How to Fine-Tune a Reasoning Model?
A Teacher–Student Cooperation Framework to Synthesize Student-Consistent SFT Data

Zixian Huang, Kaichen Yang, Xu Huang, Feiyang Hao, Qiming Ge, Bowen Li, He Du, Kai Chen and Qipeng Guo
huangzixian@pjlab.org.cn
Shanghai AI Laboratory

SFT a well-trained reasoning model would lead to catastrophic forgetting.

Given prompts, training smaller models on data generated by stronger teachers is a widely used strategy for model enhancement.

However, for reasoning models already capable of deep thinking, further SFT can actually hurt performance.

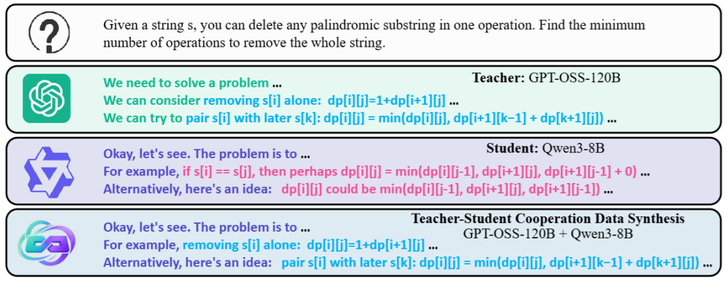
On tasks where models already have strong performance (e.g., code/math reasoning), superior teacher data leads to drops.

Despite response generated by strong teacher GPT-OSS-120B (44.61%), direct SFT causes Qwen3-8B drop from 18.75% to 8.73% on OJBench.

The style gap between SFT data and the target model is key.

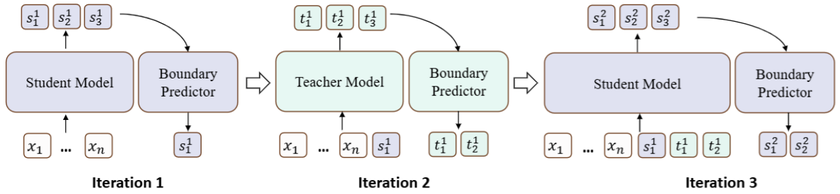
Distinct Model "Accents": GPT-OSS-120B uses formal structures like "We need to ...", while Qwen3-8B favors conversational phrases like "Okay, let's see ...".

Learning the Accent, Losing the Knowledge: Direct SFT forces the student to waste capacity mimicking the teacher's "accent" rather than mastering the reasoning.



Can we keep the Teacher's brain while using the Student's 'accent'?

TESSY: Teacher–Student Cooperation Data Synthesis framework.



Alternating Generation: Interleaves Teacher and Student models to generate reasoning trajectories.

Role Division: Teacher provides core reasoning capability; Student maintains stylistic consistency.

Generate-then-Rollback: Generates a fixed-length span first, then employs boundary predictors to truncate the span and assign the next generation role.

Boundary Predictor: Trains 0.6B sequence labeling models using random segments auto-annotated by the teacher model.

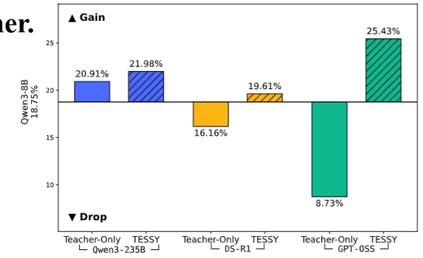
TESSY enables GPT-OSS-120B to synthesize on-policy SFT data for Qwen3-8B.

Boosts by 11.25% on LCB-Pro, while Teacher-Only SFT a 3.25% drop.

	In-domain Test Sets				Out-of-domain Test Sets			
	LCB-V5	LCB-V6	LCB-Pro	OJBench	GPQA-Diamond	AIME2024	AIME2025	OlympiadBench
Qwen3-8B	55.09	49.58	25.35	18.75	60.16	76.67	69.17	**69.81**
+ Reject-Sampling	54.55	48.00	26.06	18.32	**60.61**	75.42	66.77	69.53
+ Self-Distillation	51.50	43.43	19.26	14.22	**60.61**	71.98	62.29	67.80
+ Teacher-Only	41.32	36.57	22.10	8.73	55.81	76.88	66.46	67.76
+ TESSY	**62.87**	**55.43**	**36.69**	**25.43**	59.34	**80.42**	**70.10**	69.80

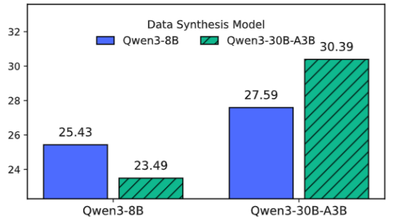
Consistent gains across diverse teacher.

TESSY Gains over Teacher-Only SFT on OJBench:
+1.07% (with Qwen3-235B Teacher)
+3.45% (with DeepSeek-R1 Teacher)
+16.70% (with GPT-OSS-120B Teacher)



Consistency > Capability: Use the Target Model as the Student

TESSY Framework (Target 8B Training):
Matched 8B Student > Superior 30B-A3B Student (+1.94%)

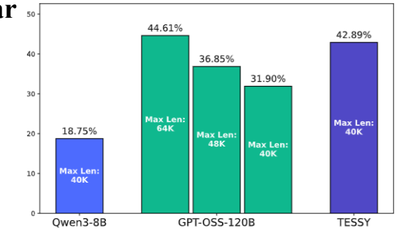


Achieves high-quality synthesis on par with Direct Teacher Generation.

At a 40K token limit, TESSY outperforms GPT-OSS-120B by 10.99%.

GPT-OSS-120B requires 64K tokens to only marginally surpass TESSY at 40K.

TESSY shorter reasoning traces.



Paper Link

Github Link

Synthetic Training Set