasing model size
- Teacher domain adaptation plays a more critical role than teacher performance
- Surprisingly, higher teacher-student agreement did not always correlate with better student performance
- Results underscore the need for task-aware KD strategies and adaptive distillation techniques tailored to student learning dynamics.
- Future research should explore alternative KD objectives, self-distillation mechanisms, and refined teacher-student alignment strategies to improve both performance and reasoning fidelity.

Paper: Code:

Primary contact ayan.sengupta@ee.iitd.ac.in

ACL 2025 VIENNA JULY 27 - AUGUST 1