DISTILLING ACTIONS FROM BLACK-BOX TEACHERS: ACTION-POSTERIOR ON-POLICY DISTILLATION FOR AGENTIC TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a language model with teacher supervision on states induced by the model itself. Standard OPD relies on the teacher's next-token logits, which provide a rich supervision signal when available. However, many of the strongest models are proprietary black boxes that expose only sampled text, preventing them from serving as OPD teachers. Token-level supervision is also poorly matched to interactive environments, where task outcomes are determined by executable actions: different responses may induce the same action, while similar responses may lead to different state transitions. We therefore argue that black-box distillation should recover the teacher's distribution over environment-defined decisions rather than imitate its surface text. We introduce Action-Posterior On-Policy Distillation (AP-OPD), a reward-free method that pushes repeated teacher samples through the environment's parser or tool schema, aggregates decision-equivalent responses, and estimates an action posterior at each student-visited state. Because these posteriors concentrate rapidly at most states but remain diffuse at a few decision points, a consensus-gated sampler stops when responses agree and allocates additional queries to ambiguous states. We derive the query-optimal oracle allocation for a prescribed estimation accuracy, whose plug-in form defines the sampler, and train the student model by matching the estimated posterior with action-level cross-entropy. Across ALFWorld, ScienceWorld, and τ^2-bench, AP-OPD improves over single-sample on-policy imitation by 12.6% relative on ALFWorld and over white-box OPD by 18.9%, while matching the ALFWorld success of ten-sample querying with 54% fewer teacher tokens. Using the black-box DeepSeek-V4-Flash teacher further improves a Qwen3-4B student by 25.1% on ALFWorld and 120.4% on ScienceWorld relative to Qwen3-32B.

1 INTRODUCTION

Interactive tasks turn language-model generation into sequential decision making. A model issues commands or tool calls, observes their consequences, and conditions its next decision on the resulting interaction history. An early mistake therefore changes not only the immediate outcome but also every subsequent state presented to the model. Supervised fine-tuning on expert trajectories provides no guidance once the student's own decisions carry it beyond the states visited by the expert. On-policy distillation (OPD) addresses this distribution mismatch by rolling out the student model and asking a teacher to supervise the states that the student actually encounters (Ross et al., 2011; Agarwal et al., 2024; Gu et al., 2024). Standard OPD obtains this supervision from the teacher's next-token distribution, providing a dense learning signal without relying on task rewards.

Logit supervision is highly informative when it is available. The difficulty is that many of the strongest available teachers are proprietary models exposed only through sampling APIs. Because these models return generated text rather than next-token probabilities, the standard OPD objective excludes precisely the teachers one would most like to use. Recovering supervision from samples is possible, but the obvious alternatives discard important information. Imitating a single response at each state gives the student one noisy draw of the teacher's decision and hides how uncertain it is. Repeated sampling reveals variation in the teacher's behaviour, but in text space this variation remains entangled with surface realisation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

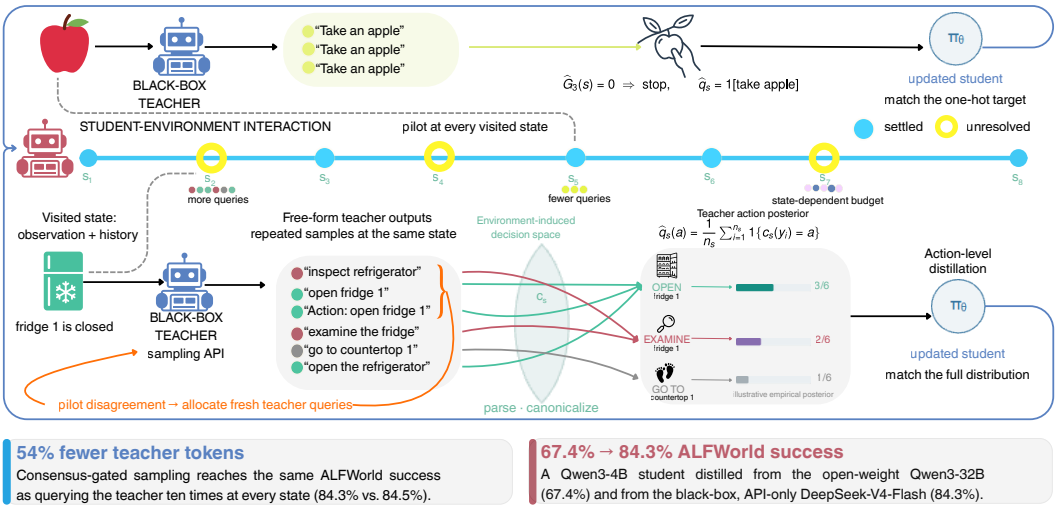


Figure 1: Overview of AP-OPD. Every state the student visits receives a pilot of three teacher samples. A unanimous pilot settles the state with a one-hot target (top); a disagreeing pilot triggers further queries (bottom). The environment's parser c_s maps paraphrased responses to the same executable action, and the resulting empirical posterior (illustrative) supervises the student. Bottom strip: key results on ALFWorld.

Interactive environments provide a more direct supervision space. Their command parsers and tool schemas expose the actions that can affect the environment, allowing teacher responses to be interpreted in terms of the decisions they induce rather than the tokens they contain. The relevant target is therefore not a distribution over textual realisations, but the teacher's posterior over executable actions. Recent black-box OPD methods estimate teacher preferences through semantic comparisons between sampled text chunks (Zhou et al., 2026a); in interactive tasks, the environment's action interface provides a task-aligned alternative.

Based on this observation, we introduce Action-Posterior On-Policy Distillation (AP-OPD). At each state visited by the student model, AP-OPD repeatedly queries the teacher, maps its responses through the environment's parser or tool schema, and estimates the resulting action posterior. The student model is then trained to match this posterior using an action-level cross-entropy objective, requiring neither teacher logits nor task rewards (Figure 1). Unlike single-sample imitation, the resulting target captures both the teacher's preferred decision and its uncertainty over plausible alternatives.

Repeated sampling introduces a new question: how many teacher queries should be allocated to each state? Empirically, the teacher's action posterior becomes concentrated after only a few samples at most states, while remaining diffuse at a small number of consequential decision points. AP-OPD therefore uses a consensus-gated sampler that stops querying when a short pilot produces sufficient agreement and allocates additional samples to unresolved states. We show that, under our target accuracy criterion, the oracle allocation of teacher queries follows a square-root rule in the teacher's disagreement, and the sampler implements its plug-in form. We further bound the probability of prematurely stopping at a state with genuine teacher disagreement and show that it decreases geometrically with the pilot size.

We evaluate AP-OPD on ALFWorld, ScienceWorld, and τ^2-bench (Shridhar et al., 2021; Wang et al., 2022; Barres et al., 2025), using Qwen3-1.7B and Qwen3-4B as student models and four teachers: the open-weight Qwen3-32B and three proprietary models accessible only through APIs. The results support two main conclusions. First, distributional action supervision improves over point supervision. With DeepSeek-V4-Flash as the teacher, AP-OPD raises the ALFWorld success rate of Qwen3-4B from 74.9% under single-sample on-policy imitation to 84.3%; with Qwen3-32B, for which white-box OPD is available, it improves over white-box OPD from 56.7% to 67.4%. Second, access to stronger black-box teachers translates into stronger student models. Replacing Qwen3-32B with DeepSeek-V4-Flash raises the same student's ALFWorld success rate from 67.4% to 84.3% and its ScienceWorld score from 29.4 to 64.8, while consensus-gated sampling matches the success of uniform ten-sample querying on ALFWorld with 54% fewer teacher tokens.

Our contributions are as follows.

- We introduce AP-OPD, an on-policy method for distilling black-box teachers into language models through executable action distributions, requiring neither teacher logits nor task rewards.
- We formulate repeated teacher sampling as action-posterior estimation and develop a consensus-gated allocation rule that concentrates queries on unresolved decisions. We derive the query-optimal oracle allocation that motivates it and bound its probability of premature stopping.
- Across three interactive benchmarks, two student models, and four teachers, AP-OPD improves over single-sample on-policy imitation and white-box OPD, enables distillation from stronger API-only teachers, and reduces teacher-query cost.

2 Related Work

On-policy and black-box distillation. On-policy distillation labels the states a learner actually visits (Ross et al., 2011), and for language models it trains students on their own generations under token-level divergences to the teacher (Agarwal et al., 2024; Gu et al., 2024). SAGE-OPD extends it to multi-turn agentic interaction by deciding for each turn how strongly token-level supervision is applied (Zhou et al., 2026b). All of these require the teacher's token distribution. Without logits, sequence-level distillation fine-tunes the student on teacher outputs (Kim & Rush, 2016), GAD trains a discriminator whose output serves as an on-policy reward (Ye et al., 2025), and OmniOPD compares sampled multi-token chunks by semantic similarity (Zhou et al., 2026a). These methods compare student and teacher in text space, whereas AP-OPD compares teacher responses through the actions they induce.

Aggregating and adaptively sampling model outputs. Self-consistency and semantic entropy group sampled outputs into equivalence classes (Wang et al., 2023; Kuhn et al., 2023), and adaptive variants stop sampling once the answers agree (Aggarwal et al., 2023; Li et al., 2024). They aggregate samples at inference time; AP-OPD aggregates teacher samples into training targets, defines equivalence through the environment's action interface, and allocates samples across states. Appendix A discusses related work in more detail.

3 Action-Posterior On-Policy Distillation

AP-OPD converts samples from a black-box teacher into distributional supervision over executable actions. During each on-policy rollout, the student determines the states on which supervision is collected. At every visited state, repeated teacher responses are mapped through the environment interface to form an empirical action posterior, which is then distilled into the student using canonical action strings. To control query cost, a consensus-gated sampler terminates sampling when the teacher's decisions agree and allocates additional queries to states with greater disagreement. We next describe the action posterior, the distillation objective, and the adaptive sampling rule.

3.1 Action Posterior

At each state s of an interactive task, the model produces a response $y \in \mathcal{Y}$, which the environment parses into an executable action. We denote this mapping by $c_s(y) \in \mathcal{A}(s)$. Each action a is represented by a canonical string $\bar{y}(a)$, such as `<action>open fridge 1</action>` followed by the end-of-sequence token. In practice, both c_s and $\bar{y}$ are specified by the environment's command grammar or tool schema.

The black-box teacher is a conditional generator $p_T(y \mid s)$ that can be sampled under a fixed decoding configuration but does not expose its logits. We assume that repeated queries at the same state are independent. The teacher's action posterior is the distribution induced by its parsable responses:

$$q_s(a) = \Pr_{Y \sim p_T(\cdot|s)} \big[c_s(Y) = a \;\big|\; c_s(Y) \in \mathcal{A}(s) \big]. \tag{1}$$

If a response cannot be parsed, we query the teacher again. Thus, the sample counts r and n_s below refer to parsable responses, each of which is a draw from q_s. Unparsable calls are excluded from the posterior estimate but included in the reported teacher-query cost $C_s \geq n_s$ (Appendix B). The student is an autoregressive model $p_\theta(y \mid s)$ whose induced action policy is defined analogously:

$$\pi_\theta(a \mid s) = \frac{1}{Z_\theta(s)} \sum_{y : c_s(y) = a} p_\theta(y \mid s), \qquad Z_\theta(s) = \Pr_{Y \sim p_\theta(\cdot \mid s)} \big[c_s(Y) \in \mathcal{A}(s)\big]. \tag{2}$$

3.2 Action-Level Distillation

In each training round, the current student interacts with the environment. Every visited state receives a pilot set of teacher queries, while only unresolved states receive additional queries. At state s, we collect n_s parsable teacher responses, with n_s chosen as described in Section 3.3. Their empirical frequencies define $\widehat{q}_s(a)$. We then train the student with the action-level cross-entropy

$$\mathcal{L}(\theta; s) = -\sum_a \widehat{q}_s(a) \log p_\theta\big(\bar{y}(a) \mid s\big), \tag{3}$$

along with a KL penalty to the initial student (Appendix B). Equation (3) places the empirical action posterior on canonical output strings. In the unconstrained distributional limit and without the KL penalty, its optimum assigns probability $\widehat{q}_s(a)$ to the canonical representation of each observed action. Proposition 1 further shows that this objective upper-bounds the KL divergence from the teacher's empirical action posterior to the student's induced action policy.

When $n_s = 1$, the target reduces to a hard action label; when $n_s > 1$, it preserves the teacher's uncertainty across actions. During both training and evaluation, the student predicts an action directly from the interaction history, and only the executable part of each teacher response enters the training target.

3.3 Consensus-Gated Sampling

A fixed query depth spends equally on states at which the teacher is certain and states at which its decisions vary. We quantify disagreement at state s by the probability that two independent teacher samples induce different actions:

$$G(s) = 1 - \sum_a q_s(a)^2. \tag{4}$$

AP-OPD allocates queries in two stages. It first draws a pilot of r responses at every state. If all pilot responses induce the same action, the state is settled and receives no further queries. Otherwise, additional queries are allocated in proportion to $\sqrt{\widehat{G}_r(s)}$, scaled by an allocation parameter ϵ and capped at $n_{\max}$ per state. Here $\widehat{G}_r(s)$ is the unbiased pairwise estimator of $G(s)$:

$$\widehat{G}_r(s) = \frac{1}{r(r-1)} \sum_{i \neq j} \mathbf{1}\{c_s(y_i) \neq c_s(y_j)\} = 1 - \sum_a \frac{N_a(N_a - 1)}{r(r-1)}, \tag{5}$$

where N_a is the number of pilot responses that induce action a. For $r = 3$, $\widehat{G}_r(s)$ takes one of three values, 0, $2/3$, or 1, yielding a simple three-level rule: stop, allocate a moderate number of queries, or sample up to the cap. The final posterior estimate pools the pilot and all additional samples.

Section 4.3 derives the oracle allocation approximated by this plug-in rule. Appendix F.6 bounds the bias introduced by reusing the pilot samples in the final estimate. The full procedure appears in Algorithm 1 in Appendix C, and Appendix B describes a fixed-budget variant.

4 Statistical Analysis

We analyse the action-level objective, the accuracy with which repeated queries recover the teacher posterior, and the optimal allocation of a finite query budget. Throughout, teacher samples are independent at each state and c_s is deterministic. All proofs are deferred to Appendix F.

4.1 Bounding the Action-Level KL

Proposition 1. *For every state s and student p_θ,*

$$\mathcal{L}(\theta; s) = H(\widehat{q}_s) + \mathrm{KL}\big(\widehat{q}_s \,\|\, \pi_\theta(\cdot \mid s)\big) + \sum_a \widehat{q}_s(a) \log \frac{\pi_\theta(a \mid s)}{p_\theta(\bar{y}(a) \mid s)} \;\geq\; H(\widehat{q}_s) + \mathrm{KL}\big(\widehat{q}_s \,\|\, \pi_\theta(\cdot \mid s)\big), \tag{6}$$

with equality if and only if $p_\theta(\bar{y}(a) \mid s) = \pi_\theta(a \mid s)$ for every action a in the support of $\widehat{q}_s$.

Because the entropy term is independent of θ, minimising Eq. (3) minimises an upper bound on the action-level KL divergence. The gap captures probability assigned to non-canonical realisations of the observed actions and to unparsable outputs; it vanishes when the student emits parsable actions in canonical form.

4.2 Estimating the Action Posterior

Theorem 2. *Let $K_s \geq 2$ be the number of actions in $\mathcal{A}(s)$. For a fixed number n of independent teacher samples, with probability at least $1 - \delta$,*

$$\mathrm{TV}(\widehat{q}_s, q_s) \leq \sqrt{\frac{\log(2^{K_s} - 2) + \log(1/\delta)}{2n}}. \tag{7}$$

This result follows from concentration for empirical distributions (Weissman et al., 2003). Crucially, the bound depends on the number of actions induced by the environment interface, rather than on the potentially much larger number of textual realisations. In the worst case, estimating a distribution over M surface strings to total-variation error ϵ requires $\Theta(M/\epsilon^2)$ samples, compared with $\Theta(K_s/\epsilon^2)$ for the action posterior (Han et al., 2015). Posterior concentration is captured more finely by $G(s)$ in the analysis below. Appendix F extends the result to an imperfect parser.

Theorem 2 assumes a fixed sample size. The adaptive estimator in Section 3.3 incurs a stopping bias because its final sample size depends on the pilot observations. This bias is bounded in Appendix F.6 and vanishes whenever the teacher is deterministic at the state.

4.3 Allocating Teacher Queries

Lemma 3. *Fix a state s and let $g_a = -\nabla_\theta \log p_\theta(\bar{y}(a) \mid s)$. For a fixed number n of independent teacher samples, the stochastic gradient $\widehat{g}_s = \sum_a \widehat{q}_s(a) g_a$ is unbiased for $\bar{g}_s = \sum_a q_s(a) g_a$, and*

$$\operatorname{tr} \mathrm{Cov}(\widehat{g}_s) = \frac{1}{n} \sum_a q_s(a) \|g_a - \bar{g}_s\|^2 \leq \frac{D_s^2 G(s)}{2n}, \qquad D_s = \max_{a,b} \|g_a - g_b\|. \tag{8}$$

For one-hot g_a, the left-hand side is $\mathbb{E}\|\widehat{q}_s - q_s\|_2^2 = G(s)/n$.

Teacher sampling therefore contributes gradient noise proportional to $G(s)/n_s$. Query allocation determines how this noise is distributed across the states encountered during training.

Proposition 4. *Let states $i = 1, \ldots, m$ have known disagreements G_i and weights $\omega_i > 0$. Treating the sample counts n_i as continuous, the allocation that minimises $\sum_i \omega_i G_i / n_i$ under the budget $\sum_i n_i = N$ satisfies $n_i \propto \sqrt{\omega_i G_i}$ (Neyman, 1992). Equivalently, the minimum number of queries required to achieve $m^{-1} \sum_i \omega_i G_i / n_i \leq \epsilon^2$ is obtained by*

$$n_i^\star = \frac{\sqrt{\omega_i G_i} \sum_j \sqrt{\omega_j G_j}}{m\epsilon^2}, \qquad N^\star(\epsilon) = \frac{\big(\sum_j \sqrt{\omega_j G_j}\big)^2}{m\epsilon^2} \;\leq\; \frac{\sum_j \omega_j G_j}{\epsilon^2} = N_{\mathrm{unif}}(\epsilon), \tag{9}$$

where $N_{\mathrm{unif}}(\epsilon)$ is the cost of a uniform-depth allocation at the same accuracy. States with $G_i = 0$ receive no queries under the oracle allocation.

Setting $\omega_i = D_i^2/2$ recovers the variance bound in Lemma 3, while $\omega_i = 1$ yields the posterior mean-squared error. The sampler in Section 3.3 is a clipped plug-in approximation to the latter allocation: it replaces G_i with its pilot estimate, assigns at least r queries to every state, and caps the total at $n_{\max}$. This choice is appropriate when the gradient spreads D_s are comparable across states.

Proposition 4 thus motivates the allocation rule without asserting that the finite-pilot procedure is itself optimal.

The ratio $N^\star/N_{\mathrm{unif}}$ becomes small when disagreement is concentrated on a few states. It can be estimated from the pilot before additional queries are allocated within each training round. The remaining risk is that a genuinely uncertain state produces a unanimous pilot and is settled prematurely.

Lemma 5. *For $r \geq 2$, the probability that all pilot samples at state s induce the same action is*

$$\sum_a q_s(a)^r \;\leq\; \big(1 - G(s)\big)^{r/2}, \tag{10}$$

and equals one if and only if $G(s) = 0$.

Thus, the probability of premature stopping is bounded by a quantity that decreases with both the pilot size and the teacher's disagreement. The KL penalty to the initial student further regularises learning from one-hot targets at settled states (Appendix B). Appendix F.8 gives a PAC bound on the resulting action-level KL divergence.

5 Experiments

5.1 Experimental Setup

Benchmarks. ALFWorld evaluates household tasks grounded in textual environments (Shridhar et al., 2021), and ScienceWorld evaluates scientific reasoning across diverse task families (Wang et al., 2022). On τ^2-bench (Barres et al., 2025), we use the telecom domain in its solo mode, in which the agent resolves the task through tool calls with full control of the environment and without a simulated user, isolating single-agent tool use. We report success rate for ALFWorld, normalized score and success rate for ScienceWorld, and pass^1 for τ^2-bench (Table 9).

Models. The students are Qwen3-4B and Qwen3-1.7B. We use four teachers: three accessible only through sampling APIs, DeepSeek-V4-Flash, Gemini-3.8-Flash, and GPT-5.6-Luna, and the open-weight Qwen3-32B, for which logit-based on-policy distillation provides a white-box reference. All policies, including teacher rollouts, act through the same environment interface at evaluation.

Canonicalization and sampling. ALFWorld and ScienceWorld expose a fixed command grammar, so c_s is a deterministic parser followed by normalization of casing, whitespace, and object aliases. On τ^2-bench, tool calls are canonicalized by tool name and normalized arguments. Students generate actions without extended reasoning (thinking disabled). AP-OPD uses the pilot size r, cap $n_{\max}$, and target ϵ in Table 9.

Baselines. We compare off-policy imitation of teacher trajectories (SFT), on-policy imitation of a single teacher sample per state (OPD$_{n=1}$), and a matched OmniOPD implementation (Zhou et al., 2026a). For the open-weight teacher, we also report white-box on-policy distillation with teacher token log-probabilities (OPD$_{\text{logit}}$). Teacher rollouts are reported as a reference. All methods start from the same student and use the same training schedule, and all results are averaged over 3 training seeds.

5.2 Black-Box Teachers Transfer into Small Models

Across both students and all four teachers, AP-OPD outperforms every baseline on all three benchmarks and on average, and OmniOPD is the strongest baseline on average in every group (Table 1). Four comparisons stand out.

Against single-sample imitation. OPD$_{n=1}$ queries the teacher on the student's states but distils one response per state. Replacing it with the action posterior improves the ALFWorld success of Qwen3-4B by 10.1–35.3% relative across teachers, for example from 74.9% to 84.3% with DeepSeek-V4-Flash, its ScienceWorld score by 18.4–25.6%, and its τ^2-bench pass^1 by 8.7–10.9 points. Table 2 separates where this gain comes from. With the same cached teacher samples, aggregating five samples into their majority action improves on a single sample by 4.5 points on ALFWorld and 6.1

Student	Teacher	Method	ALFWorld Success (%)	ScienceWorld Score	ScienceWorld Success (%)	τ^2-bench Pass[1] (%)	Average (%)
		Reference models (teacher rollouts)					
–	Qwen3-32B	Base	66.4	28.8	12.0	75.3	56.8
–	DeepSeek-V4-Flash	Base	83.6	66.1	45.0	99.2	83.0
–	Gemini-3.8-Flash	Base	98.5	80.0	64.0	100.0	92.8
–	GPT-5.6-Luna	Base	79.1	56.0	38.0	99.2	78.1
		Student: Qwen3-1.7B					
Qwen3-1.7B	–	Base	1.5	10.9	1.8	8.7	7.0
Qwen3-1.7B	Qwen3-32B	SFT	38.8 $\pm$ 1.3	12.5 $\pm$ 1.6	2.1 $\pm$ 1.8	17.4 $\pm$ 2.2	22.9
		$\text{OPD}_{n=1}$	39.8 $\pm$ 2.2	13.1 $\pm$ 2.2	2.2 $\pm$ 1.2	26.1 $\pm$ 3.8	26.3
		$\text{OPD}_{\text{logit}}$	45.8 $\pm$ 1.9	15.4 $\pm$ 1.7	2.7 $\pm$ 1.2	21.7 $\pm$ 3.8	27.6
		OmniOPD	48.0 $\pm$ 1.1	18.2 $\pm$ 2.4	3.7 $\pm$ 0.6	30.4 $\pm$ 5.8	32.2
		AP-OPD (ours)	**52.2** $\pm$ 1.5	**21.4** $\pm$ 2.1	**4.0** $\pm$ 1.8	**34.8** $\pm$ 2.2	**36.1**
		Student: Qwen3-4B					
Qwen3-4B	–	Base	35.0	19.3	7.0	19.6	24.6
Qwen3-4B	Qwen3-32B	SFT	48.5 $\pm$ 1.3	22.1 $\pm$ 1.8	8.5 $\pm$ 1.2	43.5 $\pm$ 2.2	38.0
		$\text{OPD}_{n=1}$	49.8 $\pm$ 0.9	23.4 $\pm$ 2.4	8.7 $\pm$ 1.6	54.3 $\pm$ 4.3	42.5
		$\text{OPD}_{\text{logit}}$	56.7 $\pm$ 0.7	25.8 $\pm$ 1.9	10.4 $\pm$ 1.3	50.0 $\pm$ 3.8	44.2
		OmniOPD	63.4 $\pm$ 1.3	27.2 $\pm$ 2.7	11.0 $\pm$ 1.5	58.7 $\pm$ 4.3	49.8
		AP-OPD (ours)	**67.4** $\pm$ 1.7	**29.4** $\pm$ 1.6	**12.5** $\pm$ 1.4	**65.2** $\pm$ 3.8	**54.0**
Qwen3-4B	DeepSeek-V4-Flash	SFT	73.4 $\pm$ 1.9	52.4 $\pm$ 2.3	28.5 $\pm$ 1.7	67.4 $\pm$ 4.3	64.4
		$\text{OPD}_{n=1}$	74.9 $\pm$ 1.1	54.1 $\pm$ 2.7	29.5 $\pm$ 1.3	76.1 $\pm$ 4.3	68.4
		OmniOPD	80.3 $\pm$ 1.1	60.2 $\pm$ 1.9	37.1 $\pm$ 0.6	78.3 $\pm$ 3.8	72.9
		AP-OPD (ours)	**84.3** $\pm$ 2.2	**64.8** $\pm$ 2.2	**44.2** $\pm$ 1.9	**84.8** $\pm$ 2.2	**78.0**
Qwen3-4B	Gemini-3.8-Flash	SFT	80.1 $\pm$ 1.1	58.2 $\pm$ 2.6	35.4 $\pm$ 1.2	71.7 $\pm$ 4.3	70.0
		$\text{OPD}_{n=1}$	84.8 $\pm$ 0.9	61.5 $\pm$ 1.8	41.2 $\pm$ 0.8	80.4 $\pm$ 2.2	75.6
		OmniOPD	86.3 $\pm$ 1.1	69.1 $\pm$ 2.4	44.3 $\pm$ 1.4	78.3 $\pm$ 3.8	77.9
		AP-OPD (ours)	**94.0** $\pm$ 1.5	**75.4** $\pm$ 1.7	**58.2** $\pm$ 1.3	**89.1** $\pm$ 4.3	**86.2**
Qwen3-4B	GPT-5.6-Luna	SFT	70.9 $\pm$ 2.0	44.2 $\pm$ 2.8	26.8 $\pm$ 1.8	65.2 $\pm$ 4.3	60.1
		$\text{OPD}_{n=1}$	72.1 $\pm$ 1.9	46.8 $\pm$ 2.1	27.6 $\pm$ 1.7	73.9 $\pm$ 4.3	64.3
		OmniOPD	77.1 $\pm$ 1.1	53.5 $\pm$ 2.5	33.4 $\pm$ 1.6	76.1 $\pm$ 2.2	68.9
		AP-OPD (ours)	**79.4** $\pm$ 2.2	**55.4** $\pm$ 1.9	**37.2** $\pm$ 1.5	**82.6** $\pm$ 5.8	**72.5**

Table 1: Main results on ALFWorld, ScienceWorld, and τ^2-bench (telecom domain, solo mode). Average is the mean of ALFWorld success, ScienceWorld score, and τ^2-bench pass[1]. Reference rows report each teacher's own rollouts. SFT: off-policy imitation of teacher trajectories; $\text{OPD}_{\text{logit}}$: white-box on-policy distillation with teacher log-probabilities, available only for the open-weight teacher; $\text{OPD}_{n=1}$: on-policy imitation of a single teacher sample per state; OmniOPD: generic black-box OPD (Zhou et al., 2026a). Trained rows report means $\pm$ standard deviations over 3 training seeds. Bold marks the best result within each student–teacher group.

on ScienceWorld, keeping the full distribution adds 2.7 and 2.3, and grouping the samples by the action they induce rather than by their text adds 5.0 and 4.3 over a surface-text target. The gain does not come from the larger budget: spending the same teacher tokens on five times as many single-sample states leaves success unchanged (Appendix D.2).

Target	Samples per state	ALFWorld (success, %)	ScienceWorld (score)
Canonical action, single sample ($\text{OPD}_{n=1}$)	1	74.9 $\pm$ 1.1	54.1 $\pm$ 2.7
Surface text, all samples	5	77.1 $\pm$ 0.7	58.2 $\pm$ 1.9
Canonical action, majority	5	79.4 $\pm$ 1.3	60.2 $\pm$ 1.6
Canonical action, distribution (AP-OPD)	5	82.1 $\pm$ 1.3	62.5 $\pm$ 2.3

Table 2: Components of the target (Qwen3-4B, DeepSeek-V4-Flash; uniform five-sample rows share the same teacher samples; 3 training seeds).

Against black-box OPD. AP-OPD improves on OmniOPD by 2.3–7.7 points of ALFWorld success, 1.9–6.3 points of ScienceWorld score, and 6.5–10.8 points of τ^2-bench pass[1] for Qwen3-4B, and by 4.2, 3.2, and 4.4 points for Qwen3-1.7B. Both methods query the same teacher without logits; they differ in whether teacher responses are compared through semantic similarity or through the actions they induce.

Against white-box OPD. With the open-weight Qwen3-32B teacher, where teacher logits are available, AP-OPD improves on logit-based OPD by 10.7 points of ALFWorld success for Qwen3-4B (56.7% to 67.4%, 18.9% relative) and by 6.4 points for Qwen3-1.7B, and by 9.8 and 8.5 points on average, without using logits. Logit-based OPD improves on single-sample imitation on ALFWorld and ScienceWorld but not on τ^2-bench, and OmniOPD outperforms it on average, in line with the gains over white-box OPD that the OmniOPD authors report on mathematical reasoning (Zhou et al., 2026a).

Access to stronger teachers. Because it needs only sampled text, AP-OPD can distil from teachers that logit-based OPD cannot use. Replacing Qwen3-32B with DeepSeek-V4-Flash raises the ALFWorld success of Qwen3-4B from 67.4% to 84.3%, its ScienceWorld score from 29.4 to 64.8, and its τ^2-bench pass[1] from 65.2% to 84.8%; with Gemini-3.8-Flash, the student reaches 94.0%, 75.4, and 89.1%.

The distilled student approaches the teachers it learns from. With DeepSeek-V4-Flash and GPT-5.6-Luna, it matches the teacher's ALFWorld success (84.3% against 83.6%, and 79.4% against 79.1%) and stays within 1.3 points of its ScienceWorld score; with Qwen3-32B, it matches the teacher on ALFWorld and ScienceWorld; with the strongest teacher, Gemini-3.8-Flash, gaps of 4.5 points on ALFWorld and 4.6 points of ScienceWorld score remain. On τ^2-bench, where the API teachers nearly saturate the domain, the students remain below their teachers. Matching the teacher's own rollouts requires matching its whole action distribution, which a student sampled at the same temperature then reproduces; the several samples per state make this target far less noisy than a single draw. Single-sample imitation, which distils individual draws, stays well below the teacher, at 49.8% against 66.4% with Qwen3-32B.

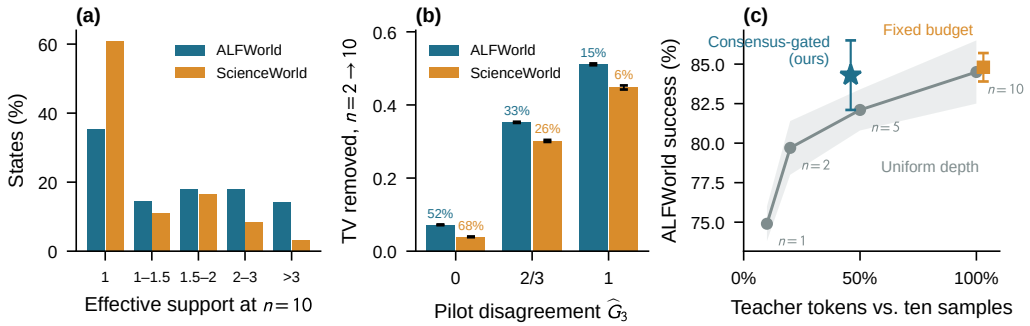


Figure 2: Teacher posteriors are narrow at most states, so stopping on consensus halves the teacher cost (DeepSeek-V4-Flash teacher). (a) Effective support $1/\|\widehat{q}\|_2^2$ of the ten-sample posterior at 32,005 ALFWorld and 16,001 ScienceWorld states. (b) Total-variation distance between the two-sample and ten-sample posteriors, grouped by the pilot disagreement $\widehat{G}_3$ of the first three samples; labels give the share of states in each group, and error bars are standard errors. (c) ALFWorld success against the teacher tokens of training, relative to ten samples per state, for uniform depths n=1, 2, 5, 10, the fixed-budget schedule, and consensus gating (4.45 samples per state); shading and error bars show standard deviations over 3 training seeds.

5.3 Teacher Posteriors Are Narrow at Most States and Broad at a Few

Table 3 compares uniform depths $n \in \{1, 2, 5, 10\}$ with the two-stage schedules of Section 3.3.

Where the samples go matters more than how many are drawn. With 4.45 samples per state, fewer than uniform sampling with five, the consensus-gated sampler reaches 84.3% ALFWorld success against 82.1%; it matches uniform sampling with ten samples (84.5%) while using 54% fewer

Sampling	$\bar{n}$	#Actions	Entropy	Success (%)
Uniform	1	1.000	0.000	74.9 ± 1.1
Uniform	2	1.320	0.222	79.7 ± 1.7
Uniform	5	1.932	0.476	82.1 ± 1.3
Uniform	10	2.574	0.620	84.5 ± 2.0
Two-stage, fixed budget	10.00	1.815	0.331	84.8 ± 0.9
Consensus-gated (ours)	4.45	2.105	0.374	84.3 ± 2.2

Table 3: Sampling depth on ALFWorld with the Qwen3-4B student and the DeepSeek-V4-Flash teacher. $\bar{n}$ is the mean number of teacher samples per state, including pilot samples for the two-stage schedule. #Actions is the mean number of distinct canonical actions in the empirical posterior of a supervised state, and entropy is its mean entropy in nats. Success is reported as means ± standard deviations over 3 training seeds.

teacher tokens. Under uniform sampling, success rises steadily with depth, from 74.9% at $n{=}1$ to 79.7%, 82.1%, and 84.5% at $n{=}2$, 5, and 10, but the gain per additional query falls from 4.8 points for the first extra sample to below one point for each sample thereafter. The mean number of distinct actions per state hides a skewed distribution (Figure 2a). With ten samples, 35% of ALFWorld states and 61% of ScienceWorld states show a single action, whereas the effective support exceeds three at 14% and 3% of states. Additional samples matter almost only where the pilot disagrees (Figure 2b). At states whose first three samples agree, 52% of ALFWorld and 68% of ScienceWorld states, going from two to ten samples moves the posterior by 0.07 and 0.04 in total variation; at states whose three samples all differ, it moves it by 0.51 and 0.45. Stopping on consensus exploits this structure: its additional queries go to states whose pilots disagree, while states with a unanimous pilot remain one-hot.

5.4 Sampling Until the Decision Is Clear Reduces Teacher Cost

Because most states are narrow, stopping on a unanimous pilot should save many queries where the teacher is certain and few where it is not. In training on ALFWorld, the consensus-gated sampler matches the success of ten-sample querying (84.3% against 84.5%) with 4.45 samples per state and 54% fewer teacher tokens, and it exceeds uniform sampling with five samples, which costs more (Figure 2c; Table 7).

The savings come from where the queries go. At a matched number of teacher calls, pilot-based allocation produces posteriors whose majority action disagrees less often with a held-out reference than uniform sampling does, on all three benchmarks and at every budget we test (Figure 3; Appendix D.5). The pilot size trades savings against fidelity: a pilot of three saves 29% of teacher tokens on ALFWorld and 32% on τ^2-bench while changing the majority action at fewer than 0.5% of states, and the savings are smaller where the teacher disagrees more often, as on ScienceWorld (Table 8).

6 Discussion and Limitations

AP-OPD relies on an action interface to define equivalence, so its benefits are largest in environments with command grammars or tool schemas. For free-form replies it keeps each response separate, and environments that can be reset could instead define equivalence by the consequences of executing each response. Our students act without extended reasoning, and our evaluation covers Qwen3 students of up to 4B parameters and the solo mode of τ^2-bench; distilling reasoning traces alongside actions and extending the evaluation to user-facing dialogue are natural next steps.

7 Conclusion

Interactive environments define which language outputs count as the same decision. Pushing a black-box teacher's samples through this definition yields a distribution over decisions that is narrow at most states and broad at a few, which determines both what to distil and where to spend teacher queries. AP-OPD samples until the teacher's decision is clear and distils the result into students that approach, and on ALFWorld and ScienceWorld reach, the level of their teachers' own rollouts.

AI USAGE STATEMENT

Large language models are the subject of this work: the teachers and students in all experiments are language models, used as described in Section 5 and Appendix E. Separately, we used Large language models as writing and editing assistants to improve the organization, terminology, and presentation of this manuscript. The authors defined the method, conducted all experiments, verified the claims and numerical results, and take full responsibility for the manuscript. Experimental evidence comes from the verified runs reported in this paper.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13649.

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025. doi: 10.48550/arXiv.2506.07982. URL https://arxiv.org/abs/2506.07982.

Sanjiban Choudhury and Paloma Sodhi. Better than your teacher: LLM agents that learn from privileged AI feedback. In *International Conference on Learning Representations*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.08543.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2291–2295. IEEE, 2015.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, 2024.

Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics: Methodology and distribution*, pp. 123–150. Springer, 1992.

Liam Paninski. Variational minimax estimation of discrete distributions under KL loss. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 627–635, 2011.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022. doi: 10.18653/v1/2022.emnlp-main.775.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical report, Hewlett-Packard Laboratories, 2003.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025. URL https://arxiv.org/abs/2511.10643.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. OmniOPD: Logit-free on-policy distillation via speculative verification, 2026a. URL https://arxiv.org/abs/2606.01476.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. SAGE-OPD: Selective agent-guided intervention for multi-turn on-policy distillation. *arXiv preprint arXiv:2606.19659*, 2026b.

A Extended Related Work

On-policy distillation. DAgger established that an expert should label the states a learner actually visits rather than only the states of expert demonstrations (Ross et al., 2011). For language models, generalized knowledge distillation and MiniLLM train students on their own generations under token-level divergences to the teacher's distribution (Agarwal et al., 2024; Gu et al., 2024). In agentic tasks, LEAP relabels student trajectories with a teacher that has access to privileged state, and finds that students trained in this way can exceed their teachers (Choudhury & Sodhi, 2025). SAGE-OPD applies OPD to multi-turn interaction and decides for each turn whether, and how strongly, token-level teacher supervision is applied, weighting it by the teacher's predictive confidence (Zhou et al., 2026b). These methods either rely on the teacher's token distribution or supervise each state with a single teacher output. AP-OPD retains on-policy data collection but supervises each visited state with a sampled distribution over executable actions.

Black-box distillation. When only the teacher's text is available, the standard approach is sequence-level knowledge distillation, which fine-tunes the student on teacher-generated outputs (Kim & Rush, 2016). GAD makes black-box distillation on-policy by training a discriminator to distinguish student from teacher responses and using it as a reward for the student (Ye et al., 2025). OmniOPD estimates which continuation the teacher favours by comparing sampled multi-token chunks with a semantic similarity measure (Zhou et al., 2026a). Both compare the student with the teacher in text space. In interactive environments, AP-OPD instead compares teacher responses through the actions they induce, which separates paraphrase from disagreement without a learned discriminator or a similarity model.

Aggregating and adaptively sampling model outputs. Grouping samples into equivalence classes before forming a distribution over them has precedents. Self-consistency selects the most frequent answer among sampled reasoning paths (Wang et al., 2023), and semantic entropy measures uncertainty over clusters of generations that share a meaning (Kuhn et al., 2023; Farquhar et al., 2024). Adaptive-Consistency and Early-Stopping Self-Consistency reduce the number of samples by stopping once the sampled answers agree sufficiently (Aggarwal et al., 2023; Li et al., 2024), in the spirit of sequential analysis (Wald, 1945). These methods aggregate samples to answer a query or to estimate uncertainty at inference time. AP-OPD aggregates teacher samples to construct training targets, defines equivalence through the environment's action interface rather than by answer matching or semantic entailment, and allocates samples across states under an explicit accuracy target. Distilling a teacher's full action distribution rather than only its most likely action follows policy distillation (Rusu et al., 2015) and soft-target distillation (Hinton et al., 2015).

B Method and Implementation Details

B.1 Canonicalization and unparsable responses

The canonicalizer applies up to three stages.

1. **Parsing:** extract the executable portion of the response and map exact valid commands or tool calls to environment actions. Any other text in the response, including reasoning, is discarded.
2. **Normalization:** collapse formatting, casing, argument aliases, argument order, and other surface variations known to preserve the executed transition.
3. **Identity for free-form text:** natural-language replies, where an environment has them, are not merged; each sampled reply forms its own class.

When a teacher response cannot be parsed, it is excluded from the posterior and the teacher is queried again, so that the pilot and the further samples consist of r and $n_s - r$ parsable responses, each an independent draw from q_s in Eq. (1). All calls, including unparsable ones, count towards the teacher-query cost C_s reported in Section 5.4. A state for which no parsable response is obtained contributes no distillation loss. Canonicalization is fixed within each comparison and does not use evaluation rewards.

B.2 Full training objective

With $\mathcal{D}_k$ the states visited by the student in round k and p_{θ_0} the initial student, the training objective is

$$\widehat{\mathcal{J}}(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{s \in \mathcal{D}_k} \Big[\mathcal{L}(\theta; s) + \lambda \, \mathrm{KL}\big(p_\theta(\cdot \mid s) \,\|\, p_{\theta_0}(\cdot \mid s)\big) \Big], \tag{11}$$

with λ listed in Table 9. The objective uses neither environment rewards nor teacher logits.

The KL penalty and unobserved actions. A finite-sample target assigns zero probability to actions that the teacher takes rarely but that were not observed. Consider the per-state objective as a function of the student's output distribution p over sequences, $-\sum_y \tilde{q}_s(y) \log p(y) + \lambda \sum_y p(y) \log\big(p(y)/p_0(y)\big)$, where $\tilde{q}_s$ places $\widehat{q}_s(a)$ on $\bar{y}(a)$. With a multiplier μ for normalisation, stationarity gives $-\tilde{q}_s(y)/p(y) + \lambda\big(\log(p(y)/p_0(y)) + 1\big) + \mu = 0$. For every y outside the support of $\tilde{q}_s$, this yields $p(y) = p_0(y) \exp(-1 - \mu/\lambda)$, so responses the teacher was not observed to produce retain probability in proportion to the initial model rather than being driven to zero.

B.3 Consensus-gated sampling implementation

The pilot draws $r \geq 2$ responses per state. With $Z_i = \mathbf{e}_{c_s(y_i)}$ the one-hot vector of the i-th pilot action, the pairwise estimator of Eq. (5) can be written as

$$\widehat{G}_r(s) = 1 - \frac{1}{r(r-1)} \sum_{i \neq j} \langle Z_i, Z_j \rangle. \tag{12}$$

This estimator is unbiased for $G(s)$, and $\widehat{G}_r(s) = 0$ exactly when the pilot is unanimous. For $r = 3$, it equals 0, $2/3$, or 1 when the pilot contains one, two, or three distinct actions. Writing $\mathcal{D}_k$ for the states of the current round, the total number of samples at a state with a non-unanimous pilot is

$$n_s = \min\Big\{n_{\max},\, \max\Big\{r,\, \Big\lceil \frac{\sqrt{\widehat{G}_r(s)}\, \sum_{u\in\mathcal{D}_k} \sqrt{\widehat{G}_r(u)}}{|\mathcal{D}_k|\, \epsilon^2} \Big\rceil \Big\}\Big\}, \tag{13}$$

the plug-in form of Proposition 4 with $\omega_i = 1$; a state with a unanimous pilot keeps $n_s = r$. The r pilot queries are issued concurrently, with further calls only to replace unparsable responses, and all pilot and further calls are counted in the teacher-query cost, with the full prompt counted for each.

Fixed-budget variant. Instead of a target accuracy, the fixed-budget variant divides a preset second-stage budget N_k across the states of round k in proportion to

$$\omega_s = \frac{\sqrt{\max\{\widehat{G}_r(s), 0\} + \gamma}}{\sum_{u\in\mathcal{D}_k} \sqrt{\max\{\widehat{G}_r(u), 0\} + \gamma}}, \tag{14}$$

where the floor $\gamma > 0$ gives every state some second-stage samples. It therefore never settles a state after the pilot and does not reduce the total number of calls.

C Training Algorithm

Algorithm 1 Action-Posterior On-Policy Distillation (AP-OPD)

Require: Initial student p_{θ_0}; black-box teacher T; environment $\mathcal{E}$; canonicalizer c; pilot size r; cap $n_{\max}$; target accuracy ϵ; KL coefficient λ; training rounds K

for $k = 0, \ldots, K-1$ **do**
  Roll out p_{θ_k} in $\mathcal{E}$ and collect the visited states $\mathcal{D}_k$
  for all $s \in \mathcal{D}_k$ **do**
    Query $T(\cdot \mid s)$ r times concurrently, re-querying for unparsable responses, until r parsable pilot responses are obtained
    Canonicalize them and compute $\widehat{G}_r(s)$ by Eq. (5)
  end for
  for all $s \in \mathcal{D}_k$ **do**
    if $\widehat{G}_r(s) = 0$ **then**
      $n_s \leftarrow r$ ▷ unanimous pilot: settled
    else
      Compute n_s by Eq. (13) and draw $n_s - r$ further responses
    end if
    $\widehat{q}_s(a) \leftarrow$ share of parsable pilot and further responses with $c_s(y) = a$
    Cache the responses, canonical actions, and $\widehat{q}_s$
  end for
  Update θ_k on Eq. (11) to obtain θ_{k+1}
end for
return p_{θ_K}

Teacher responses and their canonicalized actions are cached so that objective, canonicalization, and stopping ablations reuse identical teacher samples; the replay in Table 8 is computed from these caches.

Statistic	ALFWorld	ScienceWorld
Effective support	1.89 ± 1.11	1.39 ± 0.68
Effective support, median / 90th pct.	1.52 / 3.52	1.00 / 2.27
States with a single action (%)	35.4	60.9
States with effective support ≥ 3 (%)	15.3	4.2

Table 4: Diagnostics of the empirical teacher posterior at supervised states with $n{=}10$ with the DeepSeek-V4-Flash teacher. Effective support is $1/\|\widehat{q}\|_2^2$. The first row reports the mean and standard deviation over supervised states; the remaining rows describe its distribution across states.

D Additional Results

D.1 Posterior diagnostics

D.2 Matching the teacher budget

Deeper sampling spends more teacher tokens per state, so its benefit could reflect a larger budget rather than a better target. Table 5 compares schedules with close to the same total teacher tokens on ALFWorld. The budget-matched single-sample run collects ten rather than two rollouts per round and therefore trains on about five times as many states, each supervised by one teacher sample. Spending the budget this way leaves success essentially unchanged (75.1% against 74.9%), whereas spending it on several samples per state raises success to 82.1% with uniform sampling and to 84.3% with consensus gating. The gain of multiple samples therefore does not come from the larger budget: once the target at each state is a single draw, additional states no longer help.

Schedule	Rollouts per round	Samples per state	Teacher tokens per run (M)	Success (%)
Single sample ($\text{OPD}_{n=1}$)	2	1	141.2	74.9 ± 1.1
Single sample, budget-matched	10	1	710.3	75.1 ± 1.7
Uniform $n{=}5$	2	5	706.2	82.1 ± 1.3
Consensus-gated (AP-OPD)	2	4.45	698.7	84.3 ± 2.2

Table 5: Schedules at a matched teacher budget on ALFWorld with the Qwen3-4B student and the DeepSeek-V4-Flash teacher. Tokens count the full prompt of every sample. Success is reported as means $\pm$ standard deviations over 3 training seeds.

D.3 Validity of student actions

Proposition 1 bounds the action-level KL by a loss whose gap includes the probability of unparsable outputs, so a distributional target could in principle trade action validity for fidelity to the teacher. Table 6 measures how often the Qwen3-4B student trained with the DeepSeek-V4-Flash teacher at uniform depth n emits an action that ALFWorld cannot execute, over 134 tasks and 8 evaluation seeds, paired by task and seed ($1{,}072$ episodes per arm). Invalid actions are rare at every depth, about 6% of all actions, and become rarer with more teacher samples per state: relative to $n{=}1$, they fall by 0.16 per episode at $n{=}5$ and by 0.19 at $n{=}10$, whereas $n{=}2$ is indistinguishable from a single sample.

Depth	Invalid per episode	Invalid actions (%)	Truncated (%)	Δ per episode	Paired t-test p	Wilcoxon p
$n{=}1$	1.625	6.33	21.5	–	–	–
$n{=}2$	1.618	6.52	19.1	-0.007	0.92	0.73
$n{=}5$	1.466	6.03	19.6	-0.159	0.0057	0.0024
$n{=}10$	1.431	5.77	19.1	-0.194	0.0011	0.00064

Table 6: Invalid actions of the Qwen3-4B student on ALFWorld with the DeepSeek-V4-Flash teacher, over 134 tasks and 8 evaluation seeds. Truncated episodes reach the horizon of 50 steps. Differences and tests are paired by task and seed ($1{,}072$ pairs) against $n{=}1$.

D.4 TEACHER COST

Sampling	Samples per state	Tokens per state	Tokens per run (M)	Success (%)
Uniform n=1	1	4,342	141.2	74.9 $\pm$ 1.1
Uniform n=2	2	8,685	282.5	79.7 $\pm$ 1.7
Uniform n=5	5	21,712	706.2	82.1 $\pm$ 1.3
Uniform n=10	10	43,423	1,412.4	84.5 $\pm$ 2.0
Two-stage, fixed budget	10.00	42,314	1,318.3	84.8 $\pm$ 0.9
Consensus-gated (ours)	4.51	20,683	698.7	84.3 $\pm$ 2.2

Table 7: Teacher tokens on ALFWorld with the DeepSeek-V4-Flash teacher. Tokens count the full prompt of every sample, with output tokens counted by the student tokenizer. Uniform rows use the mean prompt length of the recorded states (4,342 tokens per sample). The fixed-budget and consensus-gated rows are measured on the supervised states of their own training runs, and their samples per state count all teacher calls, including those that replace unparsable responses; for the consensus-gated run, the prompt averages 4,485 tokens per sample and ten samples would take 44,970 tokens per state. Success is reported as means $\pm$ standard deviations over 3 training seeds.

D.5 REPLAY OF THE STOPPING RULE AND HELD-OUT COMPARISON

Table 8 replays the stopping rule on recorded teacher samples with ten samples per state, excluding states at which none of the ten samples can be parsed. A state stops after a unanimous pilot of r samples and otherwise continues to all ten, so the cap of ten is a property of the recorded data rather than of the training configuration. The table reports how much the rule saves and how far stopping moves the target from the posterior that continuing to ten samples would give. On ScienceWorld, the trained consensus-gated sampler draws 8.19 samples per state and uses 45,612 teacher tokens per state, 18.5% fewer than the 55,990 that ten samples would cost on the same states (5,589 prompt tokens per sample).

Benchmark	Rule	Settled (%)	Tokens (k)	Δ tokens	TV	Mode (%)
ALFWorld	Ten samples	–	43.4	0	0	0
	Pilot r=2	55.8	25.2	−41.9%	0.072	2.3
	Pilot r=3 (ours)	45.1	30.9	−28.9%	0.032	0.3
	Pilot r=4	39.5	34.2	−21.3%	0.017	0.0
τ^2-bench	Ten samples	–	73.0	0	0	0
	Pilot r=2	56.5	40.0	−45.2%	0.068	2.2
	Pilot r=3 (ours)	46.1	49.4	−32.3%	0.033	0.4
	Pilot r=4	40.2	55.4	−24.1%	0.018	0.0
ScienceWorld	Ten samples	–	54.7	0	0	0
	Pilot r=2	33.1	39.8	−27.3%	0.034	1.1
	Pilot r=3 (ours)	27.6	43.9	−19.7%	0.016	0.2
	Pilot r=4	24.9	46.4	−15.1%	0.008	0.0

Table 8: Replay of the stopping rule on recorded teacher samples (ALFWorld: 32,005 states and ScienceWorld: 16,001 states with the DeepSeek-V4-Flash teacher; τ^2-bench telecom: 6,650 states with the Qwen3-32B teacher). Settled is the fraction of states whose pilot is unanimous; tokens are per state in thousands with the full prompt counted for every sample, and Δ is the change relative to ten samples. TV and Mode compare the posterior after stopping with the posterior from all ten samples: TV is their total-variation distance, and Mode is the percentage of states whose majority action differs.

Figure 3 reports the held-out comparison of Section 5.4 for all pilot sizes and both metrics. Each state's ten samples are split at random into a pool of seven, from which each schedule draws at the stated mean number of calls, and a reference of three, and results average 60 splits. The pilot-based schedule draws a pilot of r samples and allocates the remaining calls with the fixed-budget rule of Eq. (14) and floor $\gamma = 0.2$, so that both schedules spend exactly the stated budget. Comparing

against samples that neither schedule uses avoids favouring a schedule whose samples overlap with the reference.

The comparison uses states at which all ten recorded samples can be parsed (on τ^2-bench, 4,561 states with the Qwen3-32B teacher). Relative to uniform sampling at the same budget, majority-action disagreements with the reference fall by 1.0–4.9% in all 21 combinations of pilot size, budget, and benchmark, with 95% intervals that exclude zero except at six calls on ScienceWorld. Total-variation distance falls in 18 of the 21, by up to 4.5%; it rises by 0.7% with a pilot of two at three calls on ALFWorld and is unchanged within its interval in the two remaining cases, a pilot of two at three calls on ScienceWorld and at six calls on τ^2-bench. The advantage shrinks at six calls, where both schedules draw nearly all of the seven pooled samples. Settled fractions in Table 8 count a pilot as unanimous only if all r calls return parsable responses inducing the same action, whereas the pilot buckets of Figure 2b are formed from the first three parsable responses; the two therefore differ most on ScienceWorld, where unparsable responses are most frequent.

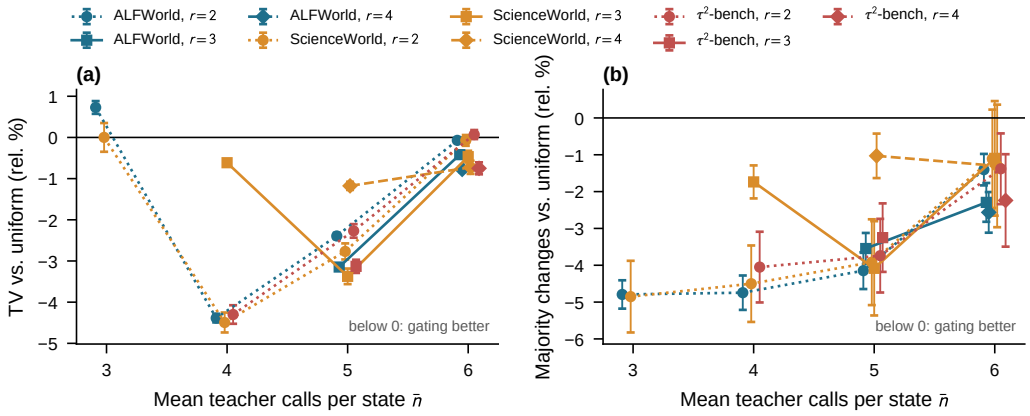


Figure 3: Pilot-based allocation against uniform sampling at a matched mean number of teacher calls, measured against a held-out reference posterior. (a) Relative change in total-variation distance and (b) relative change in the fraction of states whose majority action disagrees with the reference; negative values favour gating. Error bars are 95% intervals from the paired standard error across states.

E Experimental Configuration

White-box OPD. $\text{OPD}_{\text{logit}}$ uses the Qwen3-32B teacher's token log-probabilities on the student's own rollouts, with the same student initialization and training schedule as AP-OPD.

Differences in the OmniOPD comparison. The OmniOPD baseline uses the authors' released implementation with the same teacher, student initialization, and training schedule.

F Proofs and Additional Analysis

F.1 Proof of Proposition 1

Adding and subtracting $\log \pi_\theta(a \mid s)$ inside Eq. (3) gives

$$\mathcal{L}(\theta; s) = -\sum_a \widehat{q}_s(a) \log \pi_\theta(a \mid s) + \sum_a \widehat{q}_s(a) \log \frac{\pi_\theta(a \mid s)}{p_\theta(\bar{y}(a) \mid s)}. \tag{15}$$

The first sum is the cross-entropy $H(\widehat{q}_s) + \mathrm{KL}(\widehat{q}_s \| \pi_\theta(\cdot \mid s))$. Since $\bar{y}(a)$ is one of the responses that induce a and $Z_\theta(s) \le 1$, $p_\theta(\bar{y}(a) \mid s) \le \sum_{y: c_s(y)=a} p_\theta(y \mid s) = Z_\theta(s)\, \pi_\theta(a \mid s) \le \pi_\theta(a \mid s)$, so every term of the second sum is non-negative, and the sum vanishes if and only if $p_\theta(\bar{y}(a) \mid s) = \pi_\theta(a \mid s)$ for every a with $\widehat{q}_s(a) > 0$. □

Setting	ALFWorld	ScienceWorld	τ^2-bench
Environment and evaluation			
Environment version	ALFWorld 0.4.2	ScienceWorld 1.2.2	τ^2-bench 1.0.1 (telecom)
Evaluation split and episodes	`valid_unseen`, 134	test, all 1,308	held-out, 46 of 114 tasks
Maximum horizon (steps)	50	50	30
Evaluation action interface	`<action>` tag, admissible actions	`<action>` tag, action templates	tool calls, solo mode
Models and decoding			
Students	Qwen3-1.7B, Qwen3-4B		
Teachers	Qwen3-32B (open weights); DeepSeek-V4-Flash, Gemini-3.8-Flash, GPT-5.6-Luna (API)		
API access	via OpenRouter		
Teacher temperature and top-p	Qwen3-32B: 1.0, 1.0; API teachers: provider defaults		
Teacher maximum output tokens	64	64	128
Student extended reasoning	disabled		
Student maximum new tokens per turn	64		
Teacher sampling			
Pilot size r, cap $n_{\max}$	$r = 3$, $n_{\max} = 18$		
Allocation parameter ϵ	0.25		
Unparsable teacher responses	excluded from $\widehat{q}_s$ and re-queried; counted in the query cost		
Training			
Training rounds and rollouts per round	500×2	500×2	200×2
KL coefficient λ	0.01		
Optimizer, learning rate, batch size	AdamW, 10^{-6} (constant), 2		
Fine-tuning	LoRA adapters; base model frozen		
Context truncation (tokens)	6,144	6,144	20,480
Training seeds	3		

Table 9: Experimental configuration.

F.2 Proof of Theorem 2

The estimate $\widehat{q}_s$ is the empirical distribution of n independent draws from q_s over the K_s actions of $\mathcal{A}(s)$. For a categorical distribution over K_s outcomes, the empirical-distribution inequality of Weissman et al. (2003) states that

$$\Pr(\|\widehat{q}_s - q_s\|_1 \geq u) \leq (2^{K_s} - 2)\exp(-nu^2/2). \tag{16}$$

Since $\mathrm{TV}(p, q) = \frac{1}{2}\|p - q\|_1$, setting $u = 2\sqrt{(\log(2^{K_s} - 2) + \log(1/\delta))/(2n)}$ gives Eq. (7) with probability at least $1 - \delta$. Because $\widehat{q}_s$ is supported on the actions to which q_s assigns positive probability, K_s may be replaced by the size of that support whenever it is smaller. □

Imperfect parser. Let $c_s^\star$ be the ideal equivalence map and $q_s^\star$ its action posterior, assume that both maps accept the same responses, and let $\eta_s = \Pr[c_s(Y) \neq c_s^\star(Y) \mid c_s(Y) \in \mathcal{A}(s)]$ be the error of the implemented map among them. Coupling the ideal and implemented actions by applying both maps to the same teacher response Y, the coupling inequality gives

$$\mathrm{TV}(q_s, q_s^\star) \leq \Pr[c_s(Y) \neq c_s^\star(Y) \mid c_s(Y) \in \mathcal{A}(s)] = \eta_s, \tag{17}$$

so by the triangle inequality, with probability at least $1 - \delta$, $\mathrm{TV}(\widehat{q}_s, q_s^\star) \leq \eta_s + \sqrt{(\log(2^{K_s} - 2) + \log(1/\delta))/(2n)}$. Merging responses with different consequences increases η_s and biases the target, whereas splitting equivalent responses only refines the action space.

Minimax rate. Because total variation is one half of ℓ_1 distance, the categorical-distribution specialization of the nonasymptotic minimax ℓ_1 result of Han et al. (2015) is

$$\inf_{\widehat{q}} \sup_{q \in \Delta^{K-1}} \mathbb{E}_q[\mathrm{TV}(\widehat{q}, q)] = \Theta\left(\min\left\{1, \sqrt{\frac{K-1}{n}}\right\}\right), \tag{18}$$

so the minimax query complexity for expected TV error ϵ is $\Theta(K/\epsilon^2)$. A target defined over M surface categories replaces K by M.

F.3 Proof of Lemma 3 and the aggregation gap

Let $A_1, \ldots, A_n$ be the actions induced by the teacher samples, which are independent draws from q_s. Then $\widehat{g}_s = \sum_a \widehat{q}_s(a) g_a = n^{-1} \sum_{i=1}^n g_{A_i}$, so $\mathbb{E}[\widehat{g}_s] = \bar{g}_s$ and

$$\mathrm{Cov}(\widehat{g}_s) = \frac{1}{n}\mathrm{Cov}_{A \sim q_s}(g_A) = \frac{1}{n}\sum_a q_s(a)(g_a - \bar{g}_s)(g_a - \bar{g}_s)^\top. \tag{19}$$

Taking the trace and using $\sum_a q_s(a)\|g_a - \bar{g}_s\|^2 = \frac{1}{2}\sum_{a,b} q_s(a)q_s(b)\|g_a - g_b\|^2$,

$$\operatorname{tr}\operatorname{Cov}(\widehat{g}_s) = \frac{1}{2n}\sum_{a\neq b} q_s(a)q_s(b)\|g_a - g_b\|^2 \leq \frac{D_s^2}{2n}\sum_{a\neq b} q_s(a)q_s(b) = \frac{D_s^2\, G(s)}{2n}, \tag{20}$$

because $\sum_{a\neq b} q_s(a)q_s(b) = 1 - \sum_a q_s(a)^2 = G(s)$. For one-hot $g_a = \mathbf{e}_a$, $\bar{g}_s = q_s$ and $\sum_a q_s(a)\|\mathbf{e}_a - q_s\|^2 = 1 - \|q_s\|_2^2 = G(s)$, so the trace equals $G(s)/n = \mathbb{E}\|\widehat{q}_s - q_s\|_2^2$. □

Aggregation gap. A student distilled on surface strings is trained towards the empirical surface distribution $\widehat{p}_{s,n}$, an estimate of the surface distribution p_s, whereas AP-OPD is trained towards $\widehat{q}_s$, an estimate of its pushforward q_s. By the one-hot case above, $\mathbb{E}\|\widehat{p}_{s,n} - p_s\|_2^2 = (1 - \|p_s\|_2^2)/n$ and $\mathbb{E}\|\widehat{q}_s - q_s\|_2^2 = (1 - \|q_s\|_2^2)/n$. Expanding $q_s(a)^2 = (\sum_{y:c_s(y)=a} p_s(y))^2$ gives

$$\mathbb{E}\|\widehat{p}_{s,n} - p_s\|_2^2 - \mathbb{E}\|\widehat{q}_s - q_s\|_2^2 = \frac{2}{n}\sum_a \sum_{\substack{y<y' \\ c_s(y)=c_s(y')=a}} p_s(y)p_s(y'), \tag{21}$$

which is nonnegative and grows with the probability that two independent teacher samples express the same action in different words.

F.4 Proof of Proposition 4

Ignoring integer rounding, the budget-constrained problem has Lagrangian $\sum_i \omega_i G_i/n_i + \lambda(\sum_i n_i - N)$, and stationarity gives $n_i = \sqrt{\omega_i G_i/\lambda} \propto \sqrt{\omega_i G_i}$, the allocation of Neyman (1992). For the accuracy-constrained problem, minimise $\sum_i n_i$ subject to $m^{-1}\sum_i \omega_i G_i/n_i \leq \epsilon^2$. The constraint binds at the optimum, and the Lagrangian $\sum_i n_i + \mu(m^{-1}\sum_i \omega_i G_i/n_i - \epsilon^2)$ gives $n_i = \sqrt{\mu\omega_i G_i/m}$. Substituting into the binding constraint, $m^{-1}\sum_i \omega_i G_i/n_i = m^{-1}\sqrt{m/\mu}\sum_i \sqrt{\omega_i G_i} = \epsilon^2$, so $\sqrt{\mu/m} = \sum_j \sqrt{\omega_j G_j}/(m\epsilon^2)$ and $n_i^\star = \sqrt{\omega_i G_i}\sum_j \sqrt{\omega_j G_j}/(m\epsilon^2)$. Summing gives $N^\star(\epsilon) = (\sum_j \sqrt{\omega_j G_j})^2/(m\epsilon^2)$. A uniform depth n achieves accuracy ϵ^2 when $n = \sum_j \omega_j G_j/(m\epsilon^2)$, at total cost $mn = \sum_j \omega_j G_j/\epsilon^2$, and Cauchy–Schwarz gives $(\sum_j \sqrt{\omega_j G_j})^2 \leq m\sum_j \omega_j G_j$. Hence $N^\star(\epsilon) \leq \sum_j \omega_j G_j/\epsilon^2$, the cost of the uniform depth with the same accuracy, with a ratio $(\sum_j \sqrt{\omega_j G_j})^2/(m\sum_j \omega_j G_j)$ that is small when disagreement is concentrated on few states and can be estimated from the pilot. States with $G_i = 0$ contribute no noise and receive $n_i^\star = 0$. □

F.5 Proof of Lemma 5

The r pilot actions are independent draws from $q = q_s$, so they all equal a with probability $q(a)^r$, and the probability of a unanimous pilot is $\sum_a q(a)^r$. Since $\max_a q(a)^2 \leq \sum_a q(a)^2 = \|q\|_2^2$, we have $\max_a q(a) \leq \|q\|_2$ and, for $r \geq 2$,

$$\sum_a q(a)^r \leq \Big(\max_a q(a)\Big)^{r-2}\sum_a q(a)^2 \leq \|q\|_2^{r-2}\|q\|_2^2 = \|q\|_2^r = \big(1 - G(s)\big)^{r/2}. \tag{22}$$

The sum equals one if and only if q places all its mass on one action, that is, if and only if $G(s) = 0$. □

F.6 Stopping bias of the pooled estimator

The consensus-gated sampler chooses n_s from the pilot and then pools the pilot with the further samples, so $\widehat{q}_s$ is not the fixed-n estimator of Theorem 2. Write $\widehat{p}_s$ for the empirical distribution of the r pilot actions and $w_s = r/n_s$ for the weight of the pilot in the pooled estimate, so that $w_s = 1$ at a settled state and $w_s \in [r/n_{\max}, 1]$ in general.

Proposition 6. *The pooled estimator satisfies*

$$\big\|\mathbb{E}[\widehat{q}_s] - q_s\big\|_2 \leq \sqrt{\operatorname{Var}(w_s)}\,\sqrt{\frac{G(s)}{r}} \leq \frac{1 - r/n_{\max}}{2}\sqrt{\frac{G(s)}{r}}. \tag{23}$$

In particular, the estimator is unbiased at every state with $G(s) = 0$. Moreover, the error contributed by settled states satisfies

$$\mathbb{E}\big[\mathrm{TV}(\widehat{q}_s, q_s)\,\mathbf{1}\{s \text{ settled}\}\big] = \sum_a q_s(a)^r\big(1 - q_s(a)\big) \le G(s)\big(1 - G(s)\big)^{(r-1)/2}. \tag{24}$$

Proof. Let $\widehat{f}_s$ be the empirical distribution of the further samples when $n_s > r$. Then $\widehat{q}_s = w_s\widehat{p}_s + (1 - w_s)\widehat{f}_s$. The number of further samples depends only on the pilots, and the further samples are fresh, so $\mathbb{E}[\widehat{f}_s \mid \text{pilots}] = q_s$ and $\mathbb{E}[(1-w_s)\widehat{f}_s] = \mathbb{E}[1-w_s]\,q_s$. Hence $\mathbb{E}[\widehat{q}_s]-q_s = \mathbb{E}[w_s(\widehat{p}_s-q_s)] = \mathbb{E}[(w_s-\mathbb{E}w_s)(\widehat{p}_s-q_s)]$, using $\mathbb{E}[\widehat{p}_s] = q_s$. By Cauchy–Schwarz, $\|\mathbb{E}[\widehat{q}_s]-q_s\|_2 \le \mathbb{E}[|w_s-\mathbb{E}w_s|\,\|\widehat{p}_s-q_s\|_2] \le \sqrt{\mathrm{Var}(w_s)}\sqrt{\mathbb{E}\|\widehat{p}_s-q_s\|_2^2}$, and $\mathbb{E}\|\widehat{p}_s-q_s\|_2^2 = G(s)/r$ by the one-hot case of Lemma 3. Finally, a variable supported on $[r/n_{\max}, 1]$ has variance at most $(1 - r/n_{\max})^2/4$. If $G(s) = 0$, the pilot equals q_s almost surely and the bias vanishes.

For Eq. (24), a state is settled on action a with probability $q_s(a)^r$, in which case $\widehat{q}_s = \mathbf{e}_a$ and $\mathrm{TV}(\mathbf{e}_a, q_s) = 1-q_s(a)$. Summing over a gives the equality. Writing $q_s(a)^r(1-q_s(a)) = q_s(a)^{r-1}\cdot q_s(a)(1-q_s(a))$ and bounding $q_s(a)^{r-1} \le (\max_b q_s(b))^{r-1} \le \|q_s\|_2^{r-1} = (1-G(s))^{(r-1)/2}$, as in the proof of Lemma 5, gives the inequality because $\sum_a q_s(a)(1-q_s(a)) = G(s)$. □

Because a state stops exactly when its pilot is unanimous, the bias moves mass towards the pilot's modal action, which makes the target more confident than q_s. With $r = 3$ and the cap $n_{\max} = 18$ used in our experiments, the bound is $0.42\sqrt{G(s)/3}$, and exact computation for a two-action posterior with $q_s(a) = 0.7$, when every unresolved state is sampled up to the cap, gives a bias of 0.10 in ℓ_2 norm against a bound of 0.16. Since $\sum_a(\mathbb{E}[\widehat{q}_s(a)] - q_s(a)) = 0$, the induced bias of the stochastic gradient satisfies $\|\mathbb{E}[\widehat{g}_s] - \bar{g}_s\| \le \frac{1}{2}D_s\|\mathbb{E}[\widehat{q}_s] - q_s\|_1$. The KL penalty toward the initial student (Appendix B) tempers the overconfidence that this bias introduces. Together, the two statements control the two ways in which adaptive stopping distorts the target: pooling biases the estimate at states that continue, and settling replaces the posterior by a one-hot label at states that stop. The second contribution vanishes both when the teacher is deterministic and, by Lemma 5, when disagreement is so high that a unanimous pilot is unlikely; with $r = 3$ it is at most $G(s)(1 - G(s)) \le 1/4$.

F.7 Pilot disagreement estimator

For $i \ne j$, independence gives $\mathbb{E}[\langle Z_i, Z_j\rangle] = \langle \mathbb{E}Z_i, \mathbb{E}Z_j\rangle = \|q_s\|_2^2$. Averaging over the $r(r-1)$ ordered pairs proves $\mathbb{E}[\widehat{G}_r(s)] = G(s)$. Each Z_i is a one-hot vector, so $\widehat{G}_r(s) = 0$ exactly when all pilot samples induce the same action.

F.8 Action-level KL convergence

For a fixed-round student state distribution d, and with $q_s^\star$ and η_s as in the imperfect-parser paragraph above, define $\mathcal{K}_d(\pi) = \mathbb{E}_{s\sim d}[\mathrm{KL}(q_s^\star\|\pi(\cdot \mid s))]$ and $\bar{\eta} = \mathbb{E}_{s\sim d}[\eta_s]$. When the student emits actions in canonical form, Eq. (3) coincides with the empirical action log loss by Proposition 1, so the following result applies to the objective that AP-OPD optimises.

Corollary 7. *Let Π be finite with $\pi(a \mid s) \ge \beta > 0$ and $B = \log(1/\beta)$. Draw m independent states from d and n_i fresh teacher actions from q_{s_i} at state s_i. If $\widehat{\pi}$ minimizes empirical action log loss up to ξ_{opt}, then with probability at least $1 - \delta$,*

$$\begin{aligned}\mathcal{K}_d(\widehat{\pi}) \le\ & \inf_{\pi\in\Pi}\mathcal{K}_d(\pi) + \xi_{\mathrm{opt}} + 2B\sqrt{\frac{\log(4|\Pi|/\delta)}{2m}} \\ & + \frac{2B}{m}\sqrt{\frac{\log(4|\Pi|/\delta)}{2}\sum_{i=1}^{m}\frac{1}{n_i}} + 2B\bar{\eta}.\end{aligned} \tag{25}$$

Proof. Let $\ell_\pi(s,a) = -\log\pi(a \mid s) \in [0, B]$ and define the implemented population and empirical risks $\widetilde{R}(\pi) = \mathbb{E}_{S\sim d, A\sim q_S}[\ell_\pi(S, A)]$ and $\widehat{R}(\pi) = m^{-1}\sum_{i=1}^m n_i^{-1}\sum_{j=1}^{n_i}\ell_\pi(s_i, A_{ij})$. For fixed π,

Hoeffding's inequality over states and a union bound over Π imply, simultaneously for every $\pi \in \Pi$,

$$\left| \widetilde{R}(\pi) - \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{A \sim q_{s_i}} [\ell_\pi(s_i, A)] \right| \leq B \sqrt{\frac{\log(4|\Pi|/\delta)}{2m}}. \tag{26}$$

Conditional on the sampled states, teacher samples are independent, and each sample at s_i has weight $1/(mn_i)$. A weighted Hoeffding inequality and the same union bound give

$$\left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{A \sim q_{s_i}} [\ell_\pi(s_i, A)] - \widehat{R}(\pi) \right| \leq \frac{B}{m} \sqrt{\frac{\log(4|\Pi|/\delta)}{2} \sum_{i=1}^{m} \frac{1}{n_i}}. \tag{27}$$

Let D be the sum of the two right-hand sides. Approximate empirical risk minimization gives $\widetilde{R}(\widehat{\pi}) \leq \inf_{\pi \in \Pi} \widetilde{R}(\pi) + \xi_{\mathrm{opt}} + 2D$. The coupling in Eq. (17) and boundedness of the loss imply $|R^\star(\pi) - \widetilde{R}(\pi)| \leq B\bar{\eta}$ for every π, where $R^\star(\pi) = \mathbb{E}_{S \sim d, A \sim q^\star_S}[\ell_\pi(S, A)]$, so transferring the bound to $R^\star$ adds at most $2B\bar{\eta}$. Finally, $R^\star(\pi) = \mathbb{E}_{S \sim d}[H(q^\star_S)] + \mathcal{K}_d(\pi)$, and the entropy term cancels from the excess risk. We state the result for bounded student log loss because unsmoothed empirical posteriors can assign zero mass to unobserved teacher actions (Paninski, 2004). The bound treats supervised states as independent; consecutive states within a trajectory are correlated, so additional states from the same rollouts contribute less than the $m^{-1/2}$ term suggests. □