COPD JD.com

Ask the teacher.
Then learn to act alone.

Communicative On-Policy Distillation: distilling black-box teacher capabilities through communication for long-horizon tasks

Lirui Luo · Kelong Mao · Xinwei Yang · Rongqing Li · Luyu Chen · Kieran Wong
Yudong Guo · Xinrui Wang · Jiayin Zhu · Gu Simiu · Sulong Xu · Cong Fang

Peking University · JD.com · Sichuan University · Beijing Institute of Technology
Renmin University of China · The Hong Kong University of Science and Technology

---

Language is the channel. The student decides when to use it.

liruiluo.github.io/communicative-opsd/

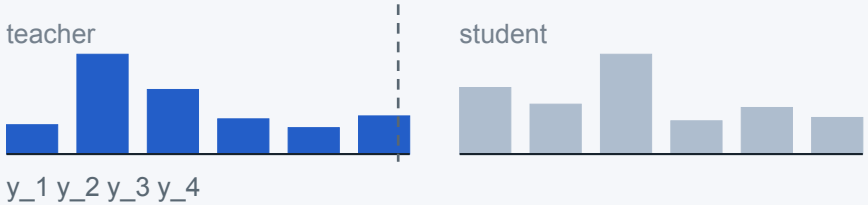
Roadmap

01	**Background**	What OPD needs, what black-box OPD does today, and how people learn instead.
02	**Method**	How does one conversation become training supervision?
03	**Results**	Does it work, what carries the gain, does it transfer?
04	**Takeaways**	What should an OPD practitioner take away?

A black-box teacher cannot score tokens

On-policy distillation compares two distributions at every sampled token.

White-box OPD: what it needs at one token



`per-token gap  log p_student − log p_teacher`

Every token needs the teacher distribution.

That means scores from inside the teacher.

The teacher must run in the training loop.

Black-box OPD today: supervision from generations

chunk-level verification of sampled continuations	OmniOPD
action-level rewards from teacher decisions	ActionOPD
rubric scores from teacher-student contrasts	ROPD, GAD
teacher completion of the student prefix	OLIVE

The teacher scores or completes. The student never asks.

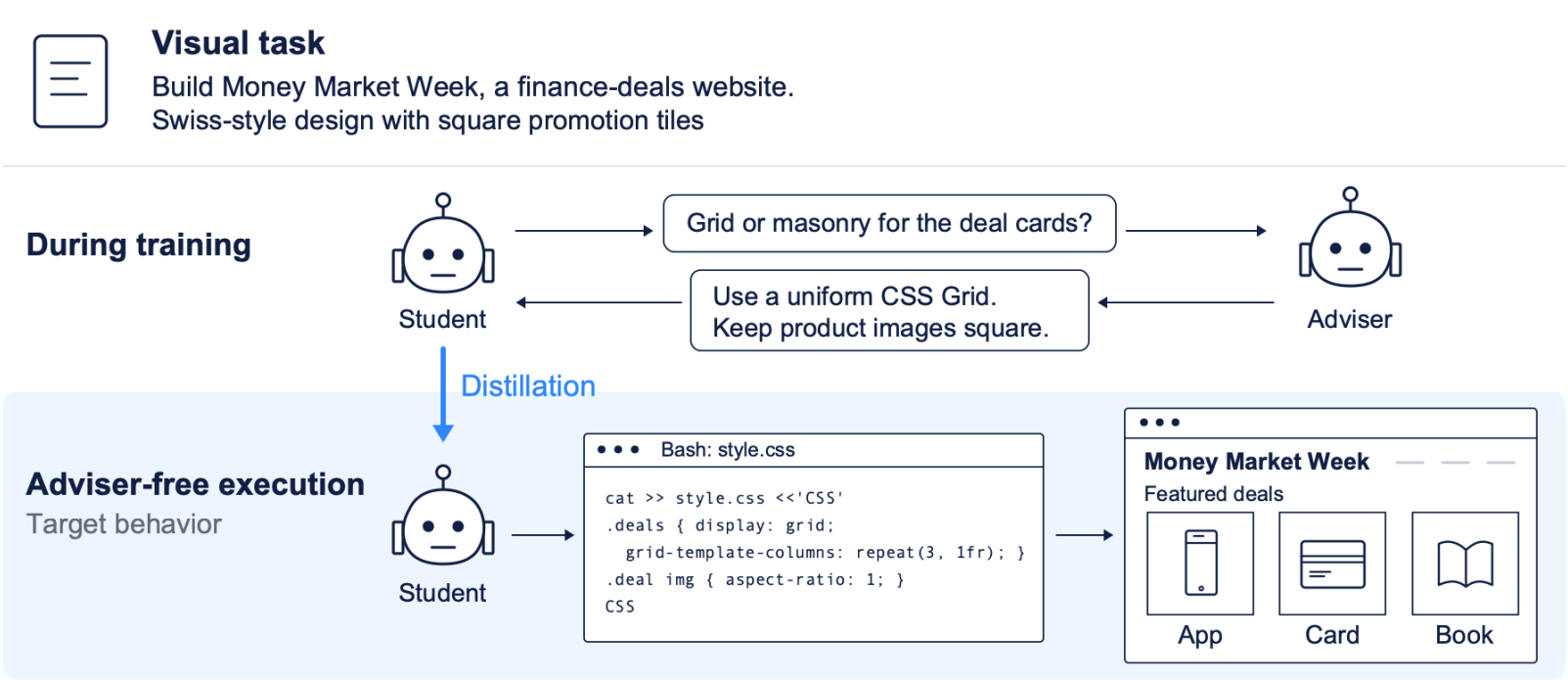
Text channel, but one direction: no consultation during execution.

Existing black-box OPD supervises from what the teacher produces, not from a dialogue.

So the questions are chosen by the pipeline, not by the student.

Human learning never needed logits

We consult those more experienced and internalize their guidance.



01

the learner decides when to ask

02

the answer comes back as language

03

guidance becomes skill, the adviser goes

A person learns exactly this way: ask the person who knows, hear it in words, then keep the answer in your own hands.

So can a model learn this way from a teacher it cannot see inside?

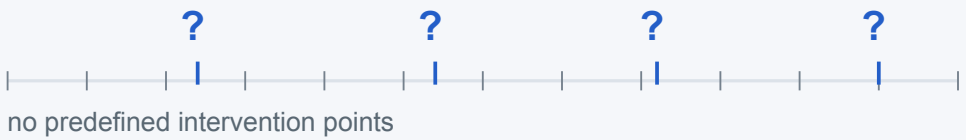
Figure 1 · a consultation during training, and the same task without one

Imitating it needs two things to be true

The learner decides when to ask, and the guidance has to stick.

FIRST CHALLENGE

When should it ask, and what should it say?

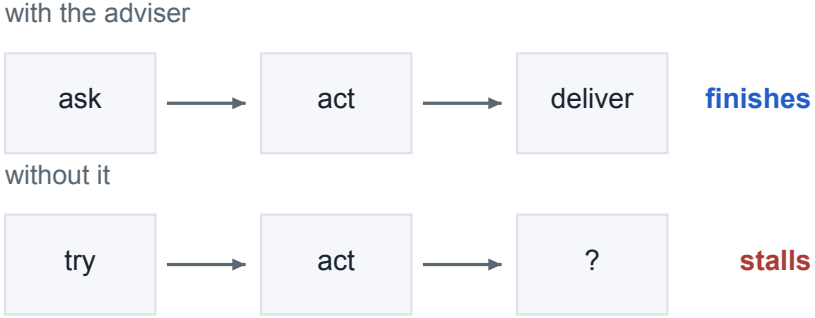


Two runs need the teacher at different steps.

The student must decide both while acting.

SECOND CHALLENGE

Does a successful consultation transfer?

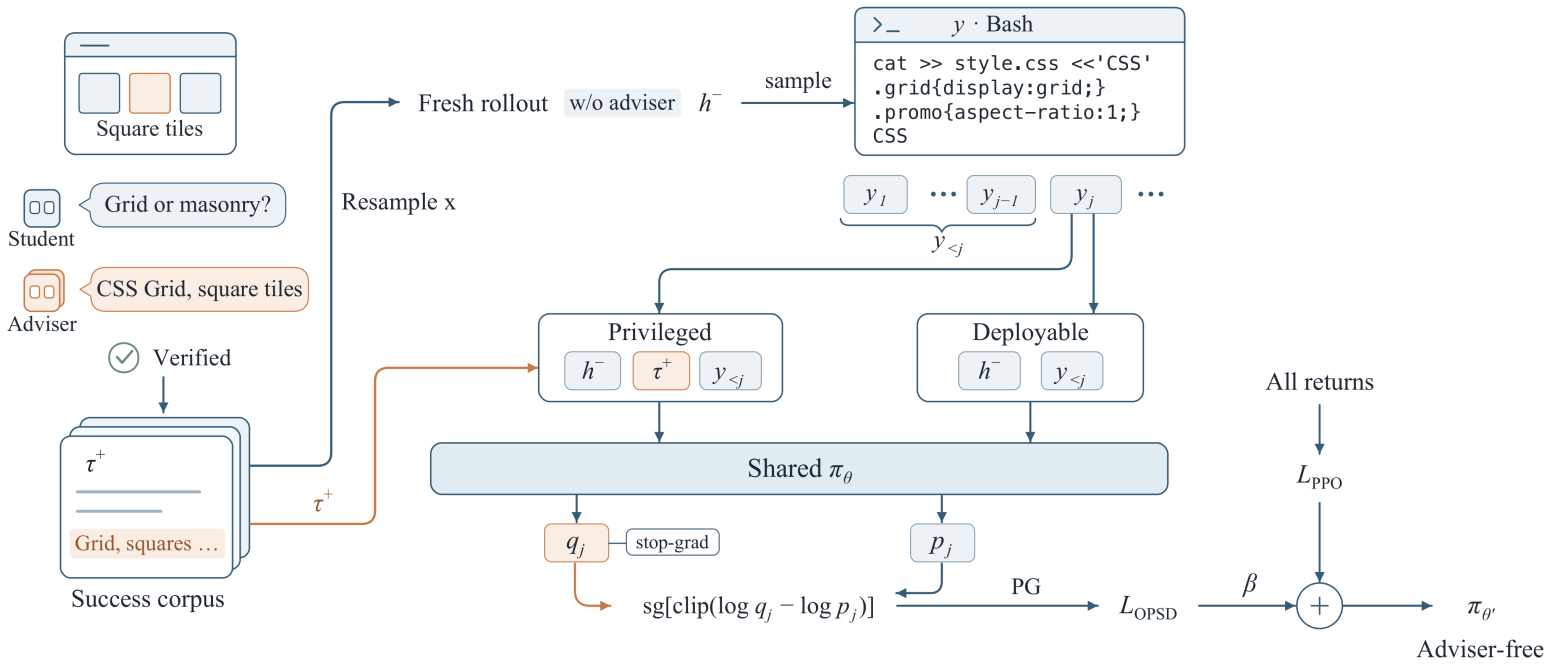


Finishing with help is not repeating it alone.

Neither is solved by having the teacher on the call.

Communication is the supervision channel

One policy acts, asks, and is distilled from its own successful consultations.



01 Ask

A free-text request is a policy action, trained by the task reward.

02 Keep

Task-verified trajectories that used the adviser form the reference corpus.

03 Distill

Two views score a fresh prefix, and only the privileged view sees the success.

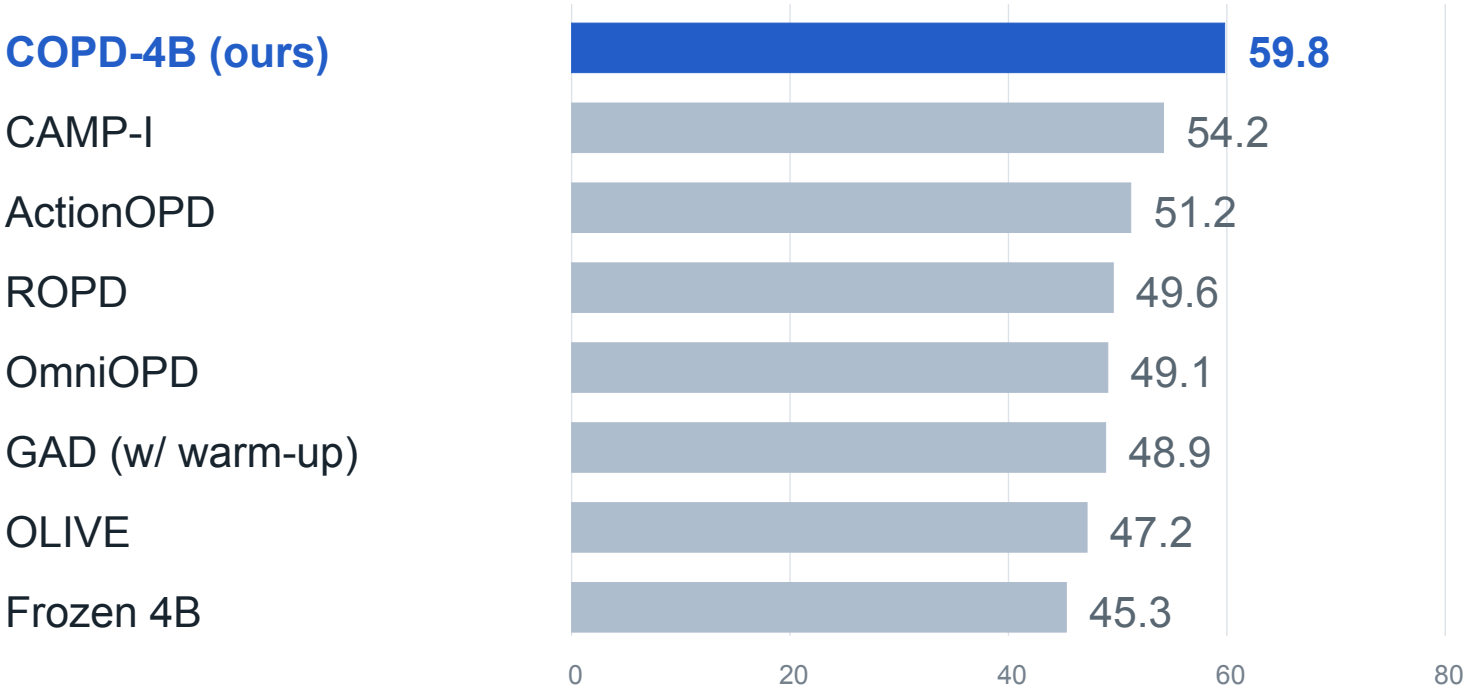
```
p = pi(. | c)
q = sg[pi(. | c, tau+)]
```

A free-text request is a policy action. The privileged view never acts.

Figure 2 · method overview

Best average of the 4B systems without an adviser

Experiment: eight 4B policies on the same five domains and task panels, with adviser access disabled for every policy.



+14.5

points over its own 45.3 starting point

2.46×

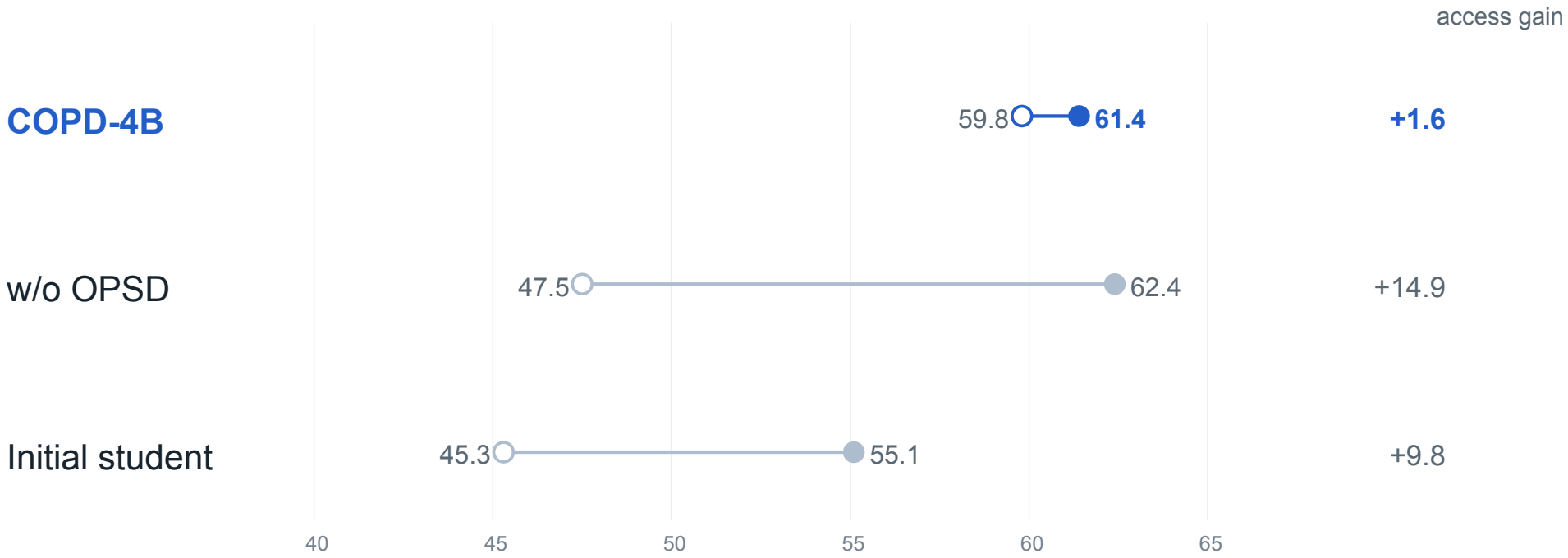
the gain of ActionOPD, the strongest black-box baseline

COPD gives the best adviser-free average of the 4B systems.

Table 1

The capability survives without the adviser

Experiment: score each policy twice on the same tasks, once with the adviser removed and once with it available.

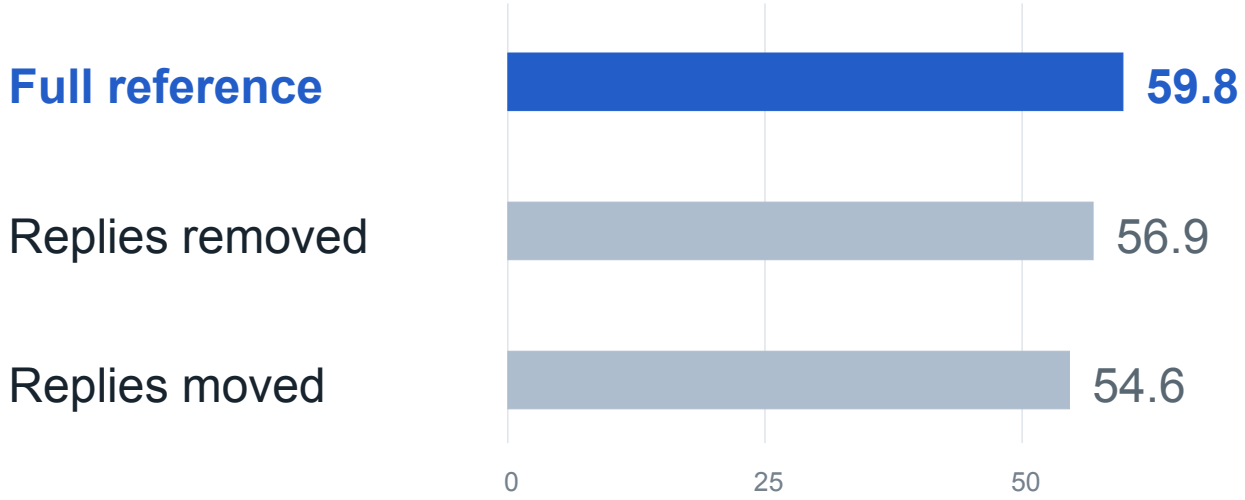


COPD internalizes the channel: dropping the adviser costs it least.

Figure 3

The words carry supervision the actions do not

Experiment: three runs that differ only inside the privileged reference — replies intact, replies deleted, replies moved to another task.



−2.9
replies deleted

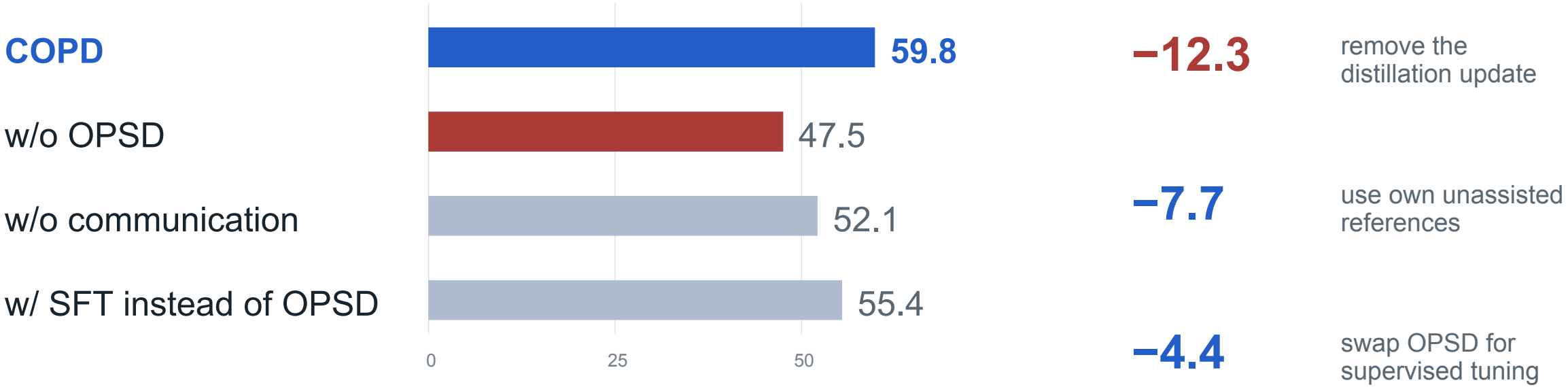
−5.2
replies moved to another task

The replies are part of the supervision, and they only help when they fit the task.

Figure 5

Every component carries part of the gain

Experiment: remove one piece at a time — the update, the privileged references, the loss form — and rescore every arm without an adviser.



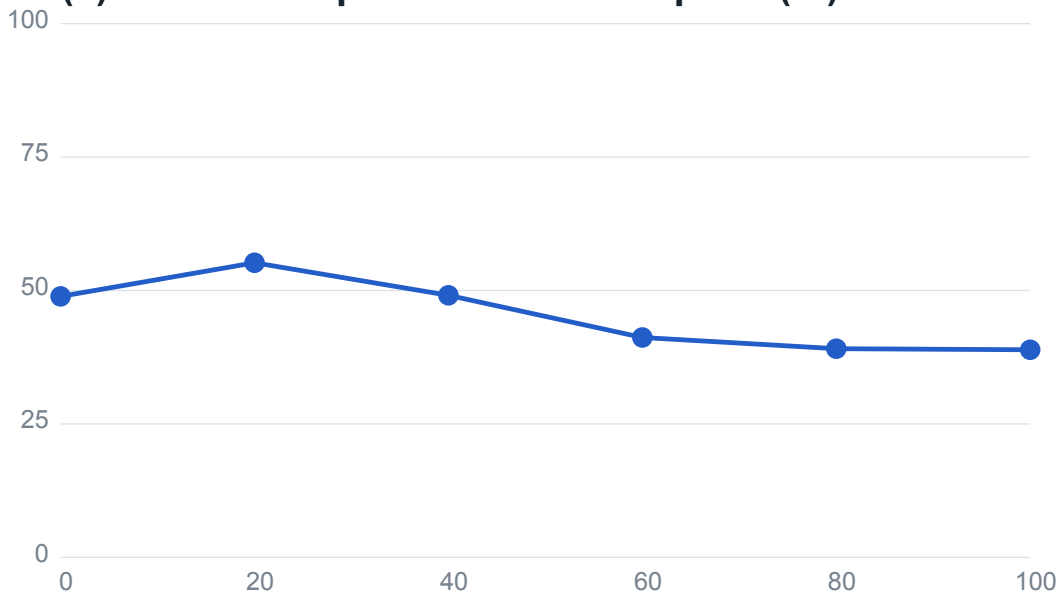
The update carries most of the gain, and the channel alone does not substitute.

Figure 7

The student asks less often and asks better

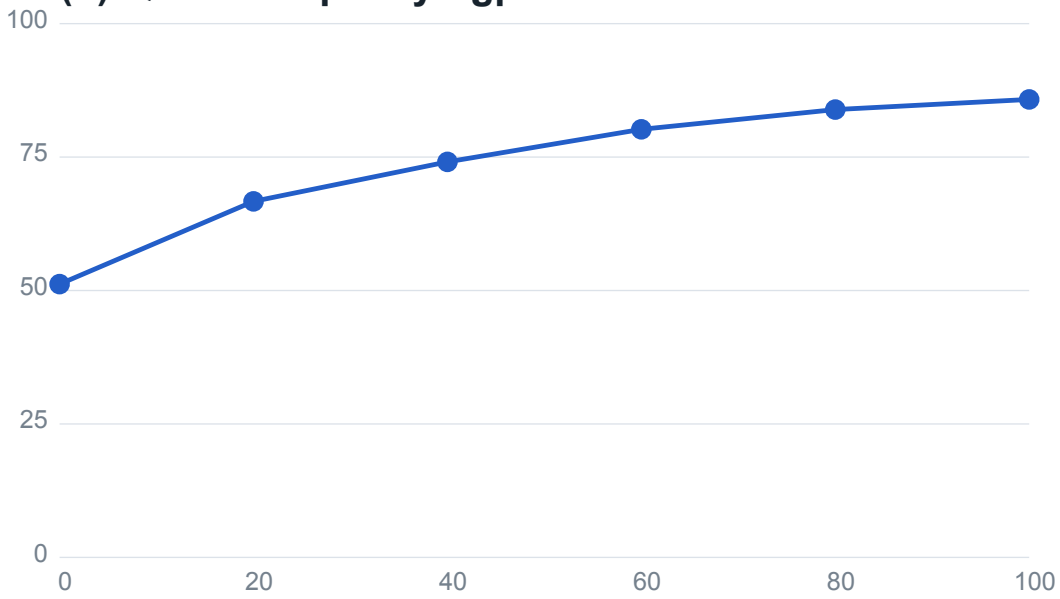
Experiment: measure ask rate and question quality on fixed probe tasks every twenty updates of one run.

(a) Ask rate · episodes with a request (%)



48.9 → 55.2 → 38.9

(b) Question quality · gpt-6-astra rubric



51.2 → 85.8, most of the rise by update 40

Training teaches when to ask and how to ask, not to stop asking.

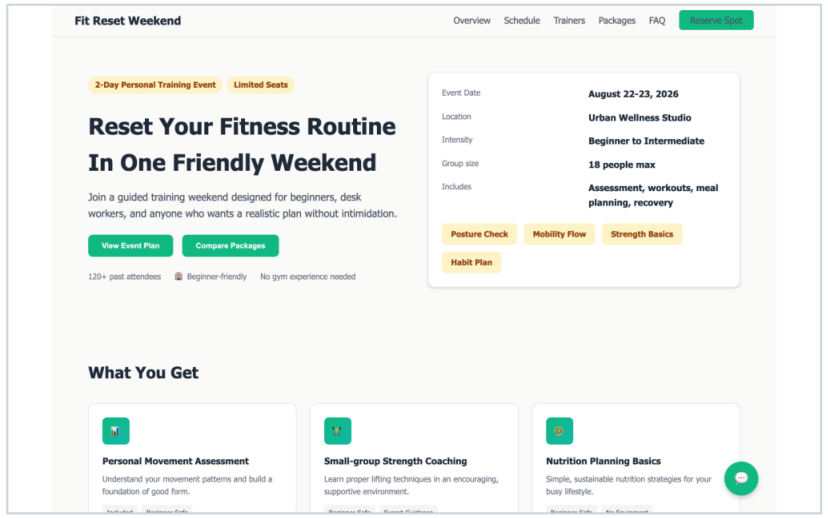
Figure 6

One brief, delivered twice

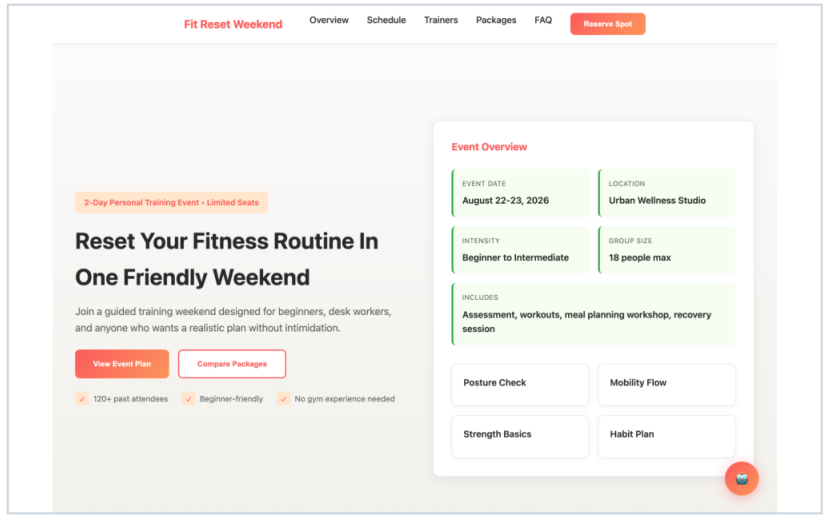
Experiment: one Visual brief executed twice by the trained policy, with the adviser available and with the channel removed.

Task: Build a two-day training-event page, delivered with the adviser available and with the channel removed.

Adviser available



Adviser disabled



The two pages differ in detail and agree in structure.

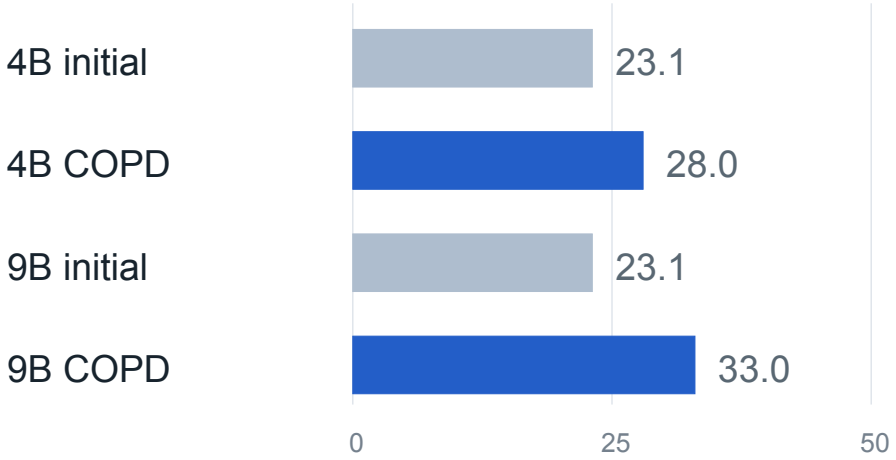
A two-day training-event page, and both runs return a finished page without access at test time.

Figure 8

Both students improve on both benchmarks

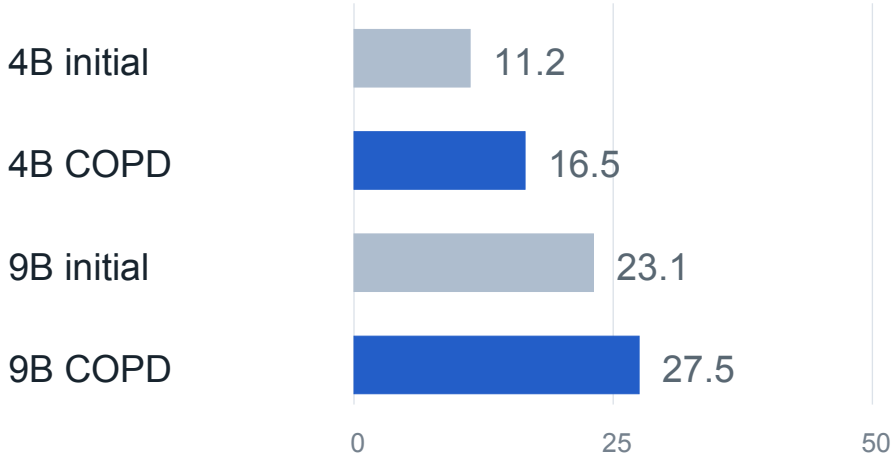
Experiment: both scales on two held-out benchmarks with no adviser at test time, compared with the same students before training.

Terminal-Bench 2.1



larger same-generation reference: 33.8

SWE-Bench Pro V2



larger same-generation reference: 27.4

The gain transfers to tasks and environments the student never trained on.

Table 2

Takeaways

01 **A black-box teacher can be distilled through the student's own questions.**

The student asks on its own schedule, and successful consultations become supervision.

02 **Communication alone does not make the student independent.**

The guidance has to be task-relevant and internalized into the student policy.