DECISIONOPD: ALLOCATING TEACHER VIEWS AND STATE COVERAGE FOR DECISION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training compact language agents requires effective use of costly teacher supervision. Additional queries can refine targets for familiar decisions or extend supervision to new states, and the value of either choice depends on the training available to the student. We introduce DecisionOPD, a framework for comparing these allocations through executable action distributions while accounting for teacher cost and student exposure. The resulting comparisons connect supervision quality and breadth to downstream policy performance. Refined targets retain an empirical advantage over broader coverage as training increases, although the size of that advantage remains uncertain. On ALFWorld, aligned mean targets improve task success from 38.9% to 50.0% under matched state coverage and student updates.

1 INTRODUCTION

Distillation turns capable language teachers into compact agents that are cheaper to deploy. Its efficiency also depends on how supervision is acquired. A teacher query can improve the target for a familiar decision or provide supervision for a new state. Choosing between these uses determines both what the student learns from and how far the teacher budget reaches.

This choice is especially relevant when agents act through menus of executable actions. A teacher's preferences can depend on how the same actions are ordered and labeled. Aligning predictions across presentations and averaging them produces a richer target for the same decision (Liusie et al., 2024). Such refinement competes for resources with labeling additional decisions. Prior studies of reasoning and distillation report benefits from both broader coverage and repeated teacher sampling under different conditions (Tong et al., 2026, Bansal et al., 2025, Guha et al., 2025). For agent training, the practical question is how to allocate that supervision given the optimization budget available to the student.

Acquiring supervision and learning from it are distinct stages. At a fixed update budget, a larger labeled pool gives the student fewer opportunities to revisit each record. Refining targets concentrates teacher effort on records that receive more repeated exposure. The value of a supervision allocation may therefore change as training proceeds. Teacher cost and student exposure together define the comparison relevant to efficient distillation.

We introduce **DecisionOPD** as a framework for studying this joint allocation. It aligns teacher judgments over executable actions and compares target refinement, recorded state coverage, and optimization exposure through a shared student policy interface. The central comparisons hold the behavior stream fixed, separating changes to supervision from changes to data collection.

Controlled ALFWorld experiments show that target refinement improves policy success on a fixed pool. At matched teacher cost, it also retains an empirical advantage over broader supervision after further training. The modest narrowing of this advantage remains uncertain. These findings motivate evaluating supervision choices through the policy performance attainable at the intended training budget.

Our contributions are:

1. **Target refinement improves downstream decisions.** Aligned teacher targets improve policy success on the same recorded states. The direction persists across repeated executions, connecting the supervision change to deployed behavior.

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.

2. **Longer training preserves the observed allocation ranking.** Broader coverage benefits from additional updates, while refined targets remain ahead at matched teacher cost. The narrowing of the difference is modest and uncertain.
3. **Supervision allocation depends on the learning protocol.** Teacher views, labeled coverage, and student exposure jointly define the comparison. Frozen-checkpoint evaluations further show that the observed advantage depends on which tasks measure policy performance.

2 Related work

From debiasing judgments to allocating supervision. Knowledge distillation exploits information beyond the winning label (Hinton et al., 2015). PriDe addresses option-ID bias through a reusable prior (Zheng et al., 2024). Liusie et al. align and average permutation-dependent teacher predictions to train a compact student (Liusie et al., 2024). Their method is our closest target-construction precedent. Repeated-label acquisition also compares improving existing labels with labeling new examples through downstream classifier quality (Sheng et al., 2008). These approaches motivate separating target refinement on a fixed pool from its value relative to broader supervision at equal teacher cost.

Supervision quality, breadth, and learning compute. Bansal et al. (2025) show that more samples from a cheaper teacher can improve reasoning at a specified sampling budget and fixed student steps. Tong et al. (2026) find gains from broader question coverage at fixed record budgets. OpenThoughts (Guha et al., 2025) finds that multiple answers per question can preserve or improve performance at a fixed dataset size. Together, these findings make the allocation conditions central to interpretation. Tiapkin et al. (2025) connect prompt diversity and completion budgets to distillation dynamics. Distillation scaling laws separate teacher and student compute (Busbridge et al., 2025), while budget-aware selection conditions data quantity and distribution on training compute (Wan et al., 2025). We compare an additional presentation of the same executable decision with supervision for another visit, matching logical teacher-input cost and specifying student exposure.

From learner-state coverage to usable coverage. DAgger queries learner states (Ross et al., 2011), Actor-Mimic transfers soft action preferences on student trajectories (Parisotto et al., 2016), and GKD supervises self-generated sequences (Agarwal et al., 2024). Policy-distillation analysis and agentic extensions link supervision to the data distribution (Czarnecki et al., 2019, Li et al., 2026, Wang et al., 2026). A September 2026 preprint by Fu et al. (2026) separates input-query count from rollout coverage measured by teacher-representation clusters, and observes slow alignment even with fixed token-prefix states. Our frozen behavior stream provides a complementary action-level comparison, measuring recorded-visit exposure and policy return. Earlier learner-refresh controls remain separate. SmartAD selects correct teacher trajectories by student likelihood (Tang & Zhao, 2026). AggreVaTe uses continuation cost-to-go for training (Ross & Bagnell, 2014), whereas our frozen branches diagnose local action replacements alongside the distributional intervention.

From language generation to executable decisions. DRRN scores language-described actions (He et al., 2016). DecisionOPD uses categorical executable-choice targets and matched student interfaces for the allocation comparison. Independent scorers and constrained action-text generation provide supporting evidence. Paired task outcomes and execution sensitivity distinguish task aggregation from finite-run uncertainty (Agarwal et al., 2021).

3 Problem and method

Executable action sets provide a common supervision space for comparing teacher allocations. We represent each recorded visit x_i by its visible history and finite menu $C(x_i)$ of executable choices, retaining distinct visit IDs for repeated inputs. To combine judgments across presentations, let g specify the menu order and labels, and let M_g map the teacher's label probabilities $\widetilde{q}_T^g$ back to action identities. Averaging over the queried presentations G_i gives the training target

$$\bar{q}_i = \frac{1}{|G_i|} \sum_{g \in G_i} M_g \widetilde{q}_T^g(\cdot \mid x_i).$$

The core comparison uses the original and reversed menus. Alignment places their predictions over the same action identities, while target correctness and student invariance remain empirical questions.

The alignment step is essential because a menu label has meaning only within its presentation. Reversing the menu changes which label denotes an executable command, while preserving the commands available to the agent. We therefore aggregate probability mass after recovering those command identities. This construction lets presentation variation affect the supervision distribution without changing the underlying decision being supervised.

3.1 One teacher allocation, two resource budgets

Efficient allocation depends on both the supervision acquired and the training used to learn from it. We describe the supervision purchase by $\mathcal{A} = (\mathcal{D}, \{G_i\}_{i\in\mathcal{D}})$, which selects visit IDs and teacher views. Its logical teacher-input cost is

$$B_T(\mathcal{A}) = \sum_{i\in\mathcal{D}} \sum_{g\in G_i} \ell_T(P_g(x_i, C(x_i))),$$

where ℓ_T is teacher-tokenizer length. Costs are incremental after the shared warm stage. Cache reuse saves physical calls during training while retaining the logical supervision cost.

Optimization exposure is represented by $\mathcal{T} = (\mathcal{B}_1, \ldots, \mathcal{B}_U)$, a sequence of U actual batches of size b, with record multiplicities m_i satisfying $\sum_i m_i = Ub$. The trained checkpoint is $\theta_U(\mathcal{A}, \mathcal{T}, \xi)$, where ξ includes initialization and execution conditions. We compare its mean task success $\widehat{J}_{\mathcal{E}}$ on fixed tasks $\mathcal{E}$ across allocations at specified $(B_T, U, \mathcal{E})$.

At fixed B_T, additional views can displace labeled coverage. At fixed Ub, a larger pool reduces repeated exposure. Following the distinction between teacher and student resources in distillation scaling (Busbridge et al., 2025), we report labeled visits, distinct inputs, and record multiplicities separately. Teacher tokens and student updates define the matched resources. A full compute comparison would also account for student tokens, physical calls, collection cost, and wall time.

These budgets describe different opportunities for learning. A teacher query can refine a target already in the pool or label another visit. A student update determines how often acquired supervision influences the parameters. Reusing a cached target increases optimization exposure without purchasing additional information from the teacher. Consequently, an allocation comparison has to specify both the supervision pool and the training endpoint to make its resource tradeoff interpretable.

3.2 Distribution supervision and single-readout deployment

To connect supervision allocation to deployed behavior, we train the student directly on distributions over executable actions. Jointly reading the history and full menu produces label-token logits, which are normalized into $p_\theta(c \mid x_i, C(x_i))$. Reverse KL is an established language-model distillation objective (Gu et al., 2024). Here update u minimizes it over executable choices,

$$\mathcal{L}_u(\theta) = \frac{1}{b} \sum_{i\in\mathcal{B}_u} D_{\mathrm{KL}}(p_\theta(\cdot \mid x_i, C(x_i)) \,\|\, \bar{q}_i)\,.$$

Batch records specify the actual exposure used by this objective. Appendix F gives target flooring and normalization. At deployment the student selects an executable choice using a single readout of its own model. For fixed student probabilities and the strictly positive implemented targets, replacing identity target q_{id} with mean target q_{mean} changes the loss by

$$D_{\mathrm{KL}}(p_\theta \| q_{\mathrm{mean}}) - D_{\mathrm{KL}}(p_\theta \| q_{\mathrm{id}}) = \sum_a p_\theta(a) \log \frac{q_{\mathrm{id}}(a)}{q_{\mathrm{mean}}(a)}.$$

Even when the preferred action is unchanged, the rest of the target distribution can alter the learning objective. Whether that change improves the trained policy is an empirical question.

The loss difference depends on probability ratios across the full menu, weighted by the student's current distribution. Changes among alternatives can therefore matter when the student assigns them appreciable mass. An unchanged teacher argmax alone does not imply an unchanged training signal, and agreement of greedy choices cannot characterize the entire intervention.

3.3 Fixed-pool, allocation, and target interventions

Separating target refinement from its coverage cost requires comparisons that hold the surrounding training process fixed. We construct allocations from a shared warm checkpoint and frozen collection stream. **Single-view, 800 visits** uses identity-presentation targets on the first 800 visits. **Two-view mean, 800 visits** averages aligned original and reversed predictions on the same visits. **Single-view, expanded pool** uses identity targets on the longest stream prefix within the two-view teacher-token budget: 1,544 visits. The first comparison isolates target construction. The second includes the different repetition rates induced by pool size at a specified update count.

To examine which target changes contribute to learning, we additionally separate visits according to whether averaging changes or preserves the identity-target argmax:

$$\mathcal{F} = \{i \in \mathcal{D}_{800} : \arg\max q_{\mathrm{mean},i} \neq \arg\max q_{\mathrm{id},i}\}, \qquad \mathcal{N} = \mathcal{D}_{800} \setminus \mathcal{F}.$$

This partition yields 200 changed-argmax and 600 unchanged-argmax visits. **Mean on argmax-changed visits** uses mean targets only on $\mathcal{F}$. **Mean on argmax-unchanged visits** uses them only on $\mathcal{N}$, retaining identity targets elsewhere. Together with the all-identity and all-mean targets, they form a 2×2 matrix. Comparing the unchanged-only intervention with single-view training, and all-mean training with the changed-only intervention, evaluates the unchanged-argmax target change in two contexts. Both interventions require both teacher views to determine membership and incur the full two-view cost. Changed argmax denotes a different preferred action. Its correctness is evaluated separately.

Freezing the collection stream makes the allocations comparable at a shared source of experience. It also fixes the scope of the comparison: downstream policy learning is measured on these acquired visits. The local probe below addresses a separate question by holding the continuation policy fixed and changing an action at a saved decision point. Its outcome does not by itself predict the result of training a new policy.

3.4 A separate local execution diagnostic

A local execution probe complements the training intervention by measuring the consequence of changing a chosen action. For history h, first action a, and frozen continuation μ, define $R^{\mu}(h, a)$ as the terminal return obtained by replaying the saved prefix, taking a, and following μ within the remaining original horizon. The contrast $\Delta^{\mu}(h) = R^{\mu}(h, a_{\mathrm{mean}}) - R^{\mu}(h, a_{\mathrm{id}})$ measures the consequence of a local greedy-action replacement. Distribution supervision instead changes a student's training objective, including when teacher argmax is unchanged. Appendix C reports the complete branch diagnostic. Appendix A.2 retains prospective query-allocation rules based on first-view features.

4 Experimental protocol

ALFWorld (Shridhar et al., 2021) uses internal partitions of the official training split, leaving the official test tasks unused. A frozen Qwen3.5-27B teacher supervises a Qwen3.5-4B student with LoRA (Hu et al., 2022). Inputs contain only visible history and the full executable menu. The original methods share a warm checkpoint, a frozen collection stream, and optimizer settings. The allocation follow-up uses two independent data-order seeds and saves checkpoints at 200 and 400 updates, with initialization, runtime, and evaluation settings held fixed. We label these executions Run I and Run II. Appendix F gives the training configuration.

The core evaluation contains 72 tasks balanced across task families and separate from the earlier development set. Selection excludes recorded historical exposure and was frozen after the original

training but before evaluation and score inspection. Failures and truncations remain in every denominator. The follow-up reuses these already analyzed tasks. Appendix A.3 documents selection and the scope of the history audit.

The primary contrasts measure the same-pool target effect, the equal-token allocation difference, and mean on argmax-unchanged visits versus single-view targets. A statistical amendment recorded after evaluation termination and before score inspection specified paired bootstrap resampling within task families and exact McNemar tests with Holm adjustment across the primary contrasts. Other target comparisons support interpretation. Task intervals are conditional on the trained checkpoints, and repeated executions assess sensitivity under shared initialization. The budget follow-up reports changes in each arm and in their contrast without altering the original test family.

Supporting studies connect target changes to local action consequences, examine static-domain adaptation on PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), and ARC-Challenge (Clark et al., 2018), and compare teachers on ALFWorld. Their protocols and sample counts appear in the appendices alongside the earlier learner-refresh and query-allocation studies.

5 RESULTS

5.1 TARGET REFINEMENT IMPROVES THE FIXED-POOL ENDPOINT

Aligned mean targets improve policy success over single-view targets on the same recorded visits. Table 1 places this comparison alongside expanded coverage and the auxiliary target interventions. All evaluations complete without policy failures, and teacher costs are measured after the common warm stage.

Table 1: Original configurations and endpoints. "Changed-only" and "unchanged-only" apply the mean target on the corresponding argmax group and identity targets elsewhere. All five methods use 200 updates. Success is out of 72 tasks.

Method	**Visits**	**Target**	**Teacher tokens**	**Success**
Single-view, 800 visits	800	Identity	755,863	28/72
Two-view mean, 800 visits	800	Mean throughout	1,511,726	36/72
Single-view, expanded pool	1,544	Identity	1,510,554	26/72
Mean on argmax-changed visits	800	Changed-only	1,511,726	32/72
Mean on argmax-unchanged visits	800	Unchanged-only	1,511,726	33/72

Table 2: Original prespecified effects. Gains compare successes on the same 72 tasks. The first three rows are the primary family. The fourth compares two-view mean targets with mean on argmax-changed visits and is supporting. Single-view refers to the 800-visit pool unless the expanded pool is named. These are conditional point estimates.

Contrast	**Gain: tasks / total**
Two-view mean versus single-view, 800 visits	+8/72
Two-view mean versus single-view, expanded pool	+10/72
Mean on argmax-unchanged visits versus single-view	+5/72
Two-view mean versus mean on argmax-changed visits	+4/72

The same-pool advantage persists across repeated executions, even as individual task outcomes change (Table 3). This agreement strengthens the directional observation under shared initialization. Appendix A.3 reports device assignments and task-level changes.

Table 3: Same-pool execution checks. The methods share recorded visits, initialization, update budget, and evaluation tasks. Repeated executions measure sensitivity under this common protocol.

Execution	Single-view / 72	Two-view mean / 72	Gain (tasks)
Original	28	36	+8
Repeated assignments	31	37	+6
Swapped assignments	31	38	+7

5.2 The allocation advantage remains positive with more updates

At approximately matched teacher-input cost, refined targets outperform single-view supervision over the expanded pool in the original comparison (Table 2). Both pools are fully used, but the smaller pool receives more repeated exposure. To examine how this contrast evolves with training, we follow both allocations from 200 to 400 updates.

Longer training improves expanded-pool success in both runs. Two-view success improves in one run and remains stable in the other, retaining the observed advantage at each endpoint (Figure 1). The comparison uses the same evaluation tasks throughout, with no policy failures.

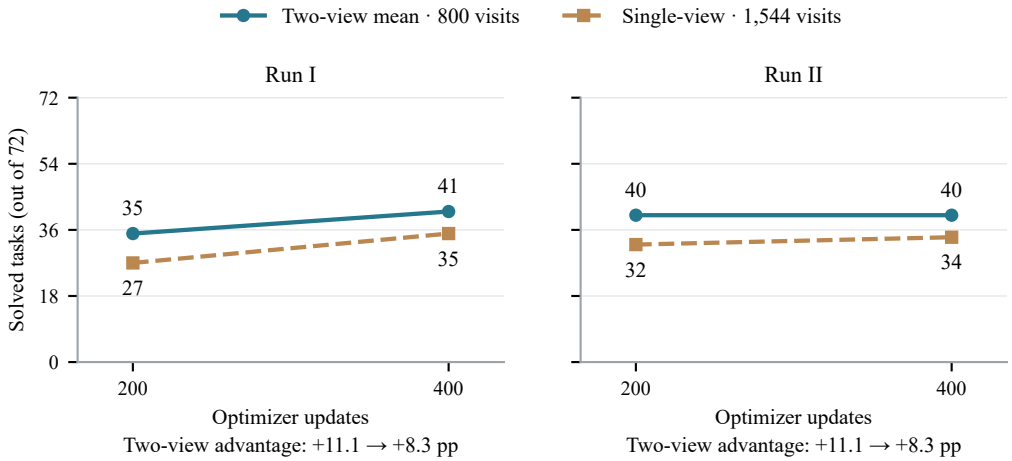

Figure 1: Longer training preserves the observed allocation ranking. Two-view mean targets and expanded-pool single-view targets share approximately the same teacher-input budget. Lines connect prespecified training endpoints, with success measured on the same tasks under shared initialization. Appendix A.5 reports paired outcomes and exposure.

The gap narrows modestly in both runs, but the paired task intervals for its change include zero. The evidence therefore supports a persistent empirical advantage at the measured budgets, with uncertainty about how that advantage evolves.

Additional updates increase exposure to both pools while preserving their different repetition rates. Appendix A.5 reports the complete accounting, using batch fingerprints for the original comparison and saved schedules for the follow-up. These records connect the policy outcomes to the training each allocation actually receives.

The distinction between absolute improvement and a changing allocation contrast is central to interpretation. Both methods can benefit from further training even when their gap narrows. The endpoint comparison therefore asks which supervision purchase yields the better policy at a specified training budget. It does not identify a general crossover point or establish that one allocation

eventually overtakes the other. Reporting the trajectories of both arms makes this distinction visible and avoids attributing a relative change to deterioration in either policy.

The fixed-pool and equal-cost comparisons answer complementary questions. Holding visits fixed measures the value of replacing the supervision target. Holding teacher cost fixed additionally introduces a coverage difference and the repetition pattern that follows from it. The longer-training follow-up traces this latter comparison. Its interpretation remains tied to those allocations, while the original fixed-pool intervention establishes a separate reference for understanding the value of target refinement.

5.3 Auxiliary target and action diagnostics

Applying mean targets where the teacher's preferred action is unchanged gives a positive policy-success estimate in both target contexts (Table 2). This pattern motivates studying the full supervision distribution alongside its argmax. The subgroup evidence remains exploratory, and unequal group sizes and nonlinear training preclude an additive interpretation of per-visit contributions.

The frozen execution probe asks what happens when averaging changes the chosen action. Such changes occur on a minority of selected decisions, and most changed-action heldout branches reach equal terminal returns. Joint failures account for much of that equality. These local outcomes are conditional on the continuation policy, while the training comparisons measure the downstream value of distribution supervision. Appendix C reports branch outcomes and the original-horizon accounting.

These diagnostics distinguish two routes through which a teacher view can matter. A different greedy action can change the immediate trajectory. A different distribution can also change the gradients received by the student, even when the teacher would execute the same action. The target interventions address this second route by changing supervision on selected groups of visits. Their positive estimates motivate studying information beyond the winning label, without assigning an independent contribution to every visit or treating the groups as interchangeable.

Equal terminal returns in the execution probe also require care. They can arise from different trajectories that ultimately fail, so they do not establish behavioral equivalence. Conversely, a favorable local replacement under a frozen continuation does not establish that training on the corresponding target improves the whole policy. Keeping the branch diagnostic separate from the learning comparison allows each to answer its own question.

5.4 Protocol sensitivity and external boundaries

Changing the evaluation tasks while holding checkpoints fixed alters the magnitude of the allocation contrast and reverses the ranking in an earlier low-budget comparison. The historical grid and core study share a warm checkpoint and frozen collection stream, while their allocations and evaluation tasks differ. The cross-evaluation in Appendix A.4 isolates the contribution of task-set composition.

Earlier matched controls place aligned mean targets above entropy matching and student presentation augmentation. This comparison supports examining the content of the teacher distribution beyond its overall concentration or the student's exposure to alternative presentations. Prospective query-allocation rules give mixed results across budgets (Appendix A.2).

The external studies provide qualified support. Mistral preserves the direction of the ALFWorld single-to-mean improvement. On PIQA, HellaSwag, and ARC-Challenge, averaging benefits teacher predictions more than student accuracy. Appendices D and E report the complete static-domain and cross-teacher comparisons.

6 Discussion

Refined targets can improve the policy learned from a fixed pool. The same-pool comparisons connect aligned teacher distributions to downstream success while holding recorded visits and student updates fixed. The auxiliary target interventions suggest that the relevant information may extend beyond the teacher's preferred action. The reverse-KL identity explains how such infor-

mation can affect the objective, while its influence on the learning dynamics remains an empirical question.

Supervision choices should be evaluated at the intended training budget. Broader coverage benefits from further optimization, yet refined targets retain their observed advantage. The modest narrowing of the gap offers limited evidence about its longer-term trajectory. This perspective complements budget-aware data selection (Wan et al., 2025) and repeated exposure to filtered pretraining data (Fang et al., 2025): acquiring useful supervision and allocating enough training to learn from it are coupled decisions. Record exposure makes this relationship explicit, with the resource match defined by teacher inputs and student updates.

The value of an allocation depends on where it is learned and evaluated. Frozen-checkpoint cross-evaluation shows that task composition can materially change the comparison. The additional teachers and domains reveal further variation in how improvements in supervision translate into policy performance. Together, these studies motivate reporting the collection stream, optimization exposure, and evaluation tasks as part of the allocation result.

For practical allocation, the relevant outcome is the policy obtained after the available training has been spent. A supervision target may be more informative while requiring repeated exposure to influence behavior. A broader pool may offer more distinct decisions while giving each one less attention. Evaluating these choices jointly provides a way to compare their realized value without assuming that either target quality or coverage alone determines the outcome. The same accounting also makes comparisons easier to interpret when teacher calls are cached, since logical supervision cost and physical execution cost remain distinct.

7 Limitations

Future research could explore supervision allocation in continual learning, where goals, tools, and environments change over time. Another direction is to combine teachers with complementary expertise and learn when an agent should request additional guidance. These extensions could support agents that adapt to unfamiliar situations while balancing learning cost and long-term utility.

8 Conclusion

DecisionOPD connects teacher-query allocation to the policy learned from the resulting supervision. Aligned mean targets improve success on a fixed pool and retain an empirical advantage over broader single-view coverage at the training budgets studied. Further optimization benefits the expanded pool without reversing that ranking, while the change in the gap remains uncertain. These findings support evaluating supervision quality and breadth together with student exposure, using downstream policy performance to guide the allocation of teacher effort.

AI Assistance

Generative AI assisted manuscript organization, language and figure preparation. Reported numerical results come from recorded experimental outputs.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 29304 to 29320, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

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

ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q. Tran, and Mehran Kazemi. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=3OyaXFQuDl.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432 to 7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL https://doi.org/10.1609/aaai.v34i05.6239.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/busbridge25a.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Wojciech M. Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant M. Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1331 to 1340. PMLR, 2019. URL http://proceedings.mlr.press/v89/czarnecki19a.html.

Alex Fang, Hadi Pouransari, Matt Jordan, Alexander Toshev, Vaishaal Shankar, Ludwig Schmidt, and Tom Gunter. Datasets, documents, and repetitions: The practicalities of unequal data quality. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/2aba6ec20299931d46ddeefd5ddcb442-Abstract-Conference.html.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models ii: One training example, 2026. URL https://arxiv.org/abs/2609.04172.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M. Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran

Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. *CoRR*, abs/2506.04178, 2025. doi: 10.48550/ARXIV.2506.04178. URL https://doi.org/10.48550/arXiv.2506.04178.

Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari Ostendorf. Deep reinforcement learning with a natural language action space. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1153. URL https://doi.org/10.18653/v1/p16-1153.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-policy distillation with curriculum turn-level guidance for multi-turn agents. *CoRR*, abs/2606.15912, 2026. doi: 10.48550/ARXIV.2606.15912. URL https://doi.org/10.48550/arXiv.2606.15912.

Adian Liusie, Yassir Fathullah, and Mark J. F. Gales. Teacher-student training for debiasing: General permutation debiasing for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 1376 to 1387. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.81. URL https://doi.org/10.18653/v1/2024.findings-acl.81.

Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06342.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *CoRR*, abs/1406.5979, 2014. URL http://arxiv.org/abs/1406.5979.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 627 to 635. JMLR.org, 2011. URL http://proceedings.mlr.press/v15/ross11a/ross11a.pdf.

Victor S. Sheng, Foster J. Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In Ying Li, Bing Liu, and Sunita Sarawagi (eds.), *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pp. 614 to 622. ACM, 2008. doi: 10.1145/1401890.1401965. URL https://doi.org/10.1145/1401890.1401965.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Guokai Tang and Feng Zhao. Smartad: Capacity-aligned agent distillation for small language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 27045 to 27057. Association for Computational Linguistics, 2026.

doi: 10.18653/V1/2026.FINDINGS-ACL.1349. URL https://doi.org/10.18653/v1/2026.findings-acl.1349.

Daniil Tiapkin, Daniele Calandriello, Johan Ferret, Sarah Perrin, Nino Vieillard, Alexandre Ramé, and Mathieu Blondel. On teacher hacking in language model distillation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/tiapkin25a.html.

Yao Tong, Jiayuan Ye, Anastasia Borovykh, and Reza Shokri. Generalization in LLM problem solving: The case of the shortest path. *CoRR*, abs/2604.15306, 2026. doi: 10.48550/ARXIV.2604.15306. URL https://doi.org/10.48550/arXiv.2604.15306.

Weilin Wan, Weizhong Zhang, and Cheng Jin. Computational budget should be considered in data selection. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/22c799f287fd05e7174fd65a3ce134af-Abstract-Conference.html.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. MAD-OPD: breaking the ceiling in on-policy distillation via multi-agent debate. *CoRR*, abs/2605.01347, 2026. doi: 10.48550/ARXIV.2605.01347. URL https://doi.org/10.48550/arXiv.2605.01347.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4791 to 4800. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1472. URL https://doi.org/10.18653/v1/p19-1472.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

A Complete fixed-budget and allocation results

A.1 View-depth and state-coverage endpoints

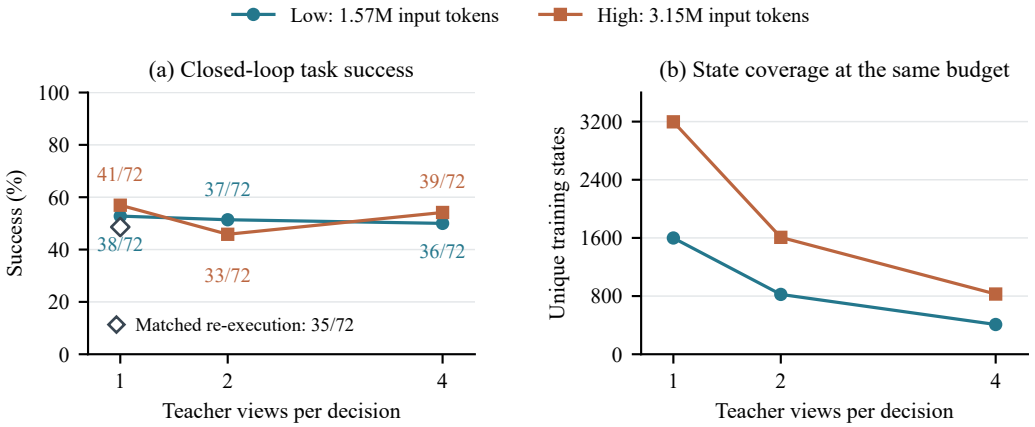

Figure 2: Historical fixed-budget view-depth and coverage grid. All students use 200 updates and 3,200 training presentations. Lines connect the measured endpoints at prespecified view counts. The hollow diamond represents a matched re-execution under the same initialization.

Table 4: Historical teacher-budget allocations and policy endpoints. Unique inputs use exact history and ordered-menu equality. All successes are out of 72.

Family	Condition	Visits	Unique inputs	Queries	Input tokens	Success
Depth, low	1 view	1,600	1,599	1,600	1,566,874	38/72
Depth, low	2 views	824	823	1,648	1,565,128	37/72
Depth, low	4 views	409	409	1,636	1,566,132	36/72
Depth, high	1 view	3,200	3,197	3,200	3,149,070	41/72
Depth, high	2 views	1,609	1,608	3,218	3,148,040	33/72
Depth, high	4 views	829	828	3,316	3,146,684	39/72
Coverage	Single, first 800	800	799	800	755,863	37/72
Coverage	Single, full 3,200	3,200	3,197	3,200	3,149,070	35/72
Coverage	Mean, first 800	800	799	1,600	1,511,726	42/72
Coverage	Mean, full 3,200	3,200	3,197	6,400	6,298,140	42/72

The paired comparisons show mixed returns to view depth and coverage. At the low budget, two-view and four-view targets have 4/5 and 6/8 exclusive successes against one-view targets. At the high budget, these counts are 2/10 and 4/6. Expanding the pool gives 5/7 exclusive successes for single-view targets and 4/4 for mean targets.

A.2 Prospective and post-hoc allocation endpoints

The random, first-view uncertainty, and calibrated return-score rules solve 41/39, 38/42, and 40/34 tasks at the low/high budgets. Post-hoc rate-matched controls differ by at most two tasks. Their

realized coverage and token costs also differ. The prospective rules use only first-view entropy, margin, candidate count, step, and history length before purchasing another view.

The original prospective package uses logical input-token budgets of approximately 1.57M and 3.15M. Its complete shared cache contains 12,800 teacher forwards and 12,596,280 input tokens. Reuse adds zero physical teacher forwards during student training. Each policy retains its logical supervision cost.

At the higher budget, random buys 1,053 second views and covers 2,181 instances. The return-score rule buys 1,318 and covers 1,837. Paired task switches for uncertainty versus random are 4 wins/7 losses at low budget and 5/2 at high. Return-score switches are 4/5 and 0/5. The zero-versus-five contrast has a small discordant count. We report its paired counts descriptively.

A.3 Common-protocol setup and execution checks

Experiment identifiers. The appendix retains the artifact identifiers for compact tables and traceability. A denotes Single-view, 800 visits. B denotes Two-view mean, 800 visits. C denotes Single-view, expanded pool. F and N apply mean targets only on argmax-changed and argmax-unchanged visits, respectively.

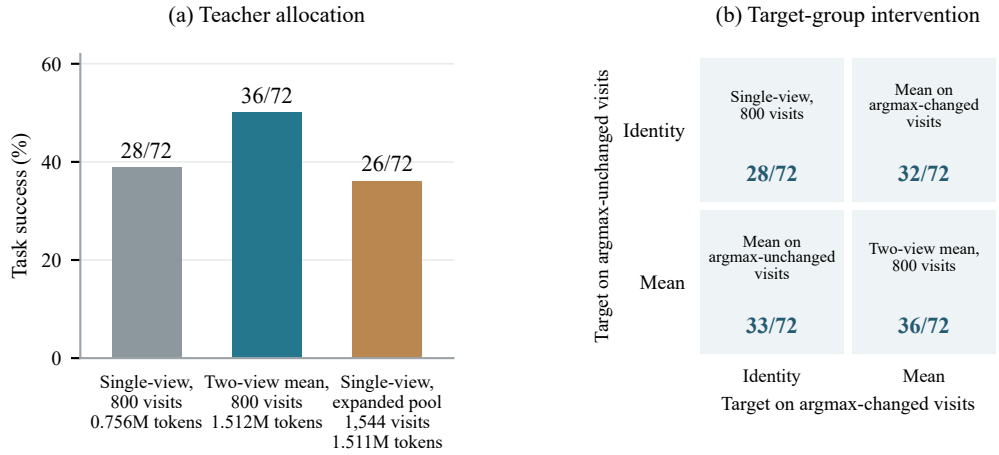

Figure 3: Original common-protocol allocation and target interventions. The left panel compares single-view and two-view mean targets on 800 visits with single-view targets on an expanded 1,544-visit pool under the two-view teacher-token cap. Costs are logical incremental supervision after the common warm checkpoint. The right panel separates mean targets on argmax-changed and argmax-unchanged visits. Repeat executions are reported separately. All failures remain in the 72-task denominators.

The frozen evaluation selection contains 72 official-train tasks, balanced across the six families. Selection excluded 1,016 task identities in recoverable structured history, comprising 844 internal-train and 172 internal-development identities. The remaining family pools contain 145, 619, 479, 366, 288, and 640 tasks. The first 12 tasks per family were selected in partition source order before any method's score was read. No official test tasks were used. All methods completed evaluation without policy failures.

The post hoc alias audit maps `task_id`, `requested_task_id`, and game-file paths to official task/trial identities, including repeated trailing trial aliases. It covers the source files listed in the frozen receipt and its imported receipts, excluding the new confirmation, repeat, and counter-

balanced outputs. Combining local and remote scans makes all 431 sources readable. The audit finds no selected-task overlap, parse failure, or unresolved identity within that recorded history. The evaluation selection and denominators remain unchanged.

Coverage counts use exact history plus ordered-candidate equality. A/B/F/N each have 800 labeled visits and 799 distinct complete inputs. C has 1,544 and 1,543. Every arm starts from the same accepted warm-50 LoRA checkpoint, constructs a new AdamW optimizer, and keeps the same optimizer and RNG stream continuous for four 50-update blocks. The frozen 3,200-visit collection and cached teacher records are shared. A/B/F/N take the first 800 visits, while C takes the longest token-fitting prefix. The update logger fingerprints the actual group after each optimizer step. Reconstructing the fixed prefix-cycle schedule matches all 200 batch fingerprints for every arm. A/B/F/N use 800 record IDs four times each, and C uses 1,432 IDs twice plus 112 three times. Thus every planned record is verified in the completed optimization trace. The 1,511,726/1,510,554 B/C token costs are logical cached teacher supervision for this branch stage, excluding the shared warm stage. Training performs zero new teacher forwards.

The execution-repeat pairs were defined before score inspection and retain the frozen tasks, initialization, and protocol. The two-view advantage persists when repeating the original GPU assignments and when swapping devices, although individual task outcomes change. Table 7 reports these endpoints. Chronology is confounded with the device counterbalance, and F/N each retain one execution. Appendix A.5 reports the separate data-order follow-up.

Table 5: Original experiment-ID configurations. The mapping above defines each method. Costs are logical teacher-input tokens and success is out of 72.

Arm	State pool and target	Teacher input tokens	Success / 72
A	First 800 visits, identity	755,863	28
B	Same 800, two-view mean	1,511,726	36
C	1,544-visit frozen prefix, identity	1,510,554	26
F	Mean on 200 argmax-changed visits	1,511,726	32
N	Mean on 600 argmax-unchanged visits	1,511,726	33

Table 6: Complete original paired contrasts. All comparisons retain 72 tasks. L/R only counts exclusive successes for the left/right method.

Contrast	Role	Tasks (pp)	L/R only
B−A	Primary	+8 (+11.11 pp)	11 / 3
B−C	Primary	+10 (+13.89 pp)	13 / 3
N−A	Primary	+5 (+6.94 pp)	7 / 2
F−A	Support	+4 (+5.56 pp)	6 / 2
B−F	Support	+4 (+5.56 pp)	7 / 3
B−N	Support	+3 (+4.17 pp)	6 / 3

Table 7: Device assignments for the same-pool execution checks. All runs retain the same initialization and 72 tasks.

Execution pair	A mapping and success	B mapping and success	B−A
Original	GPU 1, 28/72	GPU 2, 36/72	+8
Repeated original assignments	GPU 1, 31/72	GPU 2, 37/72	+6
Swapped assignments	GPU 2, 31/72	GPU 1, 38/72	+7

Repeating the original assignments changes nine task outcomes in each method. Comparing the repeated and swapped assignments, six outcomes change for single-view training and eleven for two-view training.

A.4 Frozen-checkpoint sensitivity to the evaluation task set

Table 8: Recorded relationship between the earlier budget grid and the common protocol.

Item	Earlier fixed-budget grid	Common A/B/C protocol
Start and collection	Same accepted warm checkpoint and four-round, 3,200-visit stream collected by the frozen SFT branch-200 policy	Same accepted warm checkpoint and frozen stream
Student optimization	200 updates and 3,200 presentations with a new optimizer continuous across four blocks	Same settings with a deterministic prefix-cycle schedule
Allocation	One/two/four views at about 1.57M and 3.15M tokens, plus separate first-800/full-3,200 coverage cells	A/B use 800 visits. C uses the 1,544-visit prefix under B's exact 1.512M-token cap
Evaluation	Earlier 72-task internal-development set	Separate 72-task internal-training set frozen after training and before evaluation/score inspection
Result	View-depth and coverage directions are mixed	B−A=+8 and B−C=+10 in the original five-arm execution

Cross-evaluation applies each frozen checkpoint to both existing task selections. Table 9 combines the original endpoints with the complementary inference-only evaluations, retaining every selected task and failure. B and C refer to the original common-protocol checkpoints. The repeated executions remain a separate comparison.

Table 9: Frozen-checkpoint cross-evaluation. Every entry retains all 72 tasks in the named evaluation set.

Fixed checkpoint	Earlier 72-task set	New 72-task set
Earlier low budget, one view (1,600 visits)	38/72	28/72
Earlier low budget, two views (824 visits)	37/72	39/72
Common-protocol B, two views (800 visits)	39/72	36/72
Common-protocol C, one view (1,544 visits)	36/72	26/72

Evaluation-task composition changes the observed allocation contrast, including the ranking of the earlier low-budget checkpoints. For earlier two-view versus one-view checkpoints on the new set, the paired outcomes are 27 joint successes, 12/1 exclusive successes, and 32 joint failures. For B versus C on the earlier set, they are 33, 6/3, and 30. The evaluation sets are mutually disjoint and disjoint from all four training-task sets. Training inputs are shared or nested: B's visits occur in both earlier pools, C lies within earlier one-view, and earlier two-view lies within C. Teacher costs are approximately matched, as reported in Tables 4 and 5. This comparison isolates task-set sensitivity at fixed checkpoints. Selection timing and the shared initialization define its scope, while the training factors behind the remaining protocol difference remain unresolved. The original primary test family is unchanged.

A.5 Data-order and optimization-budget follow-up

The completed follow-up varies the ordering of the existing pools through independent data-order seeds. The warm checkpoint, initialization, runtime and evaluation settings, teacher cache, collection stream, and optimizer settings remain fixed. Each cycle permutes the shared 1,544-record universe. C consumes the full permutation and B the subsequence belonging to its original pool, preserving relative record order while forming minibatches separately. A continuous AdamW optimizer

saves both prespecified endpoints during training. Within each run, the methods execute serially on one device, with B then C for Run I and the reverse order for Run II. The runs use different devices, and their ordering schedules are retained in the experiment records.

Both endpoints use the unchanged evaluation tasks and episode cap. These tasks had already informed earlier analyses. The training contract fixed the runs, endpoints, and within-method comparisons before execution. The reporting schema and bootstrap implementation were recorded before score inspection, with the complete-result analysis following evaluation. Run I's B endpoint association incorporates an author correction exchanging the archived 200- and 400-update evaluation records. Paired summaries were recomputed accordingly. The result supplement retains the original labels, correction provenance, and timestamps distinguishing the training contract, analysis specifications, and reporting correction.

Table 10: Complete B/C endpoints. B-only/C-only counts retain task pairing. Initialization, runtime, and evaluation settings remain fixed. The follow-up runs vary data ordering only.

Run	Updates	B / 72	C / 72	B/C only	Gain (pp)
I	200	35	27	12 / 4	+11.11
I	400	41	35	14 / 8	+8.33
II	200	40	32	12 / 4	+11.11
II	400	40	34	11 / 5	+8.33

Table 11: Absolute arm changes and change in the allocation contrast. $\Gamma = (B-C)_{400}-(B-C)_{200}$.

Run	Contrast	Change (tasks)	Change (pp)
I	B400−B200	+6	+8.33
I	C400−C200	+8	+11.11
I	Γ	-2	-2.78
II	B400−B200	0	0.00
II	C400−C200	+2	+2.78
II	Γ	-2	-2.78

The corrected statistical analysis retains the original paired within-family procedure and primary test family. The task intervals for Γ include zero in both runs. Every task remains a paired unit across repeated observations. These comparisons hold initialization, runtime, and evaluation settings fixed and do not establish variability across independent initializations.

Each 200-update endpoint has 3,200 student presentations and each 400-update endpoint 6,400. B covers all 800 record IDs (799 distinct complete inputs), with multiplicity four/eight at 200/400. C covers all 1,544 IDs (1,543 distinct inputs), with 1,432 IDs twice and 112 three times at 200, then 1,320 four times and 224 five times at 400. These follow-up multiplicities are derived from the saved schedule and observed completed-update counts. Per-batch fingerprint matching applies to the original five-method experiment. Logical B/C teacher-input costs remain 1,511,726/1,510,554 tokens after the shared warm stage. Cached training requires zero new teacher forwards. Student input-token totals are unrecorded and remain unavailable in this report.

The complete follow-up retains a positive two-view advantage after longer training. Run I improves in both methods, while Run II improves only with the expanded pool. These policy outcomes are reported together at the prespecified endpoints. Direct alignment and absorption rates remain unmeasured.

B Earlier ALFWorld evidence

This appendix reports historical runs with different initializations under a separate protocol. Run I and Run II distinguish these executions. They are separate from the main study and its data-order follow-up.

Table 12: Complete termination and timing records. Success and cap truncation exhaust each 72-task denominator. Policy failures are zero. Update time is cumulative and excludes loading/checkpointing. Full training wall time includes loading, saving, and post-training work for the entire 400-update run, and each repeated entry refers to the same complete run. Evaluation wall time is separate. All times are seconds.

Run / arm	Updates	Success / truncation	Update time	Full train wall	Eval wall
I / B	200	35 / 37	1,032.8	2,173.1	623.6
I / B	400	41 / 31	2,070.6	2,173.1	614.9
I / C	200	27 / 45	1,041.2	2,258.7	666.1
I / C	400	35 / 37	2,076.5	2,258.7	645.9
II / B	200	40 / 32	1,008.6	2,108.3	615.0
II / B	400	40 / 32	2,007.4	2,108.3	604.7
II / C	200	32 / 40	1,030.2	2,244.4	654.7
II / C	400	34 / 38	2,061.9	2,244.4	648.2

B.1 Main policy endpoints

Table 13: Earlier policy endpoints. Each run column reports success, cap truncation, and policy failure out of 72 tasks.

Method	Run I success / truncation / failure	Run II	Mean success
SFT	41 / 31 / 0	33 / 39 / 0	37/72 (51.39%)
DPO	26 / 46 / 0	23 / 49 / 0	24.5/72 (34.03%)
Action ID	38 / 34 / 0	41 / 31 / 0	39.5/72 (54.86%)
DecisionOPD			
Semantic scorer	0 / 72 / 0	1 / 71 / 0	0.5/72 (0.69%)
GenOPD	9 / 63 / 0	1 / 53 / 18	5/72 (6.94%)

These cross-interface results describe the tested implementations. The matched action-ID factorial is the tighter intervention, while its subgroup interpretation remains exploratory.

B.2 Presentation target by state source

Table 14: Presentation targets by collection protocol. Entries are successes out of 72 tasks for each historical run.

State distribution	Teacher target	Run I	Run II
Fixed 800 decisions	Identity	29	33
Fixed 800 decisions	Aligned mean	36	37
Learner-refreshed	Identity	33	32
Learner-refreshed	Aligned mean	**38**	**41**

B.3 Supervision and refresh

Aligned averaging and soft supervision improve success under both collection protocols. Refreshed collection also improves the mean outcome for hard and soft targets. The interaction between averaging and refresh changes direction across initializations.

Table 15: Hard and soft supervision by collection protocol. Entries are successes out of 72 tasks.

State distribution	Supervision	Run I	Run II
Fixed 800 decisions	Hard CE	31	30
Fixed 800 decisions	Soft reverse KL	36	37
Learner-refreshed	Hard CE	33	35
Learner-refreshed	Soft reverse KL	**38**	**41**

Table 16: Earlier paired task comparisons. Left/right are successes out of 72. Difference and exclusive successes are task counts.

Comparison	Run	Left	Right	Gain	Left-only	Right-only
Full vs SFT	I	38	41	-3	6	9
Full vs SFT	II	41	33	+8	12	4
Soft vs hard, fixed	I	36	31	+5	9	4
Soft vs hard, fixed	II	37	30	+7	10	3
Refresh vs fixed mean	I	38	36	+2	6	4
Refresh vs fixed mean	II	41	37	+4	8	4
Soft vs hard, refreshed	I	38	33	+5	10	5
Soft vs hard, refreshed	II	41	35	+6	11	5
Mean vs identity, fixed	I	36	29	+7	11	4
Mean vs identity, fixed	II	37	33	+4	6	2
Mean vs identity, refreshed	I	38	33	+5	8	3
Mean vs identity, refreshed	II	41	32	+9	11	2

B.4 PAIRED TASK OUTCOMES

These historical comparisons vary initialization under the earlier protocol. The comparison with SFT changes direction across runs. They remain separate from the main study, which fixes initialization, runtime, and evaluation settings and varies data order only.

C PRESENTATION RETENTION AND ACTION-CONSEQUENCE DIAGNOSTICS

C.1 COMMON-STATE CROSS-EVALUATION

Under the historical protocol of Appendix B, each of four source policies contributes two deterministic states from all 72 saved trajectories in each run, yielding 1,152 records. All four final checkpoints answer the same states in both identity and reversed presentations.

Table 17: Common-state presentation sensitivity. TV is total variation. Consistency is a fraction, and reverse KL is in nats.

Run	Checkpoint	TV	Argmax consistency	Reverse KL
I	Fixed identity	0.331	0.604	0.599
I	Fixed mean	0.265	0.639	0.553
I	Refreshed identity	0.321	0.622	0.474
I	Refreshed mean	0.253	0.660	0.430
II	Fixed identity	0.341	0.595	0.656
II	Fixed mean	0.277	0.684	0.625
II	Refreshed identity	0.326	0.602	0.527
II	Refreshed mean	0.232	0.722	0.416

Mean targets reduce TV and reverse KL in all 16 source-by-run-by-collection comparisons. Argmax consistency improves in 15 of 16. In the highest teacher-conflict quartile, fixed TV falls from 0.576/0.530 to 0.429/0.409 and refreshed TV from 0.538/0.582 to 0.425/0.386. The relationship between larger TV reduction and task wins varies across outcome strata.

C.2 FULL TEACHER-PRESENTATION AUDIT

Across 1,903 saved ALFWorld decisions, mapped teacher argmax changes in 779 (40.94%). Mean Jensen-Shannon divergence is 0.1340 nats and mean total variation is 0.3329. Their 95th percentiles are 0.3779 and 0.7604. Candidate-count strata have 32.1% conflict for 9 to 16 candidates (212 records) and 42.0% for at least 17 (1,691).

C.3 COMPLETE ACTION-BRANCH ACCOUNTING

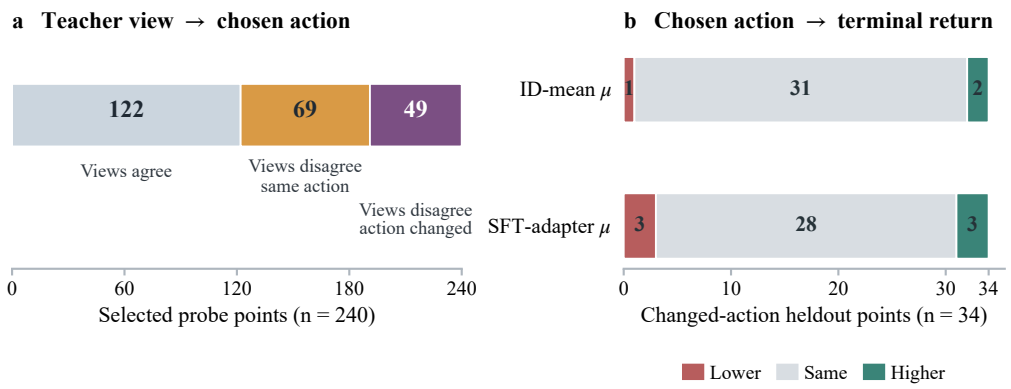

Figure 4: Greedy-action diagnostics measure a separate execution path. Panel (a) separates cross-view argmax disagreement from aggregation-induced action changes on the selected 240-point probe. Panel (b) reports terminal consequences only for 34 changed-action heldout points. These panels describe the local execution path. Distribution learning is evaluated through the separate student-training comparisons. Counts are probe-specific, and equal terminal returns can conceal different trajectories.

Table 18: Terminal consequences at 34 changed-action diagnostic-heldout points.

Frozen continuation	Mean action higher	Same return	Mean action lower	Mean paired difference
ID-mean	2	31	1	+0.0294
SFT adapter	3	28	3	0.0000

Common failure dominates the equal-return cells. ID-continuation ties contain 21 double failures and 10 double successes. SFT-continuation ties contain 17 and 11. Task-weighted summaries are +0.0577 when averaging only changed-action points within the 26 represented tasks, versus +0.00694 for ID and 0.0000 for SFT when all three selected points in all 48 heldout tasks enter. Appendix C gives the estimands. Branches preserve the original 30-step episode cap, so a saved prefix consumes part of that cap and only the remaining steps are available after the intervention. These returns are conditional on the selected points and frozen continuations. Action equivalence and distribution-learning value require their respective trajectory and training evidence.

Let I_t^c be the changed-action points for task t, $\mathcal{T}_c$ the 26 heldout tasks containing at least one such point, and Δ_{ti}^{μ} the return difference at point i. The changed-only estimate is

$$\widehat{\Delta}^{\mu}_{\text{changed}} = \frac{1}{26} \sum_{t \in \mathcal{T}_c} \frac{1}{|I_t^c|} \sum_{i \in I_t^c} \Delta_{ti}^{\mu}.$$

The all-heldout estimate averages all three selected points for each of 48 tasks,

$$\widehat{\Delta}^{\mu}_{\text{all}} = \frac{1}{48} \sum_{t=1}^{48} \frac{1}{3} \sum_{i=1}^{3} \Delta_{ti}^{\mu}.$$

The first estimates the mean within-task contrast among changed-action points. The second includes all selected points. Their different denominators account for the different effect estimates.

The selected probe contains 118 cross-presentation argmax disagreements and 49 cases where the aligned mean changes the identity-view action. Each continuation executes 395 distinct branches

without replay or runtime failures. Calibration uses 24 tasks with four points each, producing 92 zero, four positive, and no negative mean-return labels. The diagnostic holdout uses 48 tasks with three points each. The prospective allocator has a separate 72-task confirmation set. Earlier attempts that executed no branches because of a missing partition remain recorded as operational failures.

C.4 Student presentation stability on the fixed training pool

Evaluation retains every fixed training decision and completes all 6,400 checkpoint-view comparisons.

Table 19: Student presentation retention on the fixed training pool. Top-1 is agreement with the canonical action. JS is Jensen-Shannon divergence in nats.

Transformation	Mean-target top-1	Identity-target top-1	Mean-target JS	Identity-target JS
Reverse	67.75%	54.63%	0.0725	0.1680
Left cyclic shift	81.63%	74.88%	0.0333	0.0550
Label-only reassignment	83.13%	78.38%	0.0269	0.0427

Aggregate metrics favor mean-target training under all three transformations. The ordering within individual teacher-JS buckets varies.

D Static PIQA, HellaSwag, and ARC-Challenge results

D.1 Complete v2 answer counts

Table 20: Complete v2 static answer counts. Each cell gives correct answers and accuracy out of 512 domain-heldout items per dataset.

Dataset	Teacher identity	Teacher mean	Base 4B	Student identity	Student mean
PIQA	472 (92.19%)	481 (93.95%)	439 (85.74%)	453 (88.48%)	453 (88.48%)
HellaSwag	464 (90.62%)	468 (91.41%)	378 (73.83%)	429 (83.79%)	432 (84.38%)
ARC-Challenge	497 (97.07%)	503 (98.24%)	463 (90.43%)	478 (93.36%)	480 (93.75%)

Teacher averaging improves accuracy across all three datasets, while student gains are smaller and vary by domain. Newly corrected teacher answers rarely correspond to newly correct student answers. Table 21 gives the complete transitions on items reserved for evaluation.

D.2 Complete teacher/student transition table

Here **correct** and **wrong** denote correctness under the identity target followed by the aligned-mean target. The four student columns use the same transition order (C: correct. W: wrong).

D.3 Complete nested v1 pilot

The v1 and v2 splits are nested, so their outcomes share evaluation items. Pool size, updates, and repetition all change between versions.

Table 21: Teacher/student correctness transitions on the v2 domain-heldout items. The four student columns count correct (C) and wrong (W) transitions.

Domain	Teacher transition	Items	C to C	W to C	C to W	W to W
PIQA	correct to correct	466	431	2	2	31
PIQA	wrong to correct	15	9	0	0	6
PIQA	correct to wrong	6	3	0	0	3
PIQA	wrong to wrong	25	8	0	0	17
HellaSwag	correct to correct	462	414	5	3	40
HellaSwag	wrong to correct	6	2	0	0	4
HellaSwag	correct to wrong	2	0	0	0	2
HellaSwag	wrong to wrong	42	10	1	0	31
ARC-Challenge	correct to correct	497	475	1	0	21
ARC-Challenge	wrong to correct	6	3	1	0	2
ARC-Challenge	correct to wrong	0	0	0	0	0
ARC-Challenge	wrong to wrong	9	0	0	0	9

Table 22: Complete nested v1 static answer counts. Each cell gives correct answers and accuracy out of 128 domain-heldout items per dataset.

Dataset	Teacher identity	Teacher mean	Base 4B	Student identity	Student mean
PIQA	117 (91.41%)	122 (95.31%)	108 (84.38%)	108 (84.38%)	107 (83.59%)
HellaSwag	112 (87.50%)	113 (88.28%)	95 (74.22%)	104 (81.25%)	103 (80.47%)
ARC-Challenge	125 (97.66%)	126 (98.44%)	108 (84.38%)	115 (89.84%)	115 (89.84%)

E CROSS-TEACHER COMPARISON

ALFWorld Mistral single/mean students solve 10/72 and 13/72 tasks, with 5 mean-only and 2 single-only successes. Qwen matched arms solve 29/72 and 36/72, with 11/4 discordance. Mistral is therefore a directional replication only.

The Mistral ALFWorld teacher cache uses 1,600 forwards and 1,490,822 input tokens. Student deployment requires no teacher calls.

F IMPLEMENTATION, OBJECTIVES, AND ACCOUNTING

F.1 MODELS AND OPTIMIZATION

The second teacher is Mistral-Small-3.2-24B-Instruct-2506, revision `95a6d26c4bfb886c58daf9d3f7332c857cb27b43`.

Qwen3.5-27B revision `fc05daec18b0a78c049392ed2e771dde82bdf654` and Qwen3.5-4B revision `851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a` are fixed. The student trains 32,464,896 FP32 LoRA parameters across 496 tensors with rank 16, scale 32, zero dropout, and bias disabled. Backbone and language-model head remain frozen in BF16. ALFWorld uses AdamW at 10^{-5}, batch 16, zero weight decay, $\epsilon = 10^{-8}$, and gradient clipping at 1.0. The original core endpoints are fixed update 200. The allocation follow-up saves both 200 and 400 at prespecified update counts. Recorded core runs use Python 3.12.3, PyTorch 2.12.1+cu130, Transformers 5.17.0, PEFT 0.21.0, and CUDA 13.0 on RTX PRO 6000 96 GB GPUs.

F.2 VISIBLE STATE AND CANDIDATE CONTRACT

Policy input contains only the interaction transcript and complete environment-provided admissible action sequence. The complete environment-provided action order is preserved. The system instruction is "Choose the next executable action. Reply with exactly one action label from the listed options. Do not explain." The user message contains the interaction history followed by the complete labeled executable menu, rendered with the model's official non-thinking chat template. Labels are generated in Excel order and admitted only when tokenizer verification confirms a distinct ordinary single token at the assistant boundary. History and individual-action limits are 2,048 and 128 tokens. An overflow produces a typed policy failure while preserving the input-limit contract. Identity and reverse teacher views use the same history and choice identities. Reverse probabilities are mapped back before averaging.

The implemented soft loss floors target probabilities at 10^{-8}, renormalizes in FP32, and computes reverse categorical KL. Hard controls use first-index argmax and ordinary categorical cross-entropy with zero label smoothing.

F.3 COLLECTION AND EVALUATION

Refreshed rounds freeze the round-start policy, collect at least 800 eligible visits while completing the current episode, train on the first 800, and retain overshoot for audit. A decision instance is a recorded visit. Multiple visits can share an environment state.

For A/B/C/F/N, every arm loads the accepted `first-seed-sft-v2` warm-50 LoRA checkpoint through the `first-seed-sft-recovery-v4` receipt and creates a fresh AdamW optimizer. The same optimizer and RNG stream continue across four 50-update blocks. All arms read the same `shared-sft3200-v4` frozen collection and cached teacher bundle. That stream contains four 800-visit rounds collected by the frozen SFT branch-200 behavior policy. All five arms train from this shared frozen behavior stream. A/B/F/N cycle the first 800 visits in source order. C cycles the first 1,544. Actual post-update batch fingerprints match all 200 reconstructed batches per arm. The executable entry points are `experiments/core_confirmation_v2/{prepare.py,train.py}` and `experiments/offline_id_soft_kd_v1/train.py`. Prompt construction is in `src/decision_opd/policies.py`.

Evaluation uses greedy first-index argmax over the complete action set. Every selected task remains in its declared denominator. Evaluation uses 72 internal-development tasks for the earlier packages and 72 recorded-unexposed internal-training tasks for A/B/C/F/N. Success, unsuccessful environment termination, 30-step truncation, and typed policy input-bound failure are outcomes. Unexpected evaluator exceptions invalidate the artifact and preserve the declared denominator.

F.4 Earlier baseline and failure diagnostics

Warm-start success is 14/72 for generative decoding, 11/72 for action IDs, and 0/72 for the newly initialized semantic head. On 1,903 fixed-pool instances, extended ID fitting reduces KL from 0.3614 to 0.0533 and raises agreement from 71.57% to 88.39%, while task success changes 41 to 40. The scorer reduces KL from 1.1653 to 0.2633 and raises agreement from 55.49% to 75.67%, while success changes 2 to 7. On a common-505/fresh-180 diagnostic, scorer agreement is 420/505 on common states and 19/180 on fresh states, with fresh KL 2.037 and 3/24 task success. These results separate local target fit from control.

GenOPD's history-bound policy failures remain in the endpoint accounting in Table 13. Complete package wall times for historical runs I/II are 2,724.99/2,902.94 seconds for SFT, 4,015.65/4,021.66 for DPO, 5,679.53/5,119.52 for action ID, 9,324.37/9,433.77 for the scorer, and 17,271.89/15,418.03 for GenOPD. Offline cache reuse requires no incremental teacher calls while preserving logical supervision cost.

F.5 Static cost accounting

Static v2 uses 6,144 teacher forwards across training and heldout analysis, comprising 3,072 for training targets and 3,072 for teacher evaluation. The six student branches use 3,072 training presentations. Recorded successful training phases sum to 1,312.5 seconds and recorded heldout components to 618.1 seconds. Both cover the named successful phases only. Full development wall time has broader scope.

F.6 Alternate-presentation evaluations for the simple controls

Table 23: Earlier fixed-pool matched controls. All students share the earlier training pool, evaluation tasks, and implementation. Entropy matching uses the mean target to calibrate global target concentration, so its accounting includes information from that target.

Method	**Alternative explanation tested**	**Success**
Single-view identity targets	Reference: one teacher presentation per visit	29/72
Entropy-matched identity targets	Global concentration matching applied to identity targets	30/72
Student presentation augmentation	Variation only in the student's menu presentation	29/72
Aligned two-view mean targets	Full aligned distribution from a second presentation	36/72

Table 24: Alternate student-presentation endpoints for the earlier simple controls. All entries retain the 72-task denominator, including policy failures.

Endpoint	**Label shift**	**Fixed random permutation**	**Reverse**
Single view	30/72	36/72	35/72
Aligned mean	33/72	39/72	35/72
Entropy-matched single	30/72	37/72	34/72
Student presentation augmentation	30/72	35/72	26/72

The label-shift entropy run records one policy failure. The reverse augmentation run records two. All failures stay in the denominator. These alternate-presentation robustness diagnostics complement the separately retained canonical evaluation.

G ARTIFACT PROVENANCE

The result supplement groups machine-readable summaries by experiment, covering action consequences, fixed teacher budgets and state coverage, prospective allocation, post-hoc rate controls, cross-teacher transfer, and the earlier ALFWorld and static-domain studies. Each group retains its task denominators, paired outcomes, teacher costs, and failure accounting. The accompanying inventory maps these groups to their source paths and analysis programs. The core coverage reconstruction is recorded in `artifacts/results/v20-sprint-protocol-evidence/coverage-reconstruction.json`. It matches reconstructed batches to the fingerprints written from actual update groups after optimizer steps.

The research-source supplement retains the implementation and fixed experiment definitions. It excludes credentials, machine configuration, model weights, and licensed data. Earlier failed attempts and the version-17 manuscript remain in the project history. Their results remain alongside the new experiments.