Matched-Control Tests of Strategy Alignment in Routed Distillation

Rui Li **Shuang Cao** **Ruihua Liu**
Hill Research

Abstract

Reasoning distillation must turn diverse teacher solutions into consistent supervision for shared experts. We study strategy alignment as two coupled requirements: grouping traces by reasoning strategy and assigning those groups consistently across inputs. Our matched-control protocol tests both requirements while fixing the teacher-trace budget and routed capacity. Auto-v4-Open implements them with a shared strategy representation and domain-wide adapter labels. At 4.80B teacher-trace tokens, it reaches 58.62% on a five-domain out-of-distribution average, exceeding routing-only and matched-shuffle controls by 2.46 and 2.04 percentage points. Coherence-preserving controls remain 1.68–1.91 points behind, removing cross-input alignment costs 2.60 points, and verifier-free strategy discovery retains a 2.12-point gain over routing alone. The assignment advantage transfers to native sparse-MoE and soft-prefix routing and complements modern distillation recipes, reaching 60.31% with Uni-OPD. The fixed pipeline also exceeds routing-only MoA on MMLU-Pro, LiveBench, BBEH, and OlympiadBench by 4.91, 3.99, 3.70, and 2.87 points. These results establish strategy-consistent supervision as an actionable source of transfer gains across routed distillation designs.

1 Introduction

A teacher can solve a mathematics problem algebraically or geometrically, implement a program iteratively or recursively, and complete an embodied task through exploration or a shortcut. A routed student must turn these alternatives into supervision shared across problems. This introduces two distinct choices: which traces belong together, and which expert receives each group. Even a coherent local partition has arbitrary cluster indices. Reusing those indices across inputs can send different strategies to the same expert, obscuring the specialization that routing is intended to preserve. Cross-input strategy assignment therefore determines how reasoning diversity becomes useful training signal.

Our study isolates this supervision choice through matched interventions. Latent-Routing MoA measures the benefit of allocating several routes; Matched-Partition-Shuffle (MPS) retains coarse partition structure while changing trace membership; coherence-preserving controls retain ordinary similarity while varying strategy information. A separate alignment ablation sends local cluster indices directly to shared adapters. Together, these comparisons test the content of each group and the consistency of its expert assignment. They use 4.80B teacher-trace tokens with routed capacity and router depth matched within each stack; verifier-budget-equalized controls additionally match the discovery-time query allowance.

We operationalize strategy assignment with Auto-v4-Open, a zero-manual partition pipeline. A verifier labels same-strategy trace pairs; a shared contrastive encoder maps those judgments into a domain-wide inventory; global medoid labels then index the adapters. The local-to-global step is essential: aligned training reaches 58.62% on the equal-domain primary-shift average A1, compared with 56.02% when local cluster indices directly address shared adapters. The aligned pipeline

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

improves over MoA and MPS by 2.46 and 2.04 percentage points (pp), respectively. The coherence-preserving controls remain 1.68–1.91 pp behind, while verifier-free strategy discovery retains a 2.12 pp gain over MoA. This combination of interventions ties the benefit to strategy-consistent supervision across inputs.

The intervention transfers across mechanisms and complements stronger distillation recipes. Native top-8 routing in Qwen3-30B-A3B and routed soft prefixes in Gemma-2-9B retain gains of 1.92 and 1.86 pp over their routing-only controls. Within modern on-policy and multi-teacher recipes, semantic assignment adds 2.13–2.46 pp over routing-only assignment; Uni-OPD with semantic assignment reaches 60.31%. This connects a controlled mechanism result to a concrete improvement in competitive students.

The contribution has three connected parts. First, the control protocol distinguishes strategy membership from cross-input label consistency and tests both against capacity, coherence, and verifier-budget explanations. Second, the shared-label construction operationalizes this distinction, with an alignment ablation and functional route interventions linking assignment to specialization. Third, the cross-mechanism and within-recipe comparisons establish assignment as a reusable design choice; the external slices and full/simplified operating points characterize its strategy-diversity and inference-cost tradeoffs.

2 Strategy membership and cross-input alignment

For input x_i, let $\mathcal{T}_i$ be the retained teacher traces. We distinguish local groups $C_\ell^{(i)}$, used to construct pair queries, from domain-wide labels $z_d(t) \in \{1, \ldots, m_d\}$, which index shared adapters. For partition inducer q, teacher-trace budget B, routed trainable capacity P, and router depth R, the assignment effect on distribution $\mathcal{D}$ is

$$\Delta_{\mathrm{src}}(q, \mathcal{D}) = \mathrm{Acc}(S(q; B, P, R), \mathcal{D}) - \mathrm{Acc}(S(q^{\mathrm{shuffle}}; B, P, R), \mathcal{D}). \tag{1}$$

Here q^{shuffle} is the constrained MPS transformation of q. MoA provides a complementary routing-only comparison. A common permutation of global labels simply renames the adapters when their parameters and gate outputs are permuted with them. Input-specific permutations instead change the traces pooled into each shared adapter, even when every local cluster keeps its members and coherence. This distinction motivates a separate alignment ablation alongside the partition-content controls; Appendix B spells out the corresponding supervision sets.

2.1 From local trace groups to global adapter labels

The main teacher is Qwen-2.5-72B-Instruct [Qwen Team, 2024]. Best-of-32 generation supplies candidates; the discovery stage retains 24 verified traces and constructs 128 ordered pair queries per input. Local k-medoids groups guide query formation. GPT-OSS-120B-Instruct judges whether each pair uses the same strategy at temperature 0; uncertain judgments below confidence 0.6 are excluded from encoder training. A six-layer, 384-dimensional InfoNCE encoder uses 128-token step summaries, contrastive temperature 0.07, batch 256, and 8,000 steps. This learned representation defines Auto-v4-Open; frozen E5-Mistral embeddings define the generic-embedding controls.

The trained encoder embeds pooled training traces within each domain. Global k-medoids then selects $m_d \in \{2, 3, 4, 5\}$ using normalized spectral gap with a silhouette floor of 0.25 and smaller-m tie breaking. Each local medoid receives the nearest global medoid under cosine distance. A Kuhn–Munkres consistency check aligns the domain medoid set; if the top-two cosine margin is below 0.02, the local cluster inherits the majority global label of its members. The local and global cluster counts need not agree: several local groups may map to the same global label. Only the global label addresses an adapter, so route j receives a consistent strategy assignment across inputs. Algorithm 1 states the training sequence and Appendix B reports stability.

The modal inventory is three strategies for MATH, HumanEval+, ALFWorld, and GPQA, and two for BBH. It is fixed before the ten student seeds. Across 50 re-clustering seed-domain cells, only two change their selected count; fixed-modal training reaches 58.55%, compared with 58.62% for the selected inventory. Thus fixing the mode count retains the assignment advantage across the evaluated domains.

2.2 Routed student and optimization

The main student uses a frozen Llama-3.1-8B-Instruct backbone [Grattafiori et al., 2024], rank-16 LoRA adapters [Hu et al., 2022], and a two-layer MLP gate on the last-token backbone representation. Its softmax outputs $w_j(x)$ weight adapter losses during training; top-1 routing selects the adapter at inference. Each adapter samples traces carrying its global label, preserving within-strategy variation. Matched routed arms use the same inventory, adapter shapes, and gate; rank-16 and rank-48 single-adapter KD are separate capacity comparisons.

The local-to-global map defines the supervision actually seen by each adapter:

$$G_j^{(i)} = \{t \in \mathcal{T}_i : z_d(t) = j\}, \qquad \mathcal{J}_i = \{j : G_j^{(i)} \neq \varnothing\}, \qquad \widetilde{w}_j(x_i) = \frac{w_j(x_i)}{\sum_{k \in \mathcal{J}_i} w_k(x_i)}. \quad (2)$$

Local groups $C_\ell^{(i)}$ construct pair queries; the sets $G_j^{(i)}$ collect all traces assigned to the same global label, including traces from different local groups. The denominator restricts the KD mixture to populated labels. For a sampled verified target $t = (t_1, \ldots, t_{|t|})$, its sequence-target loss is

$$\ell_{\mathrm{KD}}(\theta_j; x_i, t) = -\frac{1}{|t|} \sum_{r=1}^{|t|} \log p_{\theta_j}(t_r \mid x_i, t_{<r}). \quad (3)$$

Writing $\mathcal{L}_{\mathrm{KD}}(\theta_j, G_j^{(i)})$ for the average of these losses over sampled targets, the batch objective combines gated KD with coverage, usage, and inter-adapter separation:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{J}_i} \widetilde{w}_j(x_i) \mathcal{L}_{\mathrm{KD}}(\theta_j, G_j^{(i)}) + \lambda_{\mathrm{cov}} \mathcal{L}_{\mathrm{cov}} + \lambda_{\mathrm{use}} \mathcal{L}_{\mathrm{use}} + \lambda_{\mathrm{div}} \mathcal{L}_{\mathrm{div}}, \quad (4)$$

$$\mathcal{L}_{\mathrm{cov}} = \sum_j [0.08 - \bar{w}_j]_+, \qquad \mathcal{L}_{\mathrm{use}} = \sum_j [0.05 - u_j]_+,$$

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{\sum_{a \neq b} \cos(h_{ia}, h_{ib})}{m(m-1)}. \quad (5)$$

Here $m = m_d$ for the domain inventory, $\bar{w}_j = |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} w_j(x_i)$ is batch-average soft gate mass, and h_{ij} is adapter j's hidden-state representation on input x_i. The coverage and usage terms use the full gate distribution; the KD term uses the normalized weights in Eq. 2. The training usage surrogate is $u_j^{(t)} = \beta \, \mathrm{stopgrad}(u_j^{(t-1)}) + (1-\beta)\bar{w}_j^{(t)}$; its current-batch term carries gate gradients. Hard top-1 selection frequencies are evaluation statistics. Minimizing the positive cosine term reduces inter-adapter similarity. The KD targets are verified token sequences; contrastive temperature 0.07 belongs to discovery, not to this sequence-target loss. A hard-negative gate batch is applied every 200 steps to discourage collapsed assignments.

2.3 Training sequence

Algorithm 1 separates query formation, global inventory construction, and student training. The handoff between discovery and distillation is the assignment map z_d: local indices guide pair queries, and the global labels define the shared supervision sets in Eq. 2.

2.4 Controls and resource matching

MoA trains every adapter on the undivided retained trace pool, using the same gate and adapter inventory. MPS instead changes the trace-to-label assignments in Eq. 2 while retaining teacher traces, answers, KD targets, and routed capacity. It performs 10,000 accepted pairwise swaps per domain and seed, accepting a swap only if none of the matched marginal distances increases. Cluster cardinality stays within two traces, normalized length Wasserstein-1 distance is at most 0.04, difficulty-mix and cardinality checks have $p > 0.85$ and $p > 0.97$, and verifier pass rates stay within 0.3 pp. Domain-specific metadata supplies randomization strata, not answer-label supervision. This is a finite constrained randomization procedure.

SOCP maximizes within-cluster coherence subject to external-strategy AMI at most 0.10 and the MPS marginal constraints. ECPS uses frozen generic E5 embeddings with coherence matching.

Algorithm 1 Global-label routed distillation.

Require: Common retained teacher traces, frozen backbone, gate, domain adapter inventories
Form local k-medoids groups and 128 ordered pair queries per input
Obtain verifier pair labels; train the shared InfoNCE trace encoder
Pool training embeddings per domain; select global medoids and m_d
Assign local medoids to nearest global labels; check global-medoid permutation consistency
If the top-two margin is below 0.02, use majority member global labels
Form the global-label target sets $G_j^{(i)}$ in Eq. 2
Initialize the gate uniformly and the adapters from the backbone
for step $= 1, \ldots, T$ **do**
    Sample a batch and targets from $G_j^{(i)}$; normalize KD weights over populated labels
    Compute coverage, differentiable EMA-usage, and positive mean-cosine penalties
    Update gate and adapters by Eq. 4; keep backbone frozen
    Apply the hard-negative gate batch every 200 steps
end for
return Global-label adapters with top-1 inference routing

Task-metadata partitioning groups by difficulty, code complexity, episode length, task family, or subject. The verifier-equalized MoA control spends the same query allowance on re-filtering under the unchanged token cap; verifier-equalized MPS uses it for difficulty re-stratification. Appendix B gives their semantic and functional diagnostics.

The matched quantity is the 4.80B-token teacher-trace pool used for KD, with student capacity and router depth matched separately. Candidate generation, retained targets, and repeated optimization exposures are distinct stages: repeating training batches does not create additional unique teacher traces. Discovery-time verifier work is accounted separately, including 84.2 A100-hours for Open. The strengthened controls specifically test whether that additional query budget explains the accuracy gain.

3 Experiments

3.1 Evaluation and training setup

A1 is the equal-domain mean of five fixed primary OOD shifts: MATH-500 difficulty, HumanEval+ Python-to-TypeScript, ALFWorld combined, BBH task-family, and GPQA subject-family [Hendrycks et al., 2021, Liu et al., 2023, Shridhar et al., 2021, Suzgun et al., 2023, Rein et al., 2023]. The complete original grid has three shifts per domain. Scorers are exact match for mathematics and BBH, test-suite pass@1 for code, episode success for ALFWorld, and multiple-choice correctness for GPQA. External evaluations use MMLU-Pro, LiveBench, BBEH, and OlympiadBench [Wang et al., 2024, White et al., 2025, Kazemi et al., 2025, He et al., 2024]. The discovery choice is fixed before external evaluation; the manual canonicalizer provides a reference and Auto-v4-Ensemble is a separately named sensitivity variant.

Stack A uses ten student seeds; stacks B/C and the two additional routing mechanisms use eight. MMLU-Pro uses ten main-stack seeds; LiveBench, BBEH, and OlympiadBench use five per slice. AdamW training uses 20K steps, learning rate 2×10^{-4} to 2×10^{-5} with cosine decay, batch 256, sequence length 2,048, and bf16. Hardware is $8\times$NVIDIA A100 80GB SXM4 with NVLink, CUDA 12.4, driver 550.90.07, PyTorch 2.3.1, and FlashAttention-2. Reported paired 95% BCa intervals use 10,000 resamples of evaluation examples with system pairing preserved. A1 retains equal-domain weighting; intervals quantify test-pool uncertainty. Per-slice intervals are marginal, not simultaneous coverage statements.

3.2 Strategy content and consistent labels explain the gain

Auto-v4-Open improves A1 over MoA by $+2.46$ pp (paired CI $[1.76, 3.16]$) and MPS by $+2.04$ pp ($[1.36, 2.72]$), with positive gaps on all five primary shifts (Table 1). In-distribution averages differ

Table 1: Matched teacher-trace comparisons on the five fixed primary shifts (accuracy, %). Routed controls share the adapter inventory and gate; A1 weights domains equally. The manual pipeline is a reference.

Method	MATH	HE+	ALF	BBH	GPQA	ID avg.	A1 OOD (95% CI)
Standard KD	58.24	61.68	54.17	58.06	36.83	64.82	53.80 [53.14, 54.45]
BM-SC-32	60.37	63.11	55.76	59.38	38.14	65.88	55.35 [54.68, 56.02]
Rand.-Part.	59.55	62.62	55.08	58.67	37.64	65.32	54.71 [54.04, 55.38]
Embed.-Part.	60.28	63.31	55.76	59.19	38.22	65.52	55.35 [54.66, 56.04]
MPS ctrl.	61.12	64.15	57.08	60.40	40.15	65.52	56.58 [55.88, 57.27]
MoA ctrl.	60.93	63.86	56.58	60.11	39.32	65.64	56.16 [55.43, 56.87]
MODEDISTILL-*Auto-v4-Open*	63.69	65.77	58.87	62.79	41.98	65.82	58.62 [57.91, 59.31]
MODEDISTILL-manual	63.74	65.81	58.92	62.83	42.05	65.80	58.67 [57.96, 59.37]

by at most 0.30 pp among these routed arms. The separation therefore appears primarily under distribution shift.

Table 2: Controls separating strategy assignment, coherence, and verifier budget. All rows use stack A and ten seeds. Gaps are Open minus the control, with paired 95% BCa intervals. Cosine and AMI describe the induced partition. Unaligned uses local cluster indices for shared adapters; Free uses verifier-free discovery. Dashes denote metrics outside the reported comparison.

Assignment/control	A1	Open gap [95% CI]	Cosine	AMI
MoA	56.16	2.46 [1.76,3.16]	—	—
MPS	56.58	2.04 [1.36,2.72]	0.436	0.528
MoA, equal verifier budget	56.44	2.18 [1.48,2.88]	—	—
MPS, equal verifier budget	56.81	1.81 [1.13,2.49]	—	—
Task metadata	56.63	1.99 [1.31,2.67]	—	—
SOCP	56.71	1.91 [1.23,2.59]	0.609	0.094
ECPS	56.94	1.68 [1.01,2.35]	0.618	0.549
Unaligned local indices	56.02	2.60 [1.91,3.29]	—	—
Auto-v5-Free	58.28	—	—	0.641
Auto-v4-Open	58.62	—	0.612	0.702

Ordinary coherence does not recover the strategy-assignment gain (Table 2). SOCP matches Open's within-cluster cosine to within 0.003 while suppressing strategy AMI, yet remains 1.91 pp behind. ECPS has slightly higher cosine than Open but remains 1.68 pp behind. Equalizing verifier budget improves both nulls, leaving residual gaps of 2.18 and 1.81 pp. These controls distinguish aligned strategy supervision from similarity, metadata, and additional verifier work.

Cross-input consistency provides the second discriminating test. Sending local indices directly to shared adapters reduces A1 from 58.62% to 56.02%; the paired alignment effect is $+2.60$ pp ($[1.91, 3.29]$). Fixing the modal inventory gives 58.55%, and per-seed reselection changes A1 by just 0.02 pp ($[-0.11, 0.15]$; Appendix B). The aligned-versus-unaligned contrast holds the budget, loss, schedule, and parameter count fixed; the count-selection ablations separately show that the gain persists when selection is removed.

The gain also survives verifier-free discovery. Auto-v5-Free combines trace-graph minimum-description-length structure and frozen student activations. It reaches 58.28% (CI $[57.57, 58.98]$), $+2.12$ pp over MoA ($[1.42, 2.82]$), with external-tag AMI 0.641. Strategy-sensitive grouping can therefore preserve most of the gain using a different discovery signal.

A blinded pairwise audit gives Open a same-strategy AUC of 0.867, compared with 0.712 for MoA and 0.749 for generic embedding clusters (Appendix H). Functional tests connect this separation to expert behavior. On 2,000 externally authored, answer-preserving strategy-switch pairs, Open

changes route on 51.6% of pairs, versus 27.4% for SOCP and 30.6% for ECPS. Forcing the externally tagged wrong strategy incurs penalties of -5.03, -1.57, and -1.83 pp, respectively (Table 10). The stronger response to a strategy change and the larger cost of a mismatched route support strategy specialization beyond generic cluster coherence.

The manual-reference ablations hold the discovered partition fixed and vary student training. Uniform routing and removing the diversity loss reduce A1 by 3.44 and 3.13 pp, respectively (Table 3). Stable global labels and specialized student routes therefore support complementary parts of the transfer gain.

Table 3: Key routed-student ablations on A1. All runs use the manual reference configuration and test whether routed capacity, anti-collapse losses, and multi-trace supervision remain necessary after partition induction.

Variant	A1 OOD	Δ vs. full
Full routed student	58.67	0.00
Uniform routing	55.23	-3.44
$-$ Diversity loss	55.54	-3.13
$-$ Coverage loss	57.30	-1.37
$-$ Usage loss	56.83	-1.84
Single teacher trace	54.76	-3.91

3.3 Transfer across routing mechanisms

Table 4 applies the complete control ladder to native sparse-MoE and prefix routing. Stack D uses Qwen3-30B-A3B (30.5B total and 3.3B active parameters, 128 experts), with mode-conditioned expert-group bias in native top-8 token routing in the last 12 MoE blocks. This bias acts on the native router without an additional adapter gate. Stack E uses Gemma-2-9B with routed 64-token soft-prefix experts. Each stack matches trainable capacity within its routed arms, uses 4.80B teacher-trace tokens, and evaluates eight paired seeds on the unchanged A1 shifts. Stacks B/C separately test teacher/student changes within the adapter family (Appendix J).

Table 4: Assignment effects across three routing mechanisms. Open denotes semantic assignment using the fixed discovery pipeline. Intervals are paired 95% BCa intervals on A1 differences (pp).

Routing mechanism	KD	MoA	MPS	Open
A: LoRA adapters	53.80	56.16	56.58	58.62
D: native MoE	57.34	59.61	60.02	61.53
E: soft prefixes	55.12	57.30	57.71	59.16
	Open $-$ MoA		Open $-$ MPS	
A	2.46 [1.76,3.16]		2.04 [1.36,2.72]	
D	1.92 [1.21,2.63]		1.51 [0.84,2.18]	
E	1.86 [1.15,2.57]		1.45 [0.79,2.11]	

The semantic arm exceeds both routing-only and shuffled controls under token-level expert routing and sequence-level prefix routing. Repeating the control ladder within each stack establishes that the assignment benefit transfers across these mechanisms, with effect sizes specific to each architecture.

3.4 Complementarity with modern distillation recipes

We compare implementations of GKD-style OPD, rubric-based OPD, Uni-OPD, compatibility-aware multi-teacher CoT, and domain-targeted augmentation [Agarwal et al., 2024, Fang et al., 2026, Hou et al., 2026, Cui et al., 2026, Stacey and Rei, 2024]. These are adaptations to the shared student and evaluation protocol. Every recipe uses Llama-3.1-8B, ten seeds, and 4.80B total teacher tokens; multi-teacher arms divide this total across Qwen-2.5-72B, GPT-OSS-120B, and the stack-C teacher. The comparisons vary the trajectory or teacher signal at the recipe level and the trace-to-route assignment within each recipe.

Table 5: Modern recipe comparisons on A1. The lower panel summarizes the within-recipe partition intervention across all five recipes; each range is over recipe-level paired point estimates.

Unpartitioned recipe	A1
GKD-style OPD	56.71
Rubric-based OPD	57.24
Uni-OPD	57.48
Compatibility-aware multi-teacher CoT	56.93
Domain-targeted multi-teacher augmentation	55.86
Semantic assignment − MoA, within recipe	2.13–2.46 pp
Semantic assignment − MPS, within recipe	1.75–2.04 pp
Uni-OPD with semantic assignment	60.31

Open exceeds the strongest unpartitioned recipe, Uni-OPD, by $+1.14$ pp ($[0.47, 1.81]$). The within-recipe comparison then holds that recipe's supervision design fixed: semantic assignment exceeds its routed MoA and MPS arms throughout (Table 5). Uni-OPD with semantic assignment reaches 60.31%, 2.83 pp above unpartitioned Uni-OPD and 1.69 pp above standalone Open. These two differences compare complete configurations; the within-recipe MoA/MPS contrasts isolate assignment at matched routed capacity. The practical consequence is to improve trajectory feedback and expert assignment together; the measured assignment gain persists after strengthening the underlying distillation recipe.

3.5 External breadth and boundary conditions

Table 6: External evaluation of the fixed *Auto-v4-Open* headline row. Columns report A1, four external benchmark aggregate means, and the A1 gap to MoA. Teacher/student transfer within the adapter family is detailed in Appendix J.

Method	A1 (95% CI)	External benchmark mean				Δ_{MoA} (A1)
		MMLU-Pro	Live-Bench	BBEH	Olymp.	
Standard KD	53.80 [53.14, 54.45]	42.18	42.71	28.04	27.18	−2.36
Trace-Restr. KD	55.21 [54.55, 55.87]	43.52	44.18	30.11	28.86	−0.95
Contr. MR-KD	55.98 [55.29, 56.65]	44.07	44.67	30.87	29.41	−0.18
MPS ctrl.	56.58 [55.88, 57.27]	45.82	46.60	32.09	29.94	+0.42
MoA ctrl.	56.16 [55.43, 56.87]	44.96	45.79	31.28	29.64	0.00
Open (Auto-v4-Open)	58.62 [57.91, 59.31]	49.87	49.78	34.98	32.51	+2.46
Simplified Open ($m = 2$)	57.97 [57.26, 58.68]	49.03	48.90	33.85	31.69	+1.81
Manual reference	58.67 [57.96, 59.37]	50.12	50.22	35.24	32.78	+2.51

The fixed Open pipeline improves over MoA by $+4.91$, $+3.99$, $+3.70$, and $+2.87$ pp on MMLU-Pro, LiveBench, BBEH, and OlympiadBench (Table 6). The exact Open slice results are positive in all 14 MMLU-Pro categories, all seven LiveBench categories and all eleven monthly slices, 21 of 24 BBEH subtasks, and seven of eight OlympiadBench categories. Appendix C reports coverage, named category contrasts, and the four negative cells.

Effective strategy-family count is $\text{ESF} = \exp(-\sum_k p_k \log p_k)$, computed from external tags in the verified teacher pool without outcome accuracy. Across 68 slices, ESF correlates with Open's gain over MoA at Spearman $\rho = 0.74$ (CI $[0.62, 0.83]$). The four negative cells have ESF below 1.25 and mean gap -0.64 pp; 1.25 is an observed boundary in this evaluation. LiveBench knowledge recall (ESF 1.38, $+1.88$ pp) and ALFWorld tool reconfiguration (1.41, $+1.09$ pp) show small positive effects at low strategy diversity. Figure 1 displays these heterogeneous effects. The association

links the benefit of specialization to the diversity of strategies available to specialize in, giving the aggregate improvement a concrete task-level interpretation.

ALFWorld tool reconfiguration has a separate precision check. The original 480-episode pool gives Open/MoA 58.02/56.87, a $+1.15$ pp gap ($[-0.12, 2.41]$). Expanding the same construction to 1,920 episodes gives 57.94/56.85, $+1.09$ pp ($[0.31, 1.87]$), and $+0.74$ pp over MPS ($[0.06, 1.42]$). The original grid retains 14 of 15 positive marginal intervals; replacing only this cell by the expanded evaluation gives 15 of 15. A1 stays 58.62 because its ALFWorld primary shift is combined.

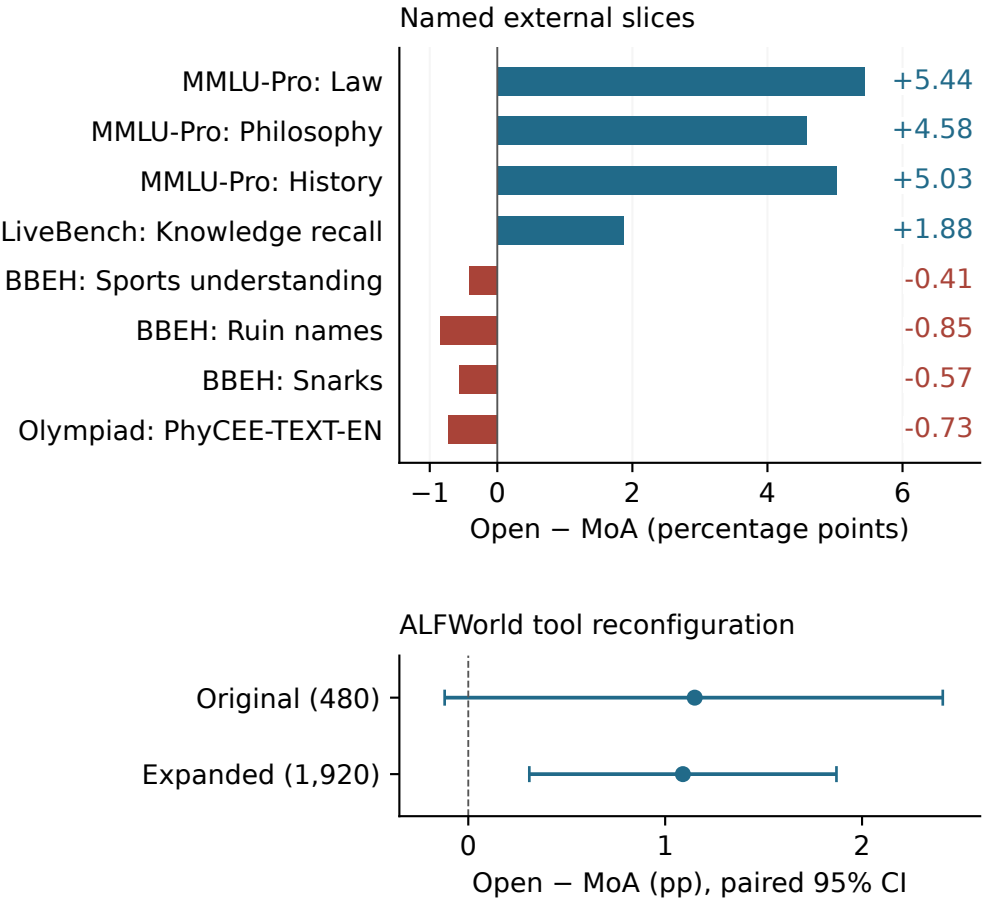


Figure 1: Observed Open effects: named external slices (top) and the original versus expanded ALFWorld tool evaluation (bottom, paired 95% BCa intervals). Negative cells remain visible. Across the 68-slice analysis, strategy diversity correlates with transfer gain ($\rho = 0.74$, CI $[0.62, 0.83]$); the displayed named slices are a subset, not the underlying correlation sample.

3.6 Off-domain routing and practical cost

On MMLU-Humanities, MMLU-Social-Sciences, ETHICS-hard, and a 20-task LegalBench reasoning subset, Open exceeds MoA by $+0.78$, $+0.87$, $+0.83$, and $+0.81$ pp, respectively. The mean is $+0.82$ pp ($[0.41, 1.23]$), with a $+0.29$ pp mean improvement over frozen Llama-3.1-8B-Instruct. These suites are excluded from training and discovery. Router entropy increases from 0.49 to 0.71 nats, and minimum EMA usage remains 0.11, above the 0.05 floor. Max-gate uncertainty yields OOD AUROC 0.813, compared with 0.662 for MoA and 0.671 for MPS (Appendix D).

Open adds 11.8% inference tokens over KD and 0.31 ms/example gate latency on an A100 at batch 16. Simplified Open with two modes reaches 57.97% at 5.8% token overhead, retaining 86.5% of the full gain over KD. The full and simplified configurations offer distinct accuracy/cost choices;

discovery is a one-time cost, whereas inference overhead recurs per request. Appendix E consolidates the evaluated operating conditions.

4 Related work

Rationale distillation transfers teacher-generated intermediate steps [Hsieh et al., 2023, Magister et al., 2023, Fu et al., 2023, Mukherjee et al., 2023]; verification and self-consistency improve which trajectories are selected [Lightman et al., 2024, Wang et al., 2023, Zelikman et al., 2022]. Sparse expert architectures change capacity allocation and routing [Shazeer et al., 2017, Lepikhin et al., 2021, Lewis et al., 2021, Fedus et al., 2022, Jiang et al., 2024, Dai et al., 2024]. Our comparison holds routed capacity and gate depth fixed within each mechanism and changes the supervision assigned to each route. The local/global distinction sharpens this contrast: preserving diversity among teacher solutions and giving a shared expert consistent strategy targets are separate operations.

GKD addresses the student/teacher trajectory mismatch with teacher feedback on student-generated sequences [Agarwal et al., 2024]. Recent OPD analysis connects success to compatible thinking patterns and informative teacher capabilities [Li et al., 2026]. Rubric-based OPD supplies structured response-level feedback [Fang et al., 2026], while Uni-OPD combines student-state balancing with outcome-guided calibration of teacher supervision [Hou et al., 2026]. COMPACT weights teacher supervision using reasoning consensus, adaptability, and difficulty [Cui et al., 2026]. Domain-targeted augmentation changes the training distribution through generated target-domain examples and selective upsampling [Stacey and Rei, 2024]. Our within-recipe comparisons isolate an additional variable: which route receives each strategy. The positive effects inside these recipes establish complementarity with trajectory selection, calibrated feedback, and teacher mixing. Coherence-matched partitions further distinguish this assignment question from clustering traces by ordinary similarity, while the unaligned ablation tests whether local group identities can support shared specialization.

5 Discussion and conclusion

Strategy diversity and label consistency guide different design decisions. The ESF analysis relates the value of specialization to the strategies present in the teacher pool, with the 1.25 boundary describing the evaluated slices (Section 3.5). The alignment ablation shows the cost of assigning those strategies inconsistently, while the manual-reference ablations show that uniform routing or removing the diversity penalty also erodes the gain (Table 3). Open's 84.2 A100-hour discovery cost and the full/simplified inference tradeoff make both one-time and recurring costs explicit.

The matched interventions establish a concrete source of routed-distillation gains: strategy-sensitive trace groups with consistent identities across inputs. Auto-v4-Open implements this assignment, coherence and budget controls distinguish its effect, and native MoE, soft-prefix, and within-recipe comparisons demonstrate its reuse. Organizing teacher supervision into stable expert responsibilities is a productive design choice for transferring diverse reasoning strategies.

Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

References

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 *International Conference on Learning Representations*, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html`.

Jin Cui, Jiaqi Guo, Ruixuan Yang, Jiayi Lu, Jiepeng Zhou, Jiajun Xu, Jiangcheng Song, Boran Zhao, and Pengju Ren. "the whole is greater than the sum of its parts": A compatibility-aware multi-teacher CoT distillation framework. *arXiv preprint arXiv:2601.13992*, 2026. doi: 10.48550/arXiv.2601.13992. URL `https://arxiv.org/abs/2601.13992`.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, et al. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. doi: 10.48550/arXiv.2401.06066. URL https://arxiv.org/abs/2401.06066.

Junfeng Fang, Zhepei Hong, Mao Zheng, Mingyang Song, Gengsheng Li, Houcheng Jiang, Dan Zhang, Haiyun Guo, Xiang Wang, and Tat-Seng Chua. Rubric-based on-policy distillation. *arXiv preprint arXiv:2605.07396*, 2026. doi: 10.48550/arXiv.2605.07396. URL https://arxiv.org/abs/2605.07396.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR, 2023. URL https://proceedings.mlr.press/v202/fu23d.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-OPD: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026. doi: 10.48550/arXiv.2605.03677. URL https://arxiv.org/abs/2605.03677.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

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 *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. doi: 10.48550/arXiv.2401.04088. URL https://arxiv.org/abs/2401.04088.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. BIG-Bench extra hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26473–26501, 2025. doi: 10.18653/v1/2025.acl-long.1285. URL https://aclanthology.org/2025.acl-long.1285/.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.16668.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6265–6274. PMLR, 2021. URL https://proceedings.mlr.press/v139/lewis21a.html.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026. doi: 10.48550/arXiv.2604.13016. URL https://arxiv.org/abs/2604.13016.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572, 2023. doi: 10.52202/075280-0943. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, 2023. doi: 10.18653/v1/2023.acl-short.151. URL https://aclanthology.org/2023.acl-short.151/.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707*, 2023. doi: 10.48550/arXiv.2306.02707. URL https://arxiv.org/abs/2306.02707.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. doi: 10.48550/arXiv.1701.06538. URL https://arxiv.org/abs/1701.06538.

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. URL https://arxiv.org/abs/2010.03768.

Joe Stacey and Marek Rei. Distilling robustness into natural language inference models with domain-targeted augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2239–2258, 2024. doi: 10.18653/v1/2024.findings-acl.132. URL https://aclanthology.org/2024.findings-acl.132/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association*

for Computational Linguistics: ACL 2023, pages 13003–13051, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

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. URL https://arxiv.org/abs/2203.11171.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e4a46394ba5378b3f9a186a5b4c650d1-Abstract-Conference.html.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488, 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

A Variant and evaluation key

Auto-v4-Open is the fixed single-verifier system in the main comparisons. Auto-v4-Ensemble adds a closed-API verifier and is reported separately; the manual canonicalizer is a reference. A1 always averages the five primary-shift domain accuracies equally. External benchmark aggregates and unweighted category summaries are distinguished below.

Table 7: Variant and evaluation key for repeated results.

System	A1	Main comparison
Auto-v4-Open	58.62	MoA 56.16; MPS 56.58
Simplified Open	57.97	86.5% of Open's gain over KD
Auto-v4-Ensemble	58.68	Verifier-source sensitivity
Manual	58.67	Canonicalization reference

B Alignment, inventory stability, and stronger controls

For a domain with fixed inventory size m_d, let D_j be the supervision assigned to shared adapter j:

$$D_j = \bigcup_i \{(x_i, t) : t \in \mathcal{T}_i,\ z_d(t) = j\}. \tag{6}$$

For a common permutation π, relabeling every trace gives $D'_{\pi(j)} = D_j$. Relabeling adapter parameters, gate coordinates, and usage state by the same permutation preserves the objective: the KD terms are reindexed, coverage and usage sums are unchanged, and the pairwise cosine sum contains the same pairs. In contrast, input-dependent permutations π_i give

$$D'_j = \bigcup_i \{(x_i, t) : t \in \mathcal{T}_i,\ z_d(t) = \pi_i^{-1}(j)\}. \tag{7}$$

These sets generally mix different original strategy labels within each adapter. Every within-input group can retain its members and cosine coherence, yet the shared training targets change. This explains why local coherence and cross-input consistency require distinct tests. The argument concerns label naming for a fixed inventory; nearest-medoid assignment handles the additional case of unequal local and global cluster counts.

A local cluster number is input-specific. In contrast, the label used by an adapter is the nearest global-medoid label in the shared domain representation. The assignment cost for medoid consistency is cosine distance, $c_{ab} = 1 - \cos(\mu_a, \mu_b^{\mathrm{ref}})$; Kuhn–Munkres matches domain medoids to the reference inventory. This permutation check operates on global inventories, not on a forced one-to-one mapping of every local cluster. Singleton groups contribute their sole trace, and absent local labels contribute no KD target. Low-pass-rate groups are filtered before assigning the common trace pool to matched arms, so the comparison changes assignment rather than retention.

Table 8: Global-label and inventory diagnostics. Ranges are across domains or re-clustering seeds as specified; paired intervals refer to A1 effects.

Diagnostic	Result
Local/global label agreement	91.4% (domain range 88.7–93.6)
Cross-seed adjusted Rand index	0.883 (range 0.851–0.907)
Identity permutation after canonical medoid ordering	47/50 seed-domain cells
Aligned / unaligned A1	58.62 / 56.02
Aligned minus unaligned	2.60 [1.91,3.29] pp
Selected inventory changes	2/50 seed-domain cells
Normalized spectral-gap margin, mean / minimum	0.187 / 0.041
Fixed modal inventory A1	58.55
Runner-up inventory A1	58.03
Runner-up minus MoA	1.87 [1.17,2.57] pp
Per-seed reselection A1	58.64
Reselection minus fixed-before-student-seeds	0.02 [-0.11,0.15] pp

The two inventory changes are ALFWorld from three to four once and BBH from two to three once. Fixing the modal count or using the runner-up preserves a positive advantage over MoA and MPS. This separates the effect of consistent labels from favorable count selection.

Constrained shuffling reduces every semantic diagnostic in Table 9, confirming that the randomized control weakens strategy information while retaining the matched marginals.

Table 9: Open MPS semantic audit. The constrained swaps preserve mode count, cardinality, length, difficulty, verifier-pass marginals, and KD targets while changing trace labels.

Metric	Aligned Open	MPS
Within-cluster cosine	0.612	0.436
External-tag AMI	0.702	0.528
Tag-to-cluster macro-F1 (%)	73.6	61.8
Held-out tag recovery (%)	79.4	66.8
Blinded same-strategy AUC	0.867	0.708

Table 10: Functional comparison against coherence-preserving controls on 2,000 externally authored strategy-switch pairs. Forced-wrong penalties use the externally reassigned strategy tag.

	SOCP	ECPS	Open
Route changes (%)	27.4	30.6	51.6
Forced-wrong penalty (pp)	-1.57	-1.83	-5.03

Open's greater route response and larger wrong-strategy penalty in Table 10 link these semantic differences to functional specialization. The generic E5 baseline in Table 1 uses an unconstrained

embedding partition; ECPS additionally matches the reference coherence. Auto-v2 uses trace-graph MDL and scores 58.35 on A1. Auto-v5-Free augments trace-graph MDL with frozen student activations and scores 58.28, $+1.70$ pp over MPS. These verifier-free controls distinguish strategy-sensitive structure from generic embedding similarity.

C Auto-v4-Open external slices and strategy diversity

The main external aggregate compares the same locked Open system against MoA. MMLU-Pro has 14 categories; Law, Philosophy, and History are held out during training. LiveBench categories and months are overlapping views of the same evaluated items, so the eleven temporal slices are not additional independent categories. BBEH has 24 subtask slices and OlympiadBench eight category slices. The strategy-diversity analysis contains 68 slices: 15 original shifts, 14 MMLU-Pro categories, seven LiveBench categories, 24 BBEH subtasks, and eight OlympiadBench categories.

For category effects δ_c with n_c evaluated items, the item-weighted effect is $\sum_c n_c\delta_c / \sum_c n_c$, while the category-unweighted effect is $C^{-1}\sum_{c=1}^{C}\delta_c$. These answer different questions: the former describes a random evaluated item and the latter gives every category equal weight. For OlympiadBench they are 2.87 and 2.62 pp, respectively. Category and month summaries for LiveBench are alternative groupings of one evaluation pool.

Table 11: Exact Open external coverage and contrasts against MoA. Aggregate gaps are item-weighted; the OlympiadBench category-unweighted gap is reported separately. Positive counts describe signed point estimates; the marginal-interval convention is defined in Appendix F.

Benchmark	Positive	Gap (pp)	Slice summary (pp)
MMLU-Pro	14/14	4.91	range 2.91–7.02
LiveBench categories	7/7	3.99	knowledge recall 1.88
LiveBench months	11/11	3.99	range 3.58–4.36
BBEH	21/24	3.70	median 3.89; range $[-0.85, 6.55]$
OlympiadBench	7/8	2.87	category-unweighted gap 2.62

Table 12: Auto-v4-Open: named category/subtask contrasts. The four negative cells are retained with their signed effects.

Benchmark	Slice	Open − MoA (pp)
MMLU-Pro	Law	5.44
	Philosophy	4.58
	History	5.03
LiveBench	Knowledge recall	1.88
BBEH	Sports understanding	-0.41
	Ruin names	-0.85
	Snarks	-0.57
OlympiadBench	PhyCEE-TEXT-EN	-0.73

The negative named slices localize the external benchmarks' failure cases rather than being removed from their aggregate scores. For ALFWorld tool reconfiguration, the expanded evaluation narrows uncertainty while retaining a similar Open-minus-MoA point estimate; it remains separate from A1's combined shift.

Table 13: ALFWorld tool-reconfiguration precision check. The expanded pool extends the original construction; this shift is separate from the combined shift used in A1. Δ denotes Open minus MoA. MPS in the expanded pool is 57.20, obtained from Open 57.94 minus the reported paired gap 0.74.

Pool	Episodes	Open	MoA	Δ [95% CI]
Original	480	58.02	56.87	1.15 [-0.12,2.41]
Expanded	1,920	57.94	56.85	1.09 [0.31,1.87]

D Transfer beyond training domains

The four suites are excluded from training, discovery, and clustering. The Humanities and Social Sciences subsets contain 13 and 12 subjects, respectively; LegalBench uses a 20-task reasoning subset. Table 14 separates preservation relative to the frozen model from improvement over routed distillation. The smaller off-domain gain is consistent across the four suites. The suite-average changes relative to

Table 14: Off-domain accuracy (%). Frozen denotes the frozen Llama model and Δ is Open minus MoA; all use the same suites.

Suite	Frozen	MoA	Open	Δ
MMLU-Humanities	58.41	57.94	58.72	0.78
MMLU-Social-Sciences	68.72	68.16	69.03	0.87
ETHICS-hard	51.36	50.79	51.62	0.83
LegalBench reasoning	54.18	53.66	54.47	0.81

the frozen model are -0.86, -0.53, -0.44, and $+0.29$ pp for KD, MoA, MPS, and Open. The paired Open-minus-MoA mean is $+0.82$ pp, CI $[0.41, 1.23]$. The OOD score increases as maximum gate confidence decreases; AUROC is 0.813 for Open, 0.662 for MoA, and 0.671 for MPS. Mean entropy increases from 0.49 to 0.71 nats and minimum EMA mass is 0.11. This indicates more diffuse routing on unfamiliar domains while retaining active modes.

E Operational evidence matrix

These are non-adversarial robustness evaluations of answer correctness and routing behavior. The evaluations use fixed scorers, training-side partition construction, and blinded external tag assignment where tags are used. The matrix connects each intervention to its comparator and observed operating condition.

Table 15: Evaluated operating conditions and evidence. Inference-token overhead is relative to Standard KD; discovery cost is separate.

Setting	Intervention	Observed effect or cost
Strategy-switch routing	Answer-preserving perturbation; SOCP/ECPS	Open reroutes 51.6%; forced-wrong penalty -5.03 pp (Table 10).
ALFWorld correctness	Tool reconfiguration; MoA/MPS	Expanded Open gap 1.09/0.74 pp; smaller gain at ESF 1.41 (Table 13).
Off-domain transfer	Unseen suites; frozen model/MoA	Mean +0.29/+0.82 pp; entropy 0.71 nats (Table 14).
Discovery source	Verifier-budget controls and verifier-free discovery	Open exceeds equal-budget MoA/MPS by 2.18/1.81 pp (Table 2).
Accuracy/token cost	Simplified/full Open	A1 57.97/58.62; overhead 5.8%/11.8%; full gate 0.31 ms (Table 6).

F Evaluation definitions and uncertainty

The original grid contains three shifts per domain, with the five A1 primary rows marked in Table 16. Let $a_{d,s}$ be the primary-shift accuracy for domain d and student seed s, and S the number of seeds in the evaluated stack. The reported point estimate is

$$\mathrm{A1} = \frac{1}{5}\sum_{d=1}^{5}\left(\frac{1}{S}\sum_{s=1}^{S} a_{d,s}\right). \tag{8}$$

The five domain weights remain equal regardless of their evaluation-pool sizes. Every accuracy comparison uses the same evaluator within a domain. BCa intervals are paired across systems and based on 10,000 example resamples; they describe test-pool uncertainty rather than a simultaneous claim across all shifts. The expanded ALFWorld tool evaluation is reported separately in Table 13. Correctness win rates use the same accuracy estimand, as derived below.

Table 16: Original fifteen-shift accuracies (%). Open is Auto-v4-Open; primary rows define A1.

ID	Shift	KD	MoA	Open
M1	Difficulty (primary)	58.24	60.93	63.69
M2	Prompt style	60.81	62.44	63.92
M3	Language (EN→CN)	54.71	55.87	56.88
H1	Py→TS (primary)	61.68	63.86	65.77
H2	Complexity escalation	63.22	64.83	66.41
H3	API style	59.44	61.26	63.31
A1	Layout variation	51.82	54.17	57.12
A2	Tool reconfiguration	55.63	56.87	58.02
A3	Combined (primary)	54.17	56.58	58.87
B1	Task family (primary)	58.06	60.11	62.79
B2	Long chain (≥ 8 steps)	62.41	63.08	66.01
B3	Few-shot format flip	60.80	61.20	64.04
G1	Subject family (primary)	36.83	39.32	41.98
G2	Distractor perturbation	38.14	40.25	42.98
G3	Expert-only hold-out	34.58	36.91	39.48

The row prefixes identify MATH-500 (M), HumanEval+ (H), ALFWorld (A), BBH-Hard (B), and GPQA-Diamond (G). Table 17 uses the same row IDs to retain both paired contrasts and their uncertainty. Open has a positive point difference on all fifteen shifts; the original tool-reconfiguration interval against MoA crosses zero.

Table 17: Original fifteen-shift contrasts (pp) and tie-split correctness win rates (WR, %). Brackets are paired 95% BCa intervals.

ID	Δ_{KD} [95% CI]	Δ_{MoA} [95% CI]	WR_{MoA}
M1	$+5.45$ $[+4.73, +6.17]$	$+2.76$ $[+1.97, +3.55]$	51.38%
M2	$+3.11$ $[+2.44, +3.78]$	$+1.48$ $[+0.71, +2.25]$	50.74%
M3	$+2.17$ $[+1.42, +2.92]$	$+1.01$ $[+0.18, +1.84]$	50.51%
H1	$+4.09$ $[+3.37, +4.81]$	$+1.91$ $[+1.12, +2.70]$	50.95%
H2	$+3.19$ $[+2.47, +3.91]$	$+1.58$ $[+0.78, +2.38]$	50.79%
H3	$+3.87$ $[+3.12, +4.62]$	$+2.05$ $[+1.23, +2.87]$	51.02%
A1	$+5.30$ $[+4.56, +6.04]$	$+2.95$ $[+2.11, +3.79]$	51.48%
A2	$+2.39$ $[+1.66, +3.12]$	$+1.15$ $[-0.12, +2.41]$	50.58%
A3	$+4.70$ $[+3.97, +5.43]$	$+2.29$ $[+1.43, +3.15]$	51.15%
B1	$+4.73$ $[+4.01, +5.45]$	$+2.68$ $[+1.88, +3.48]$	51.34%
B2	$+3.60$ $[+2.87, +4.33]$	$+2.93$ $[+2.15, +3.71]$	51.47%
B3	$+3.24$ $[+2.53, +3.95]$	$+2.84$ $[+2.04, +3.64]$	51.42%
G1	$+5.15$ $[+4.34, +5.96]$	$+2.66$ $[+1.78, +3.54]$	51.33%
G2	$+4.84$ $[+4.03, +5.65]$	$+2.73$ $[+1.83, +3.63]$	51.37%
G3	$+4.90$ $[+4.07, +5.73]$	$+2.57$ $[+1.69, +3.45]$	51.28%

G Tie-split correctness comparison

For binary correctness indicators Y_A, Y_B, a win has $Y_A = 1, Y_B = 0$ and a tie includes both-correct and both-incorrect outcomes. Splitting ties equally gives

$$\text{WR}(A, B) = \Pr(Y_A > Y_B) + \tfrac{1}{2}\Pr(Y_A = Y_B) = \tfrac{1}{2} + \tfrac{1}{2}(\text{Acc}(A) - \text{Acc}(B)).$$

Consequently, Open's equal-domain A1 tie-split win rate is 51.23% against MoA and 52.41% against KD. This is an alternative presentation of the accuracy difference, not additional evidence of example-level dominance. Table 17 reports the corresponding per-shift values; Figure 2 compares the two control contrasts on the primary shifts.

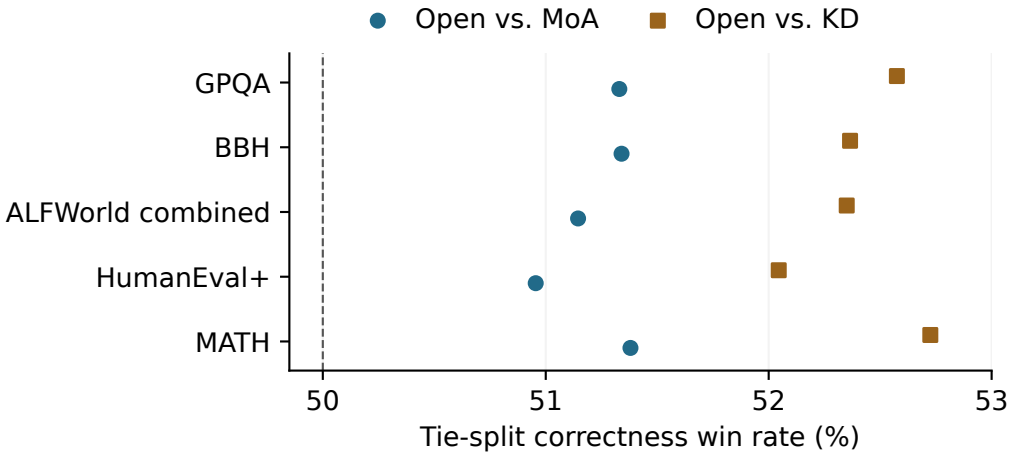


Figure 2: Tie-split correctness win rates derived from the five primary-shift accuracy differences.

H Annotation and functional strategy checks

In the annotation tables, EC denotes the generic embedding control, and Open is Auto-v4-Open. The blinded pairwise audit samples 320 pairs per system and domain: 160 within-cluster and 160 cross-cluster pairs. Three raters judge each pair across five domains and three systems, giving 14,400 judgments. Raters see whitespace-normalized traces with identifiers removed, in randomized order, without cluster, seed, model, or system labels. Contractors receive $21.50/hour, above the contracting jurisdiction's minimum wage; the task collects no personally identifying information beyond annotations and permits opting out.

Table 18: Blinded pairwise audit. Same and cross-cluster false-match rate (FMR) are percentages; AUC measures ranking and κ is Fleiss agreement.

Domain	System	AUC	Same	FMR	κ
MATH-500	Open	0.894	82.1	17.4	0.69
	MoA	0.737	66.5	35.6	0.48
	EC	0.772	68.8	32.0	0.52
HumanEval+	Open	0.868	78.6	21.5	0.65
	MoA	0.723	65.3	36.4	0.46
	EC	0.756	67.9	33.0	0.51
ALFWorld	Open	0.837	75.3	24.1	0.61
	MoA	0.684	62.0	38.6	0.43
	EC	0.728	65.1	35.2	0.48
BBH-Hard	Open	0.866	78.4	22.1	0.64
	MoA	0.702	63.5	37.8	0.45
	EC	0.741	67.0	33.9	0.49
GPQA-Diamond	Open	0.870	78.8	22.9	0.63
	MoA	0.712	63.2	37.5	0.44
	EC	0.748	67.5	33.4	0.50
Avg.	Open	0.867	78.6	21.6	0.644
	MoA	0.712	64.1	37.2	0.452
	EC	0.749	67.3	33.5	0.500

Across all five domains, Open has higher same-strategy AUC and lower cross-cluster false-match rates than both controls (Table 18).

H.1 External strategy labels

Four external domain experts define strategy tags from 80 held-out traces per domain before cluster scoring. Three independent raters label 500 successful traces per system and domain; a blinded fourth adjudicator resolves disagreement. The rubric distinguishes algebraic/geometric/analytic/casework/substitution reasoning in mathematics, imperative/functional/recursive/dynamic-programming code, explore-first/shortcut/hybrid plans, deductive-chain/case-split reasoning in BBH, and first-principles/elimination/analogical reasoning in GPQA. MMLU-Pro adds definitional, multi-hop, rule-application, and elimination tags. The audit matches tags to clusters with Kuhn–Munkres, separately from the training-time global-medoid alignment.

Table 19: External-tag audit. F1 is matched macro-F1 (%), AMI is adjusted mutual information, and κ is pre-adjudication Fleiss agreement.

Domain	System	AMI	F1	Purity	κ
MATH-500	Open	0.719	75.1	0.808	0.748
	MoA	0.446	53.8	0.579	0.491
	EC	0.526	61.0	0.634	0.534
HumanEval+	Open	0.697	73.1	0.786	0.736
	MoA	0.431	51.2	0.559	0.482
	EC	0.510	58.6	0.614	0.521
ALFWorld	Open	0.662	70.7	0.766	0.683
	MoA	0.418	48.7	0.534	0.441
	EC	0.496	56.0	0.590	0.491
BBH-Hard	Open	0.707	73.8	0.794	0.723
	MoA	0.443	52.2	0.568	0.472
	EC	0.517	59.5	0.622	0.515
GPQA-Diamond	Open	0.720	74.9	0.806	0.731
	MoA	0.452	53.7	0.576	0.462
	EC	0.525	60.8	0.630	0.521
MMLU-Pro	Open	0.710	73.4	0.802	0.708
	MoA	0.442	52.8	0.572	0.455
	EC	0.524	60.1	0.625	0.502

The external tags yield higher AMI, matched F1, and purity for Open in each domain (Table 19), supporting the strategy interpretation independently of the learned partition labels. A trace-only RoBERTa-base classifier uses a random head, three fine-tuning epochs, learning rate 2×10^{-5}, batch 32, and maximum sequence length 512. Five repeated 70/30 train/test splits yield held-out tag recovery of 79.4% for Open, 62.5% for MoA, and 65.1% for the embedding control.

In intervention tables, Wrong denotes forced-wrong routing and Oracle denotes the best-aligned route. Routing interventions use distinct conditioning populations. The original five-domain intervention in Table 20 forces a wrong paper-derived mode and gives Open a mean -3.92 pp penalty. The six-domain externally tagged intervention in Table 21 gives -4.47 pp. The answer-preserving strategy-switch evaluation in Table 10 instead uses externally reassigned tags on 2,000 paired perturbations and gives -5.03 pp. These effects have different populations and should be compared within their respective tables.

Strategy-switch examples preserve the answer while changing the tagged solution: a problem asking for the larger positive root of $x^2 - 6x + 8 = 0$ becomes a larger-positive-intercept problem for $y = x^2 - 6x + 8$, with answer 4; iterative code is recast as a recursive specification under the same unit tests; an ALFWorld retrieval target is moved into a container while the success condition is retained. BBH perturbations switch between deductive chains and case splits. An external adjudicator verifies answer preservation and the changed tag before inclusion.

Table 20: Original five-domain intervention: learned-to-Wrong and learned-to-Oracle accuracy changes (pp). A1 averages the five domains equally.

	Open		MoA		EC	
Domain	Wrong	Oracle	Wrong	Oracle	Wrong	Oracle
MATH-500	-3.63	$+1.48$	-1.03	$+3.51$	-1.72	$+2.94$
HumanEval+	-3.28	$+1.36$	-0.87	$+3.28$	-1.54	$+2.77$
ALFWorld	-4.91	$+1.81$	-1.39	$+4.05$	-2.31	$+3.42$
BBH-Hard	-3.86	$+1.68$	-1.18	$+3.78$	-1.92	$+3.18$
GPQA-Diamond	-3.92	$+1.52$	-1.23	$+3.48$	-1.86	$+3.09$
A1 avg.	-3.92	$+1.57$	-1.14	$+3.62$	-1.87	$+3.08$

Table 21: Externally tagged six-domain intervention (pp). Forced-wrong selects the adapter least aligned with the external tag; Oracle selects the best-aligned adapter.

	Open		MoA		EC	
Domain	Wrong	Oracle	Wrong	Oracle	Wrong	Oracle
MATH-500	-4.11	$+1.76$	-0.97	$+3.80$	-1.54	$+3.12$
HumanEval+	-3.93	$+1.58$	-0.86	$+3.58$	-1.39	$+2.95$
ALFWorld	-5.48	$+2.15$	-1.35	$+4.37$	-2.14	$+3.63$
BBH-Hard	-4.36	$+1.94$	-1.12	$+4.08$	-1.74	$+3.39$
GPQA-Diamond	-4.42	$+1.79$	-1.18	$+3.74$	-1.69	$+3.27$
MMLU-Pro	-4.50	$+2.12$	-1.04	$+4.07$	-1.68	$+3.06$
Avg.	-4.47	$+1.89$	-1.09	$+3.94$	-1.70	$+3.24$

I Routed-student and discovery ablations

The detailed ablations below complement the A1 summary in Table 3, separating route specialization from discovery robustness.

Table 22: MATH difficulty-shift ablations. EMC denotes effective mode count; the KD reference is 58.24%.

Variant	ID	OOD	EMC	Δ_{KD}
Manual (full)	73.6	63.74	1.63	$+5.50$
Uniform routing	73.4	59.64	3.00	$+1.40$
$-$ Diversity loss	73.7	60.08	1.08	$+1.84$
$-$ Coverage loss	73.2	61.93	1.41	$+3.69$
$-$ Usage loss	73.5	61.31	1.27	$+3.07$
Single teacher trace	73.1	59.14	1.54	$+0.90$
$m = 2$ modes	73.5	61.64	1.48	$+3.40$
$m = 5$ modes	73.2	62.58	1.72	$+4.34$

Here Auto-v2 is MDL trace-graph discovery, Auto-v1 uses AST plus deduplication, Noise denotes canonicalizer perturbation, and Surface only uses surface-form features. Table 23 retains domain accuracies; Table 24 reports their equal-domain summary.

All discovery perturbations reduce the manual reference's A1. The 20% noise setting is the weakest perturbed canonicalizer, while the surface-only partition also remains below the strategy-aware variants. The domain rows and summary describe the same runs.

Auto-v3 uses the trace-generating Qwen teacher as its pairwise verifier. In Table 25, GPT-OSS, Claude, and Mixtral abbreviate GPT-OSS-120B, Claude-3.5-Sonnet, and Mixtral-8x22B, respectively. Claude-containing configurations use a closed API; the GPT-OSS/Mixtral pair uses open-weight models. Open changes the verifier to GPT-OSS-120B while retaining the six-layer 384-dimensional

Table 23: Discovery robustness: primary-shift accuracy (%) by domain.

Variant	MATH	HE+	ALFW.	BBH	GPQA
Manual	63.74	65.81	58.92	62.83	42.05
Auto-v2	63.38	65.52	58.57	62.47	41.82
Auto-v1	62.78	65.00	58.17	62.26	41.01
Noise 10%	62.62	64.87	58.13	62.02	40.86
Noise 20%	61.39	63.78	57.06	60.59	39.44
Surface only	61.43	63.72	57.48	60.92	39.76
MoA	60.93	63.86	56.58	60.11	39.32
Random cluster	59.55	62.62	55.08	58.67	37.64
KD	58.24	61.68	54.17	58.06	36.83

Table 24: Discovery robustness: A1 and differences from the manual reference (pp).

Variant	A1	Δ_{Manual}
Manual	58.67	—
Auto-v2	58.35	-0.32
Auto-v1	57.84	-0.83
Noise 10%	57.70	-0.97
Noise 20%	56.45	-2.22
Surface only	56.66	-2.01
MoA	56.16	-2.51
Random cluster	54.71	-3.96
KD	53.80	-4.87

encoder, InfoNCE temperature 0.07, batch 256, 8,000 encoder steps, and spectral-gap selection. Auto-v4-Ensemble combines GPT-OSS with Claude-3.5-Sonnet-20240620, discarding disagreements between the two judgments. All variants use global labels to address adapters. Auto-v3 reaches 58.39 on A1; Ensemble reaches 58.68, a 0.29 pp difference. The manual reference is 58.67 and Open 58.62. The verifier configurations have a narrow A1 spread, but this aggregate stability does not eliminate

Table 25: Verifier-source sensitivity on A1. Open is fixed as the main operating point; the ensemble and alternative-verifier rows use the same downstream routed student.

Verifier	A1	95% CI	Δ_{Manual}	Δ_{MoA}
GPT-OSS	58.62	$[57.91, 59.31]$	-0.05	$+2.46$
GPT-OSS + Claude	58.68	$[57.97, 59.38]$	$+0.01$	$+2.52$
Claude	58.59	$[57.88, 59.28]$	-0.08	$+2.43$
GPT-OSS + Mixtral	58.56	$[57.86, 59.24]$	-0.11	$+2.40$
Claude + Mixtral	58.63	$[57.93, 59.32]$	-0.04	$+2.47$
Spread	0.12 pp (range 58.56–58.68)			

ambiguous strategy assignments. The held-out-seed examples in Table 26 identify boundary cases in which neighboring strategy families are confused.

Table 26: Ensemble sensitivity: held-out-seed qualitative error audit, 250 traces per domain. Ambiguous strategy boundaries remain the dominant errors.

Domain	Boundary error	Rate	Example
MATH-500	coordinate arguments near algebraic/geometric line	2.9%	parametric curve cases on coordinate grids
HumanEval+	functional split on accumulator-recursion	2.1%	tail-call recursion with explicit accumulator
ALFWorld	explore-first vs. hybrid on tool-rich layouts	3.3%	$\geq$ 6-tool rooms with short lookahead
BBH-Hard	case-split absorbed into deductive on short chains	2.8%	$\leq$ 3-step implications with one branch
GPQA-Diamond	first-principles vs. analogical on unit-heavy prompts	2.5%	dimension-only analogies read as first-principles
MMLU-Pro	multi-hop vs. rule-application on long prompts	3.0%	long rule-chain items read as multi-hop

J Teacher/student and partition-source transfer

Stack B uses GPT-OSS-120B to Mistral-7B-v0.3; stack C uses Qwen-3-72B to Phi-3.5-mini-3.8B. They change teacher and student within the adapter family, while stacks D/E in Table 4 change the routing mechanism. The B/C tables use eight seeds per stack. All ten Open-minus-MoA domain effects are positive, with a minimum of 1.89 pp on stack-C ALFWorld.

Table 27: Stack B: primary-shift accuracy (%), eight seeds. TR is Trace-Reweight KD; Manual and Open use the corresponding strategy partitions.

Domain	KD	MoA	TR	Manual	Open
MATH-500	54.72	56.98	55.94	59.18	59.14
HumanEval+	58.31	60.12	59.02	62.30	62.25
ALFWorld	50.48	52.36	51.28	54.55	54.50
BBH-Hard	54.13	56.26	55.17	58.48	58.44
GPQA-Diamond	34.12	36.57	35.48	38.74	38.70
Average	50.35	52.46	51.38	54.65	54.61

The domain-specific gains in Table 28 compare both strategy partitions with MoA under this same teacher/student stack. Their close values separate the partition benefit from the choice between manual and automatic induction.

Table 28: Stack B: primary-shift differences from MoA (pp).

Domain	Manual − MoA	Open − MoA
MATH-500	+2.20	+2.16
HumanEval+	+2.18	+2.13
ALFWorld	+2.19	+2.14
BBH-Hard	+2.22	+2.18
GPQA-Diamond	+2.17	+2.13
Average	+2.19	+2.15

Stack C retains positive assignment effects in every domain (Table 30); ALFWorld has the smallest Open gap, 1.89 pp. The cross-stack summary and paired intervals (Tables 31–32) show consistent gains over MoA with closely matched manual and Open results.

In the domain tables, MATH, HE+, ALFW., BBH, and GPQA abbreviate MATH-500, HumanEval+, ALFWorld, BBH-Hard, and GPQA-Diamond. Their primary shifts are difficulty, Python-to-TypeScript, combined, task family, and subject family, respectively.

Table 29: Stack C: primary-shift accuracy (%), eight seeds. TR is Trace-Reweight KD; Manual and Open use the corresponding strategy partitions.

Domain	KD	MoA	TR	Manual	Open
MATH-500	52.36	54.21	53.19	56.18	56.14
HumanEval+	56.73	58.41	57.44	60.44	60.40
ALFWorld	48.94	50.88	49.82	52.81	52.77
BBH-Hard	52.19	54.06	53.02	56.15	56.11
GPQA-Diamond	32.84	35.03	33.97	37.23	37.19
Average	48.61	50.52	49.49	52.56	52.52

Table 30: Stack C: primary-shift differences from MoA (pp).

Domain	Manual − MoA	Open − MoA
MATH-500	+1.97	+1.93
HumanEval+	+2.03	+1.99
ALFWorld	+1.93	+1.89
BBH-Hard	+2.09	+2.05
GPQA-Diamond	+2.20	+2.16
Average	+2.04	+2.00

Table 31: Open-minus-MoA differences (pp) on each primary shift. B/C are teacher/student transfers within the adapter family.

Stack	MATH	HE+	ALFW.	BBH	GPQA
A (Open reference)	+2.76	+1.91	+2.29	+2.68	+2.66
B	+2.16	+2.13	+2.14	+2.18	+2.13
C	+1.93	+1.99	+1.89	+2.05	+2.16

Table 32: Stack-level primary-shift differences (pp), with paired 95% BCa intervals; eight seeds per stack.

Comparison	Stack	Mean	95% CI
Open − MoA	B	+2.15	[+1.46, +2.82]
Open − MoA	C	+2.00	[+1.30, +2.70]
Manual − MoA	B	+2.19	[+1.50, +2.87]
Manual − MoA	C	+2.04	[+1.34, +2.74]
Manual − Open	B	+0.04	[−0.25, +0.34]
Manual − Open	C	+0.04	[−0.26, +0.33]

The cross-partitioner comparison holds the student fixed and changes only the source of global assignments. Mixtral-EVC uses Mixtral-8x22B-Instruct-v0.1 pair judgments with the same contrastive encoder schedule. External-Tag-Derived assigns strategies using the externally authored rubric and a blinded rater's tag-merging rule. Their A1 spread is 0.45 pp, compared with at least 2.01 pp advantage over MoA. The maximum within-domain spread is 0.60 pp.

External tags denotes External-Tag-Derived, and Open EVC is Auto-v4-Open EVC. MoA is non-semantic routing; MPS is the non-semantic structure-matched shuffle. The three semantic sources improve A1 over MoA, with the paired differences shown below.

Table 33: Alternative partition sources at 4.80B teacher tokens: accuracy (%). A1 and MMLU-Pro use ten seeds; OlympiadBench uses five.

Source	A1 [95% CI]	MMLU-Pro	Olympiad
MoA	56.16 [55.43, 56.87]	44.96	29.64
MPS	56.58 [55.88, 57.27]	45.82	29.94
External tags	58.17 [57.45, 58.87]	49.35	32.04
Mixtral-EVC	58.31 [57.59, 59.00]	49.52	32.18
Open EVC	58.62 [57.91, 59.31]	49.87	32.51

Table 34: Alternative partition-source A1 differences from MoA (pp); brackets are paired 95% BCa intervals.

Source	Source − MoA
MoA	0
MPS	+0.42
External tags	+2.01 [+1.31, +2.71]
Mixtral-EVC	+2.15 [+1.45, +2.85]
Open EVC	+2.46 [+1.76, +3.16]

Table 35: Per-domain primary-shift accuracy under alternative partition sources. MPS and Open are the same systems as in Table 1.

Partition source	MATH	HE+	ALFW.	BBH	GPQA
MoA	60.93	63.86	56.58	60.11	39.32
MPS	61.12	64.15	57.08	60.40	40.15
External tags	63.27	65.40	58.45	62.35	41.38
Mixtral-EVC	63.40	65.51	58.57	62.49	41.58
Open EVC	63.69	65.77	58.87	62.79	41.98

K Budget and inference-cost comparisons

Teacher generation and partition discovery are one-time operations. The reported teacher-generation cost is 720.4 A100-hours, and Open discovery adds 84.2 A100-hours. A100-hours denotes GPU time, not elapsed wall-clock time on the eight-GPU node. These costs are separate from the matched KD trace budget and from recurring inference-token overhead.

On MATH difficulty, the manual-reference fairness measurement uses 847 inference tokens/example for KD, 925 for MoA, and 947 for the manual routed student, corresponding to 9.2% and 11.8% increases. The 0.31 ms gate measurement uses batch 16 on an A100. Adapter ranks and counts are held fixed across matched routed arms; the nominal adapter-only capacity is m_d times the 4.2M parameters of a rank-16 adapter, with the same gate added in every routed arm. BBH's two-mode inventory is capacity-matched within that domain.

Budget-matched self-consistency samples at temperature 0.7 and top-$p = 0.95$ under a shared total decode cap per example. The cap, rather than the nominal number of samples, defines its inference cost. Incomplete samples are discarded; matching verified answers allow early stopping, and voting ties use the highest-probability completed sample. In the following cost tables, r16/r48/r96 denote single-adapter KD ranks, BM-SC denotes budget-matched self-consistency, and Open uses automatically selected mode counts with rank-16 adapters. Dominance concerns A1 accuracy and token overhead only. Parameter counts are nominal adapter capacity; the shared gate is additional.

The capacity and in-distribution scores below describe the same operating points. Open uses 12.6M adapter parameters in three-mode domains and 8.4M in two-mode BBH. The inference comparison therefore distinguishes recurring token overhead from stored adapter capacity.

Open and simplified Open offer A1 58.62/57.97 at 11.8%/5.8% overhead. Both lie on the accuracy/token-cost frontier of the reported operating points; the full system buys higher accu-

Table 36: Self-consistency decoding on MATH difficulty. Effective sample counts reflect truncation and early stopping under a shared token cap.

Variant	Tok./ex.	Cap B_k	Samples	Overhead	MATH
KD	847	—	1.0	0.0%	58.24
BM-SC-8	1,312	1,312	6.8	+55%	59.83
BM-SC-32	2,541	2,541	14.2	+200%	60.37
BM-SC-64	4,281	4,281	21.6	+405%	61.74

Table 37: A1 accuracy versus inference-token overhead. Rank-48/rank-96 KD remain nondominated alternatives.

Variant	Overhead	A1 [95% CI]	Δ_{KD}	Dominated
KD r16	+0.0%	53.80 [53.14, 54.45]	—	—
KD r48	+0.5%	55.07 [54.40, 55.73]	+1.27	no
KD r96	+0.9%	55.85 [55.16, 56.54]	+2.05	no
BM-SC-8	+55%	55.50 [54.84, 56.16]	+1.70	yes
BM-SC-32	+200%	55.35 [54.68, 56.02]	+1.55	yes
BM-SC-64	+405%	56.68 [55.99, 57.36]	+2.88	yes
Open	+11.8%	58.62 [57.91, 59.31]	+4.82	—

Table 38: Nominal adapter parameters and in-distribution accuracy for the cost comparison.

Variant	Parameters	ID (%)
KD r16	4.2M	64.82
KD r48	12.6M	65.54
KD r96	25.2M	65.70
BM-SC-8	4.2M	65.74
BM-SC-32	4.2M	65.88
BM-SC-64	4.2M	65.99
Open	8.4–12.6M	65.82

racy with more tokens. BM-SC-64 reaches 56.68 at 405% overhead. For the separate Ensemble sensitivity, the simplified/full scores are 57.98/58.68. These variants are kept distinct from Open. The additional controls in Table 39 separate multiple teacher samples, unsupervised mixture capacity, and surface-feature clustering from strategy-consistent assignment. Each remains below MoA on the shared A1 pool, whereas Open improves over it.

Table 39: Additional controls on the same A1 pool and ten seeds. Length/Format-Cluster uses trace lengths and surface-format features; Mixture Student lacks strategy-partition supervision.

Method	A1 [95% CI]	Δ_{MoA}
Multi-sample KD	54.91 [54.25, 55.56]	−1.25
Mixture Student	54.38 [53.74, 55.04]	−1.78
Length/Format-Cluster	54.78 [54.14, 55.44]	−1.38
MoA	56.16 [55.43, 56.87]	0
Open	58.62 [57.91, 59.31]	+2.46

NeurIPS Paper Checklist

1. Claims

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims in the abstract are supported by Section 3.

2. Limitations

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses strategy-diversity boundaries, global alignment, and discovery/inference costs.

3. Theory assumptions and proofs

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: No formal theorems.

4. Experimental result reproducibility

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The paper specifies the assignment mechanism, objective, and evaluation metrics, but does not report all loss coefficients, the EMA coefficient, or the hard-negative loss definition.

5. Open access to data and code

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No public code or data accompanies this submission. The experimental protocol, training sequence, and evaluation definitions are documented in Sections 2–3 and the technical appendices.

6. Experimental setting/details

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Section 3 describes the principal setup; some loss coefficients and hard-negative training details are not specified.

7. Experiment statistical significance

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 3 reports seed counts and paired BCa intervals. Table 16 preserves the original 14/15 significant shift comparisons; Table 13 separately reports the expanded ALFWorld interval. Slice intervals are marginal.

8. Experiments compute resources

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 3 specifies the eight-A100 hardware and software stack. Appendix K separates teacher generation, discovery, and inference costs.

9. Code of ethics

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Conforms to NeurIPS Code of Ethics.

10. Broader impacts

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper studies a general distillation method and does not include a dedicated analysis of societal impacts.

11. Safeguards

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release model weights or datasets.

12. Licenses for existing assets

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites the model and benchmark sources but does not provide a complete asset-by-asset license and terms-of-use inventory.

13. New assets

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new code, model, or data artifacts are released with this paper.

14. Crowdsourcing and research with human subjects

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Appendix H specifies blinded judgments and compensation; the complete participant instruction text is not included.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Annotation is limited to labeling machine-generated reasoning traces; no personally identifying information is collected beyond annotations. The task follows the institution's low-risk exempt-equivalent pathway for non-identifiable annotation of machine-generated text.

16. Declaration of LLM usage

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core methodology uses LLMs as the trace-generating teacher and as the external pairwise same-strategy verifier, and these components are described in Sections 2 and 3.