ACTION-CATALOG SHIFT REVEALS STABLE LATENT-TRACE HALLUCINATION ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination routing requires both an accurate risk score and a remedy matched to each failure, yet that pairing can become coupled to the action vocabulary used during system design. We introduce *catalog-conditioned generalization*, a controlled stress test that substitutes benchmark- or operator-supplied actions while fixing prompts, graders, utility, access, training budget, and comparator families. *TraceDispatch* (TD) addresses this shift with validation-selected multi-depth anchors, learned 64-dimensional compression, and an order-sensitive joint risk/action head executed within one generator pass. On public unseen-family evaluations spanning HaluEval-v1/Gemma-2-27B, BFCL-v3/Mixtral-8x22B, and LiveCodeBench-v3/DeepSeek-R1-32B, TD leads OO-Router by $+4.2$–$+4.4$ AUROC and $+6.2$–$+6.6$ corrected-output points. Benchmark-aligned substitution retains 94.9–96.3% of AUROC and 93.9–95.7% of intervention utility; an expanded controller frontier further places TD above output-only, matched-sketch, and richer-raw-state alternatives in every aligned catalog. TD remains top-ranked across five independently supplied catalogs spanning $K{=}3$–6; on blind human gold, it reaches 81.0 AUROC and 61.4 intervention utility at $38.0\,\mathrm{ms}$ on Llama-3.1-70B, while a human-only reconstruction of the upstream pipeline changes AUROC by just -0.9. The results establish catalog-conditioned generalization as a discriminating controller-selection criterion and show that budgeted ordered traces preserve actionable signal across changes in catalog source, model family, supervision, and serving stack.

1 INTRODUCTION

A risk estimate becomes useful only when it selects an effective response. Sending a fabricated citation to tool re-execution wastes latency, while sending a malformed call to retrieval leaves the executable error untouched. Hallucination routing therefore couples two decisions: whether to intervene and which remediation to dispatch. Scalar uncertainty estimators (Kuhn et al., 2023; Kadavath et al., 2022), binary truthfulness probes (Burns et al., 2023; Marks & Tegmark, 2024), and retrieval, citation-generation, and learned tool-use systems (Lewis et al., 2020; Gao et al., 2023; Schick et al., 2023) each resolve part of this control problem; multi-action routing must also recover the intervention semantics.

Those semantics are commonly supplied by the router designer. This creates hidden freedom to choose category boundaries that align with the training labels, representation, and downstream loop. Merging adjacent failures can erase costly confusions, while splitting convenient cases can reward a feature extractor already tuned to them. Ordinary held-out examples preserve this coupling because train and test still speak the same authored intervention language. Replacing that language with a benchmark's documented remedies directly tests whether the controller design generalizes to an action space supplied outside its construction pipeline.

We call this axis *catalog-conditioned generalization*. The paired test holds prompts, public grader, utility, train/validation/calibration partitions, example budget, threshold rule, and comparator families fixed, then substitutes the catalog source and action set. Catalog retention measures the score preserved after substitution; rank stability asks whether the controller ordering survives. Unlike label permutation or merge/split robustness, the actions acquire externally specified operational meanings.

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.

Unlike domain or model-family transfer, the request distribution and backbone can remain fixed. This controlled separation makes action-space provenance measurable rather than implicit.

TraceDispatch is designed for this regime. TD selects informative layer/token anchors on held-out validation data, maps them into a learned 64-dimensional trace, preserves early-to-late depth order with a BiGRU, and jointly predicts risk and intervention. Sixteen anchors add $4.9\,\mathrm{ms}$ to a $33.1\,\mathrm{ms}$ Llama-3.1-70B generator pass. The combination gives the controller label-aware compression, ordered multi-depth evidence, and downstream correction supervision within an explicit serving envelope.

The controller frontier reveals that catalog-conditioned generalization depends on representation design as well as input access. TD gains $+4.2$–$+4.4$ AUROC over OO-Router on three public unseen-family cells, retains approximately 95% of AUROC and intervention utility after all three vocabularies become benchmark-aligned, and remains ahead of Perceiver-Raw despite that comparator's richer pre-sketch states. Five supplied catalogs span $K{=}3$–6, dense and MoE families, and factuality, retrieval, code, and tool workloads. A disjoint $8{,}000$-example human-gold test reaches 81.0 AUROC and 61.4 intervention utility, and rebuilding anchor search, head selection, training, and calibration from human labels changes AUROC by -0.9.

Contributions We contribute a controlled catalog-shift protocol with paired retention, rank stability, request-level correction, uncertainty, and access-matched controller comparisons. We introduce TD, an efficient ordered latent-trace router whose validation-selected compression and joint risk/action objective preserve quality across catalog sources, model families, label regimes, and three serving stacks. Empirically, the cross-method frontier shows that catalog stability separates controller designs even when richer hidden-state access and substantially larger heads are available, establishing action-space provenance as a practical criterion for routing research and system selection.

2 Problem Formulation

Setting We consider a pre-trained autoregressive LLM $\mathcal{M}$ whose transformer has L layers and hidden dimension d. For prompt $x = (x_1, \ldots, x_n)$ followed by continuation $y = (y_1, \ldots, y_m)$, the hidden states emitted by $\mathcal{M}$ are $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$, indexed by layer $\ell \in [L]$ and position $t \in [n+m]$. At deployment, these states are captured while the same autoregressive decoding pass emits y; TD performs no post-generation teacher-forced replay of (x, y). We aim to detect hallucination in y and, when present, choose among $K{=}4$ action-linked *routing categories* to *route* the intervention associated with that choice:

- Unsupported factuality $(c_1) \to \iota_1$: retrieve supporting evidence and verify.
- Retrieval omission $(c_2) \to \iota_2$: re-decode with grounding-constrained sampling over the retrieved passages.
- Citation fabrication $(c_3) \to \iota_3$: verify cited references and replace fabricated citations with verified sources.
- Tool-grounding failure $(c_4) \to \iota_4$: use validated arguments to re-execute the tool call, then reconcile its returned value.

The categories form an *operational partition*: each label identifies a distinct mitigation policy and its downstream success criterion. Before final evaluation, we fixed this four-way catalog at the coarsest partition that retains every implemented intervention loop; merged, split, and independently elicited alternatives are evaluated in Appendix A10. At deployment, the action head is gated by a calibrated risk threshold: an answer below it remains unchanged, whereas a score above it triggers the intervention selected by the predicted category. In subsequent evaluation, macro-F1 serves only to diagnose consistency in predicting this operational partition; the more important criterion is intervention utility, which tests whether the dispatched action still corrects the output.

Catalog-origin shift Let $\mathcal{C}^{\mathrm{auth}} = \{(c_k, \iota_k)\}_{k=1}^{K}$ denote the authored catalog and $\mathcal{C}_b^{\mathrm{nat}}$ the vocabulary documented by benchmark b. For method m and metric q, catalog retention is

$$\rho_{m,b}^{q} = \frac{q(m; \mathcal{C}_b^{\mathrm{nat}})}{q(m; \mathcal{C}^{\mathrm{auth}})}, \qquad \Delta_{m,m',b}^{q} = q(m; \mathcal{C}_b^{\mathrm{nat}}) - q(m'; \mathcal{C}_b^{\mathrm{nat}}). \tag{1}$$

The paired protocol fixes prompts, grader, utility function, access budget, training examples, split identities, and comparator family while substituting $\mathcal{C}_b^{\mathrm{nat}}$ for $\mathcal{C}^{\mathrm{auth}}$. Rank stability holds when the sign of each predeclared $\Delta^q_{m,m',b}$ is preserved; Section 4.2 reports retention and rank stability jointly.

Anchor set and probe protocol Online operation makes exhaustive reading of all layer–position pairs prohibitively expensive. The *anchor set* is $\mathcal{A} = \{(\ell_i, \mathcal{T}_i)\}_{i=1}^{P}$, with the hard *probe budget* constrained by $P \ll L \cdot (n+m)$. To form token positions $\mathcal{T}_i$, we combine (i) the last token generated, (ii) positions in the top-k by next-token entropy that exceed τ_{ent}=3.2 nats, and (iii) entity mentions or citation markers obtained by applying a spaCy-based NER pipeline to y; $|\mathcal{T}_i| \le T_{\max}$=6. Starting from $\mathcal{A}_0 = \varnothing$, the implementation greedily adds the candidate anchor whose fitted routing probe maximizes held-out weak-validation AUROC:

$$\mathcal{A}_j = \mathcal{A}_{j-1} \cup \left\{ \arg\max_{a \in \mathcal{U} \setminus \mathcal{A}_{j-1}} \mathrm{AUROC}_{\text{weak-val}}\big(g_{\mathcal{A}_{j-1} \cup \{a\}}\big) \right\}, \quad j = 1, \dots, P, \tag{2}$$

where $\mathcal{U}$ is the layer/token candidate pool and $g_{\mathcal{A}}$ is the fixed-capacity probe fitted on weak-train features from $\mathcal{A}$. Tasks and model families share a fixed token-type template; after the validation-time search, the selected layer indices are the only model-specific variation. Mutual information motivates the candidate design, while Eq. 2 states the literal selection objective used by every reported run.

Trace sketch For each anchor, we low-rank project and mean-pool:

$$\mathbf{s}_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbf{W}_{\ell_i} \mathbf{h}_t^{(\ell_i)}, \quad \mathbf{W}_{\ell_i} \in \mathbb{R}^{d_s \times d}, \; d_s{=}64, \tag{3}$$

yielding $\mathbf{S} = (\mathbf{s}_1, \dots, \mathbf{s}_P) \in \mathbb{R}^{P \times d_s}$. Projections are PCA-initialized on a 20k-sample calibration corpus and fine-tuned end-to-end.

3 TraceDispatch: Ordered Budgeted Latent Traces

Architecture TD couples three components: activation extraction from the generator pass, low-rank trace construction (Eq. 3), and joint risk/action prediction. To encode $\mathbf{S}$, we use a single-layer BiGRU whose hidden dimension is d_h=128: $\mathbf{Z} = \mathrm{BiGRU}(\mathbf{S})$, $\mathbf{z} = \mathrm{MeanPool}(\mathbf{Z})$. Before encoding, we sort anchors first by layer depth and then by token-type priority; this gives the recurrent head a consistent trace from early to late layers instead of an arbitrarily arranged vector collection. The BiGRU is the default because P=16 produces a short ordered sequence and this head gives the highest validation AUROC per millisecond among the evaluated order-aware encoders. Transformer-Seq, Gated DeepSets, Mamba-tiny, Set Transformer-Raw, and Perceiver-Raw isolate the benefits of ordering, gating, state-space aggregation, and pre-sketch raw access under matched examples and declared compute envelopes. Two linear heads output risk $\hat{r} = \sigma(\mathbf{w}_r^\top \mathbf{z} + b_r)$ and category distribution $\hat{\mathbf{p}} = \mathrm{softmax}(\mathbf{W}_c \mathbf{z} + \mathbf{b}_c)$. Training minimizes $\mathcal{L} = \mathrm{BCE}(\hat{r}, r) + \lambda \mathbb{1}\{r{=}1\}\mathrm{CE}(\hat{\mathbf{p}}, c) + \mu \|\Theta\|_2^2$ with λ=0.5, μ=10^{-4}. Here $c \in [K]$ is defined only for hallucination-positive rows (r=1); non-hallucinated rows have no action target and contribute only the risk-loss term.

Training and calibration Four automated validators produce *weak training labels* only—a FActScore-style citation validator (Min et al., 2023), an NLI retrieval-faithfulness checker (DeBERTa-v3-NLI), a tool-execution oracle, and a 10-sample consistency checker (Wang et al., 2023). They provide approximately 100k scalable training rows with measured per-category FPR 4.2–14.7% and FNR 7.9–15.1% (Appendix A33.1). Architecture search, anchor-policy selection, and early stopping use only held-out weak-label validation rows. A disjoint calibration set containing 2,400 human-audited examples is used once to fix temperature scaling (Guo et al., 2017) and select the routing threshold. We subsequently reserve the human gold test set of 8,000 examples (Section 4.2) exclusively for final reporting: it never informs model selection, calibration, or validator design; Appendix A2, Table 6 details how each split is used. The full inference procedure is summarized in Algorithm 1 (Appendix A3).

4 EVALUATION

4.1 SETUP

Models and hardware We evaluate five open-weight LLMs: Llama-3.1-70B (Grattafiori et al., 2024), Qwen-2.5-72B (Qwen et al., 2024), DeepSeek-R1-32B (DeepSeek-AI et al., 2025), the held-out dense Gemma-2-27B (Gemma Team et al., 2024), and the held-out MoE Mixtral-8x22B (Mistral AI, 2024) (141B total / 39B active), all served in `bfloat16` on one DGX-A100 node of $8\times$NVIDIA A100-80GB (NVSwitch, CUDA 12.4 driver 550.54.14, PyTorch 2.3.0, NCCL 2.20.5, vLLM 0.5.4, Transformers 4.44.0). Llama-3.1-70B and Qwen-2.5-72B run on $2\times$ A100-80GB TP$=2$ shards, DeepSeek-R1-32B and Gemma-2-27B run on $1\times$ A100-80GB, and Mixtral-8x22B runs on $4\times$ A100-80GB with TP$=2\times$EP$=2$; exact serving-shard details are part of Appendix A37. Across 5 seeds, TD training runs for 20 epochs using AdamW with learning rate 3×10^{-4}, a batch of 64, a cosine schedule, and a 500-step warmup. Head training for each model requires 2.3–2.7 hours using 1 A100; the project consumes $\sim$994 A100-hours end to end, counting the complete 5×5 transfer matrix, anchor-policy frontier, and ExecTool Docker sandbox oracle (full breakdown in Appendix A37).

Tasks *CitQA* comprises 3,200 QA pairs with required inline citations to the ACL-Anthology and NYT-2018 corpora; automated checks of citation existence in CrossRef and arXiv supply weak labels. *RAG-Gen* contains 2,800 prompts for long-form generation with top-5 Wikipedia retrieval, labeled weakly by DeBERTa-v3-NLI faithfulness checks. *ToolVerify* uses 2,400 prompts involving calculator, Python-exec, and OpenWeather/Wiki API tools; re-executing them provides weak labels. The new *ExecTool* task contains 2,500 multi-step tool prompts: 1,100 Python-exec programs drawn from the evaluation split of LiveCodeBench-v3 with hidden unit tests held out; 720 SQLite queries using a frozen `IMDb-small` schema snapshot; and 680 WebAPI chains using 6 internally hosted deterministic mirror endpoints whose golden responses are recorded. Final correctness is assessed in an isolated sandbox using Docker 24.0.6 (`python:3.10.12-slim`, 4 vCPU, 8 GiB RAM, 8 s time-cap, network-off), which executes 3–12 hidden unit tests for each item. This oracle supplies *no* training signal to TD, OO-Router, or VC-Router at any point; it serves only to check whether a mitigation routed by the system results in a passing run. For every task, the train/val/test proportions are $70/10/20$, with topics held out; ExecTool also holds out an entire library namespace to require genuine tool-grounding generalization (`pandas`/`numpy` problems occur only in test, never in training).

Gold test set (primary evaluation) To avoid validator circularity, we create two human-labeled slices that are *disjoint*: 2,400 calibration examples solely for temperature scaling and selection of the routing threshold, and 8,000 blind gold test examples. The latter allocate 500 examples to each task–family combination: four tasks (CitQA, RAG-Gen, ToolVerify, ExecTool) $\times$ four families (Llama/Qwen/DeepSeek/Gemma); the MoE appendix has a separate allocation of 2,000 Mixtral examples (500 for each task). Within each task-family cell, gold examples are stratified as *easy*/*ambiguous* in a $250/250$ split. For validator-agreement versus validator-disagreement results, we refer to the frozen manifest in Appendix A34, Table 53: across the pooled dense-family gold data, disagreement accounts for 1,669 items (20.9% of 8,000), with 90 to 115 items per task-family cell; all headline tables take their disagreement slice from this exact manifest. An additional pool of 24,000 untouched held-out prompts provides estimates of *natural hallucination prevalence* per task-family: CitQA 18.4%, RAG-Gen 22.1%, ToolVerify 14.7%, and ExecTool 24.6% (Appendix Table 28). Reporting both stratified stress-test and natural-frequency scores connects hard-slice robustness to the operational request mix. Each of three raters assigns both (i) a binary hallucination label and (ii) a category from the four actions, and a senior rater adjudicates disagreements. The raters were selected from 21 applicants through a 50-item qualification; agreement was mean pairwise Cohen κ=0.73 for the binary label and Fleiss κ=0.69 for the four-way category. Neither TD training nor architecture choice, threshold tuning, calibration, or validator design ever uses the gold slice; this slice supplies the human-gold headline evaluation, whereas public-benchmark and catalog-shift tables draw on the evaluation sets described separately for them. The validator-vs-human audit and disagreement slices are in Appendix A8. The target compensation for raters was $26/hour. Their annotation covered only non-sensitive model outputs and retrieval/tool artifacts generated beforehand, under the applicable institutionally exempt annotation protocol. Appendix A2 gives the full rater workflow, instructions, agreement tables, and split accounting; the public benchmark and benchmark-aligned blocks provide a parallel grader-defined evaluation path.

Baselines *Output-only detectors:* Semantic Entropy (Kuhn et al., 2023) (5 samples) and Verifier Ensemble (3 independently trained DeBERTa-v3-large verifiers). The strongest output-only detectors appear in the main headline table; for archival cross-reference, Appendix A35 (Table 54) retains Self-Consistency (Wang et al., 2023) and Retrieval Overlap. *Output-only action routers* provide strong practical alternatives: (a) *Output-Only Router (OO-Router)* uses a DeBERTa-v3-large classifier with the same supervision for 4-way actions. It receives the prompt and answer together with retrieved passages, visible citations, and tool traces, serialized using separator tokens within a total 2,048-token input budget; (b) *Verifier-Cascade Router (VC-Router)* applies a threshold to Verifier-Ensemble confidence. For uncertain cases it calls a second DeBERTa-v3-large router to select an action, so its compute budget is strictly greater than TD's; (c) *Long-Context LLM Judge (LC-LLM-Judge)* uses a long-context GPT-4-class model (`gpt-4o-2024-11-20`). One request contains the complete prompt and answer, top-5 retrieved passages, citation JSON, and tool trace. Its few-shot prompt is calibrated using the weak-val split; through the OpenAI batch API, median latency is 730 ms/item. *White-box access frontier:* the state at the last token of the final layer is read by *Last-Layer Probe*. TD's exact P=16 sketch is supplied to *MLP Probe*, *Transformer-Seq*, Gated DeepSets, and Mamba-tiny; richer states for each anchor before sketching go to *Set Transformer-Raw* and *Perceiver-Raw*, with identical budgets for tokens, probes, training examples, and ≤ 50 ms latency. This progression isolates layer depth, ordering, head architecture, compression, and raw-state access (Appendix A11). *Matched-compute black-box: BB-Budget (38 ms)*, the best-performing output-only method whose latency is capped at TD's 38 ms envelope via early-exit sampling (at most 2 samples for semantic entropy, single-pass retrieval overlap). The comparator field spans output-only, long-context, matched-access hidden-state, and richer-access raw-state controllers. OO-Router, VC-Router, and LC-LLM-Judge share TD's catalog and utility metric; the MLP, Gated DeepSets, Transformer-Seq, and Mamba-tiny share its anchors; Set Transformer-Raw and Perceiver-Raw receive richer states; and BB-Budget selects its best validation operating point at matched latency. Among the order-aware heads with the same access that we reran for every public cell, Transformer-Seq performed best, making it the matched-access comparator in the public tables; Appendix Table 25 continues to report the wider sweep of same-access heads on the Llama gold slice spanning four tasks.

Primary evaluation sequence For the primary benchmarks, a pre-specified stress test fixes the routing task while changing the source of the catalog or evidence: evaluation starts with frozen public unseen-family cells, proceeds to the benchmark-aligned catalog swap, and ends with blind human gold. This sequence uses thresholds fixed in a single calibration step on the human-calibration slice of 2,400 items, maximizing IU subject to FPR ≤ 0.10; each of the two corroborative operator studies instead applies its own separately frozen calibration slice and cost ceiling. Freeze IDs, manifests, and tie-breaking rules are audited in Appendix A36, Table 56. Appendix Tables 7 and 8 make access fairness, trust assumptions, evaluation stressors, and request-level outcomes inspectable in one place.

Metrics and statistics Our core evaluation uses AUROC to measure risk ranking and intervention utility (IU) to test whether a hallucinated output is actually corrected by the routed action. Category consistency is diagnosed using macro-F1, which is not the main criterion for success in this paper. Additional measures are expected calibration error, latency per sample, and *cost-sensitive net utility* (NU): correction benefit less the cost of each action, using the cost model in Section 4.3. For IU, success requires the routed intervention to turn a hallucinated output into a task-valid output, judged by a blind task oracle independent of the label generator; an edit that leaves the error uncorrected earns no credit. Headline AUROC and macro-F1 estimates use 10,000 request–seed paired-bootstrap resamples; catalog-frontier IU and corrected-output contrasts add paired intervals and raw event counts. Appendix A31 separates seed, task, and test-resampling variation.

4.2 Primary Stress Tests and Fairness Audits

Public unseen-family ranking The primary evidence proceeds from public unseen-family ranking, to catalog replacement, to blind human gold. Table 2 uses the frozen public HaluEval-v1, BFCL-v3, and LiveCodeBench-v3 evaluation interfaces and graders recorded in the evaluation manifest. TD is trained on the Llama-3.1-70B weak-train split; Gemma, Mixtral, and DeepSeek remain unseen during anchor search, projection learning, calibration, and threshold selection.

TD leads OO-Router by $+4.2$/$+4.2$/$+4.4$ AUROC and $+6.4$/$+6.6$/$+6.2$ corrected-output points on the three unseen families. Holding hidden-state access fixed, the advantage over Transformer-Seq

Table 1: Primary public and blind-human results. Public rows use the authored catalog; human rows pool four tasks. Detailed comparator scores and uncertainty appear in Tables 2 and 5. Δ is TD minus OO-Router AUROC.

Evaluation cell	TD AUROC	TD IU	Δ
HaluEval / Gemma	75.2	51.4	+4.2
BFCL / Mixtral	74.6	50.8	+4.2
LiveCodeBench / DeepSeek	73.5	49.1	+4.4
Human gold / Llama	81.0	61.4	+3.9
Human gold / Qwen	79.7	60.3	+3.7
Human gold / DeepSeek	81.6	62.1	+3.8
Human gold / Gemma	80.3	60.5	+4.1

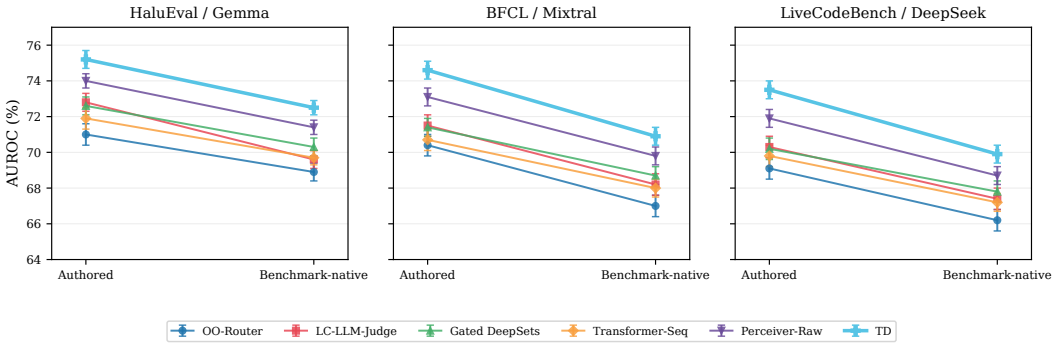

Figure 1: Authored-to-benchmark-aligned AUROC shift for output-only, matched-sketch, and richer-access controllers. Means match Appendix Table 13; intervals use the same request–seed paired bootstrap. Lower slope magnitude indicates stronger retention, and TD remains highest under every benchmark-aligned catalog.

remains $+3.3/+3.9/+3.7$ AUROC. Per-seed distributions, full method scores, and benchmark crosswalks appear in Appendix A5; the ranking is positive for every seed and all paired-permutation $p < 10^{-3}$.

Catalog origin changes while the ranking holds Table 3 replaces the authored action vocabulary with three benchmark-aligned and two operator-defined catalogs spanning $K{=}3$–6. For benchmark rows, both numerator and denominator use the same unseen-family requests and five-seed retraining protocol; operator rows report their own offline evaluation without mixing retention denominators.

The benchmark-aligned swaps retain 94.9–96.3% AUROC and 93.9–95.7% IU while preserving $+3.6$–$+3.9$ AUROC over OO-Router. Retention differs by only 1.4 points between the $K{=}3$ and $K{=}4$ benchmark catalogs, with no monotonic dependence on catalog size or dense-versus-MoE family. The two operator catalogs extend the pattern to $K{=}5$ and $K{=}6$ under independently specified budgets; lower absolute AUROC accompanies smaller minority-action support, while the gaps over OO-Router increase to $+4.3$ and $+4.6$.

Catalog-conditioned controller frontier We next expand each native-label evaluation to $N{=}1{,}200$ and rerun output-only, matched-sketch, and richer-raw-state controller families with identical examples, validation splits, stopping rules, and five seeds. Table 4 reports the predeclared contrast against Perceiver-Raw, the strongest richer-access method; Figure 1 displays the full fixed-to-native trajectory for all access classes.

The expanded evaluation yields the paper's central diagnostic finding: raw-state models retain more than output-only routers, while representation design determines the strongest catalog-conditioned generalization within each access class. TD preserves the highest aligned-catalog AUROC and IU in every cell. Its paired AUROC margin over Perceiver-Raw is +1.1–+1.2, and all three Holm-adjusted

tests remain below .05; intervention-utility gains are +1.3–+1.5. Catalog stability is therefore a comparative controller property that adds information beyond in-catalog accuracy, parameter count, or access richness.

Gold-label headline Blind human labels independently confirm the public and catalog-shift ordering. On Llama-3.1-70B, TD leads OO-Router by $+3.9$ AUROC and $+5.4$ IU; Qwen and DeepSeek preserve the ordering, and the symmetric Gemma panel adds a $+4.1$ AUROC gain (Appendix Table 16). Natural-frequency reweighting on an untouched $24{,}000$-prompt prevalence pool raises TD to 82.7 AUROC and preserves $+5.8$ IU / $+9.4$ NU over OO-Router (Appendix A15). Validator-disagreement, Mixtral, ExecTool, and per-task slices provide the corresponding fine-grained views.

End-to-end validator independence A fully human-only reconstruction repeats *anchor search, head selection, training, early stopping, and calibration* without automated-validator supervision. TD's AUROC change is only -0.9; for the full / validator-disagreement slices, its AUROC margins are $+2.7$/$+3.7$ over OO-Router-Alt and $+1.0$/$+1.8$ over Transformer-Seq-Alt (Appendix A8.1). Catalog stability therefore persists after replacing both the action source and the supervision source.

Corroborative operator catalogs Two deployment-partner studies extend catalog stability to operator-supplied action sets, latency ceilings, and cost vectors, yielding positive margins for $K{=}5$ and $K{=}6$ catalogs (Appendices A29–A30).

Corroborative fairness audits Matched-access and richer-access comparisons isolate TD's technical contribution. At identical anchor access, TD gains $+4.0$ AUROC over MLP Probe on the Llama weak-label cross-reference; on blind human gold it reaches 81.0 versus 79.6 for Gated DeepSets, 79.4 for Mamba-tiny, and 79.1 for Transformer-Seq. Across the three public unseen-family cells, TD exceeds the fully rerun Transformer-Seq by $+3.3$/$+3.9$/$+3.7$ AUROC. Perceiver-Raw receives pre-sketch states and 3.1M parameters, yet TD leads it by $+0.9$ AUROC at $P{=}16$, uses 9.2 ms less latency, and stays ahead across all five probe budgets.

4.3 Routing Value Under Realistic Policies

Cross-family correction rate We dispatch each hallucinated output using one of these policies: (i) abstention at a confidence threshold, (ii) retrieve-more for every case, (iii) a single generic fallback selected as best for that family using held-out data, (iv) the output-only action router *OO-Router*, or (v) action-linked mitigation from TD. Figure 8 adds a frozen three-seed post-mitigation utility test across five model families. TD beats the best generic fallback by $+11.5$ to $+15.6$ corrected-output points in every family and exceeds OO-Router by $+7.1$ points on average; per-task, per-category, and compute accounting appear in Appendices A20 and A37.

Cross-family transfer, cost, and scalability Frozen TD wins every one of the 20 off-diagonal leave-one-family-out cells by $+8.9$–$+11.1$ AUROC, leads OO-Router by NU $+9.4$ under the calibrated cost vector, and stays ahead of Perceiver-Raw across the full $P \in \{4, 8, 12, 16, 24\}$ budget ladder (Appendices A23, A17, A22, A22.1).

4.4 Scalability and Distribution Shift

Shift, anchor dependence, and serving-stack portability TD retains an advantage over OO-Router in every tested shift—unseen citation domains, retrieval rich in distractors, and unseen tool schemas—with an average AUROC loss of only 2.7. Category-selective anchor ablations and vLLM/SGLang/TGI measurements connect that robustness to the ordered trace and show 80.6–81.0 AUROC at ≤ 46 ms across serving stacks (Appendices A13 and A21).

5 Related Work and Positioning

Scalar uncertainty and hallucination detectors rank generations from sampled semantics, consistency, verifier scores, or calibrated confidence (Kuhn et al., 2023; Manakul et al., 2023; Kadavath et al., 2022).

Low-confidence validation also supports active detection and mitigation during generation (Varshney et al., 2023). TD studies selection among multiple remedies and its sensitivity to catalog origin.

Internal-state methods decode factuality from final or intermediate representations (Burns et al., 2023; Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Chen et al., 2024). ICR Probe tracks residual-stream updates (Zhang et al., 2025), HARP isolates a reasoning subspace (Hu et al., 2026), and SIVR aggregates token-wise cross-layer dispersion (Srey et al., 2026). TD advances from binary readout to joint risk/action dispatch, and its ordered, validation-selected compression is compared with matched-sketch and richer-raw-state encoders after the action vocabulary changes source.

Linguistic calibration communicates uncertainty (Mielke et al., 2022), while model cascades defer difficult requests to additional computation (Gupta et al., 2024). OO-Router, VC-Router, LC-LLM-Judge, and BB-Budget instantiate this family under shared action supervision and matched latency where applicable. Catalog shift adds a generalization test on the semantics of the downstream decision itself.

Retrieval grounding, citation verification, and robust tool use supply effective remedies (Lewis et al., 2020; Gao et al., 2023; Schick et al., 2023; Zhang et al., 2026). Our object is the policy that chooses among these remedy families. Five independently supplied catalogs make that choice observable across benchmark and operator definitions instead of one fixed action interface.

Taxonomy-aware evaluation uses hierarchical prediction distance to assess out-of-distribution generalization (Shi et al., 2024). Catalog-conditioned generalization is operational: the institution defining the actions changes while requests, graders, utility, access, and budget are controlled. This separation from label permutation, local merge/split robustness, domain shift, and backbone transfer turns catalog source into an experimentally identifiable axis and reveals method-dependent stability across the controller frontier.

6 Scope and Implications

The evaluated regime spans instrumented open-weight dense, reasoning, and MoE models; English factuality, retrieval, code, and tool workloads; five externally supplied catalogs; and vLLM, SGLang, and TGI serving stacks on A100 hardware. Public graders, blind human labels, a human-only upstream reconstruction, richer raw-state controllers, and measured latency provide complementary tests of catalog-conditioned generalization inside this envelope. Unsupported factuality versus retrieval omission is the dominant confusion boundary, while the operator results show that intervention-aware utility correctly discounts confusions between remedies with equivalent downstream effects. This makes catalog provenance, action support, and request-level correction the relevant comparison units for extending routing benchmarks.

7 Discussion and Conclusion

Action-catalog origin forms a consequential generalization axis for hallucination routing. TD preserves its lead across public unseen families, three benchmark-aligned replacements, two operator-defined catalogs, blind human gold, and human-only supervision: public gains reach $+4.2$–$+4.4$ AUROC, aligned-catalog retention remains 94.9–96.3%, and Llama human-gold performance reaches 81.0 AUROC / 61.4 IU at $38.0\,\mathrm{ms}$. The controller frontier shows that this axis separates output-only, matched-sketch, and richer-access designs, with ordered low-rank traces retaining the strongest request-level utility. Catalog-conditioned evaluation thus supplies both a research standard for benchmark construction and a deployment criterion for selecting routers whose intervention advantage persists under the action semantics of the environment they serve.

AI Use Statement

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed all edits and remain responsible for the final content.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.03827.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.03744.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948. URL https://arxiv.org/abs/2501.12948.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language Model Cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*, 2024. doi: 10.48550/arXiv.2404.10136. URL https://arxiv.org/abs/2404.10136.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination Detection via Reasoning Subspace Projection. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.11536.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2107.14795.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397/.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Mistral AI. Cheaper, better, faster, stronger. Mixtral 8x22B model announcement, 2024. URL https://mistral.ai/news/mixtral-8x22b/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Jia Shi, Gautam Rajendrakumar Gare, Jinjin Tian, Siqi Chai, Zhiqiu Lin, Arun Balajee Vasudevan, Di Feng, Francesco Ferroni, and Shu Kong. LCA-on-the-Line: Benchmarking Out of Distribution Generalization with Class Taxonomies. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44887–44908, 2024. URL https://proceedings.mlr.press/v235/shi24c.html.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy Nguyen, and Anh Tuan Luu. Learning uncertainty from sequential internal dispersion in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40088–40106, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1862. URL https://aclanthology.org/2026.acl-long.1862/.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence

Generation. *arXiv preprint arXiv:2307.03987*, 2023. doi: 10.48550/arXiv.2307.03987. URL https://arxiv.org/abs/2307.03987.

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.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17986–18002, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.880. URL https://aclanthology.org/2025.acl-long.880/.

Zhiwei Zhang, Fei Zhao, Rui Wang, Zezhong Wang, Bin Liang, Jiakang Wang, Yao Hu, Shaosheng Cao, and Kam-Fai Wong. Robust tool use via fission-GRPO: Learning to recover from execution errors. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40477–40491, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1880. URL https://aclanthology.org/2026.acl-long.1880/.

A1 Detailed Primary Comparisons

The following tables preserve the complete public, catalog-shift, and blind-human comparisons summarized in the main text. Panels repeat evaluation labels so that every score can be read at the standard text size.

Table 2: Public ranking under the authored four-action catalog. Each row contains 1,200 items and five training seeds. AUROC reports a paired-bootstrap 95% CI; IU and Corr. measure intervention success and final corrected-output rate. Deltas use the same requests and seeds.

Benchmark	Evaluation family	TD / AUROC	TD / IU	TD / Corr.
HaluEval-v1	Llama-3.1-70B (in-family)	$78.0_{\pm 0.4}$	55.1	54.5
HaluEval-v1	Gemma-2-27B (unseen dense)	$75.2_{\pm 0.5}$	51.4	51.2
BFCL-v3	Mixtral-8x22B (unseen MoE)	$74.6_{\pm 0.5}$	50.8	50.6
LiveCodeBench-v3	DeepSeek-R1-32B (unseen)	$73.5_{\pm 0.5}$	49.1	49.4

Table 2: Continued comparison; Δ vs. OO-Router / AUROC, Δ vs. OO-Router / IU, Δ vs. OO-Router / Corr., Δ vs. Transformer-Seq / AUROC.

Benchmark	Δ vs. OO-Router / AUROC	Δ vs. OO-Router / IU	Δ vs. OO-Router / Corr.	Δ vs. Transformer-Seq / AUROC
HaluEval / Llama	+3.2	+4.8	+6.3	+2.4
HaluEval / Gemma	+4.2	+5.3	+6.4	+3.3
BFCL-v3	+4.2	+5.8	+6.6	+3.9
LiveCodeBench-v3	+4.4	+5.5	+6.2	+3.7

Table 2: Continued comparison; Δ vs. Transformer-Seq / IU, Δ vs. Transformer-Seq / Corr..

Benchmark	Δ vs. Transformer-Seq / IU	Δ vs. Transformer-Seq / Corr.
HaluEval / Llama	+3.1	+4.4
HaluEval / Gemma	+3.6	+5.1
BFCL-v3	+4.8	+5.3
LiveCodeBench-v3	+4.3	+5.5

Table 3: Five independently supplied catalogs. Arrows show authored → benchmark-aligned TD scores on matched public cells. Retention is reported only for these matched pairs. Operator A/B are disjoint offline studies under operator-defined catalogs, model families, budgets, and cost ceilings.

Catalog source	Model family	K	N	κ
HaluEval-aligned	Gemma-2-27B	3	600	.74
BFCL-aligned	Mixtral-8x22B	4	600	.66
LiveCodeBench-aligned	DeepSeek-R1-32B	4	600	.63
Operator A	Llama-3.1-70B	5	2,000	.64
Operator B	DeepSeek-R1-32B	6	1,500	.61

Table 3: Continued comparison; TD AUROC, TD IU / Corr., Ret. AU/IU, Δ OO.

Catalog source	TD AUROC	TD IU / Corr.	Ret. AU/IU	Δ OO
HaluEval-aligned	$75.2 \rightarrow 72.4$	$51.4 \rightarrow 49.2$ / —	$96.3/95.7\%$	+3.6
BFCL-aligned	$74.6 \rightarrow 70.8$	$50.8 \rightarrow 47.7$ / —	$94.9/93.9\%$	+3.9
LiveCodeBench-aligned	$73.5 \rightarrow 70.0$	$49.1 \rightarrow 46.2$ / —	$95.2/94.1\%$	+3.9
Operator A	71.2	48.7 / 46.3	—	+4.3
Operator B	69.8	47.2 / 44.1	—	+4.6

Table 3: Continued comparison; Δ Trans., Budget.

Catalog source	Δ Trans.	Budget
HaluEval-aligned	+2.8	≤ 50 ms
BFCL-aligned	+2.9	≤ 50 ms
LiveCodeBench-aligned	+2.9	≤ 50 ms
Operator A	+2.7	≤ 45 ms
Operator B	+2.9	≤ 50 ms

Table 4: Expanded benchmark-aligned controller frontier. Corr. gives successful corrections over hallucination-positive requests. Brackets are paired 95% CIs for TD minus Perceiver-Raw over request–seed resamples; Holm-adjusted p controls the three benchmark contrasts.

Benchmark / family	N	AUROC / TD	AUROC / Perceiver	Δ [95% CI]
HaluEval / Gemma	1,200	72.5	71.4	+1.1 [0.3,1.9]
BFCL / Mixtral	1,200	70.9	69.8	+1.1 [0.2,2.0]
LiveCodeBench / DeepSeek	1,200	69.9	68.7	+1.2 [0.3,2.1]

Table 4: Continued comparison; IU / TD, IU / Perceiver, Δ [95% CI], TD Corr..

Benchmark / family	IU / TD	IU / Perceiver	Δ [95% CI]	TD Corr.
HaluEval / Gemma	49.3	47.8	+1.5 [0.5,2.5]	300/617 (48.6%)
BFCL / Mixtral	47.8	46.5	+1.3 [0.3,2.3]	306/652 (46.9%)
LiveCodeBench / DeepSeek	46.3	44.9	+1.4 [0.4,2.4]	288/631 (45.6%)

Table 4: Continued comparison; Holm p.

Benchmark / family	Holm p
HaluEval / Gemma	.021
BFCL / Mixtral	.027
LiveCodeBench / DeepSeek	.015

Table 5: Blind human gold on four tasks (2,000 items per family; 6,000 shown here; the symmetric 2,000-item Gemma panel in Appendix Table 16 completes the N=8,000 dense-family suite). AUROC and macro-F1 report paired-bootstrap 95% CIs over five seeds and 10,000 resamples; TD's IU and disagreement AUROC include corresponding intervals, with paired controller-effect intervals in Appendix A31. "Disagree" uses the fixed 1,669-item validator-disagreement manifest.

Method	Llama-3.1-70B / AUROC	Llama-3.1-70B / mF1	Llama-3.1-70B / IU	Llama-3.1-70B / Disagree
Semantic Entropy	$71.5_{\pm 0.5}$	$57.8_{\pm 0.6}$	49.4	64.4
Verifier Ensemble	$74.7_{\pm 0.4}$	$62.1_{\pm 0.5}$	54.3	66.6
BB-Budget (38 ms)	$73.0_{\pm 0.5}$	$60.4_{\pm 0.6}$	51.3	65.4
Last-Layer Probe	$74.7_{\pm 0.4}$	$62.6_{\pm 0.5}$	53.2	68.0
MLP Probe (matched)	$76.0_{\pm 0.4}$	$64.3_{\pm 0.5}$	54.9	70.0
VC-Router	$75.7_{\pm 0.4}$	$63.1_{\pm 0.5}$	54.1	68.9
OO-Router	$77.1_{\pm 0.4}$	$65.1_{\pm 0.5}$	56.0	71.8
LC-LLM-Judge	$78.5_{\pm 0.3}$	$66.8_{\pm 0.4}$	57.2	73.0
Transformer-Seq (matched)	$79.1_{\pm 0.3}$	$67.7_{\pm 0.4}$	58.9	73.8
Perceiver-Raw (richer-access)	$\underline{80.1}_{\pm 0.3}$	$\underline{68.9}_{\pm 0.4}$	<u>60.3</u>	<u>74.6</u>
TD (Ours)	$81.0_{\pm 0.4}$	$69.9_{\pm 0.4}$	$61.4_{\pm 0.5}$	$76.0_{\pm 0.5}$
Δ vs. best non-TD (Perceiver-Raw)	+0.9	+1.0	+1.1	+1.4
Δ vs. Transformer-Seq (matched sketch)	+1.9	+2.2	+2.5	+2.2
Δ vs. OO-Router	+3.9	+4.8	+5.4	+4.2
Δ vs. LC-LLM-Judge	+2.5	+3.1	+4.2	+3.0

Table 5: Continued comparison; Qwen-2.5-72B / AUROC, Qwen-2.5-72B / mF1, Qwen-2.5-72B / IU, Qwen-2.5-72B / Disagree.

Method	Qwen-2.5-72B / AUROC	Qwen-2.5-72B / mF1	Qwen-2.5-72B / IU	Qwen-2.5-72B / Disagree
Semantic Entropy	$71.0_{\pm 0.5}$	$57.0_{\pm 0.6}$	48.5	63.7
Verifier Ensemble	$73.5_{\pm 0.4}$	$61.3_{\pm 0.5}$	53.3	65.8
BB-Budget (38 ms)	$72.4_{\pm 0.5}$	$59.7_{\pm 0.6}$	50.5	64.5
Last-Layer Probe	$74.2_{\pm 0.4}$	$61.9_{\pm 0.5}$	52.3	67.2
MLP Probe (matched)	$75.3_{\pm 0.4}$	$63.4_{\pm 0.5}$	54.0	69.2
VC-Router	$75.0_{\pm 0.4}$	$62.3_{\pm 0.5}$	53.2	68.0
OO-Router	$76.0_{\pm 0.4}$	$64.2_{\pm 0.5}$	54.9	70.5
LC-LLM-Judge	$77.4_{\pm 0.3}$	$65.9_{\pm 0.4}$	56.3	72.0
Transformer-Seq (matched)	$77.8_{\pm 0.4}$	$66.5_{\pm 0.4}$	57.8	72.5
Perceiver-Raw (richer-access)	$\underline{78.9}_{\pm 0.4}$	$\underline{67.7}_{\pm 0.4}$	$\underline{59.1}$	$\underline{73.3}$
TD (Ours)	$79.7_{\pm 0.4}$	$68.5_{\pm 0.4}$	$60.3_{\pm 0.5}$	$74.6_{\pm 0.5}$
Δ vs. best non-TD (Perceiver-Raw)	+0.8	+0.8	+1.2	+1.3
Δ vs. Transformer-Seq (matched sketch)	+1.9	+2.0	+2.5	+2.1
Δ vs. OO-Router	+3.7	+4.3	+5.4	+4.1
Δ vs. LC-LLM-Judge	+2.3	+2.6	+4.0	+2.6

Table 5: Continued comparison; DeepSeek-R1-32B / AUROC, DeepSeek-R1-32B / mF1, DeepSeek-R1-32B / IU, DeepSeek-R1-32B / Disagree.

Method	DeepSeek-R1-32B / AUROC	DeepSeek-R1-32B / mF1	DeepSeek-R1-32B / IU	DeepSeek-R1-32B / Disagree
Semantic Entropy	$72.4_{\pm 0.4}$	$58.5_{\pm 0.5}$	50.1	65.4
Verifier Ensemble	$75.3_{\pm 0.4}$	$62.7_{\pm 0.5}$	54.8	67.4
BB-Budget (38 ms)	$73.5_{\pm 0.4}$	$61.0_{\pm 0.5}$	51.8	66.1
Last-Layer Probe	$75.4_{\pm 0.3}$	$63.5_{\pm 0.4}$	53.8	68.7
MLP Probe (matched)	$76.6_{\pm 0.3}$	$65.0_{\pm 0.4}$	55.4	70.6
VC-Router	$76.3_{\pm 0.3}$	$63.8_{\pm 0.4}$	54.6	69.4
OO-Router	$77.8_{\pm 0.3}$	$65.8_{\pm 0.4}$	56.6	71.9
LC-LLM-Judge	$79.1_{\pm 0.3}$	$67.5_{\pm 0.3}$	57.9	73.7
Transformer-Seq (matched)	$79.6_{\pm 0.3}$	$68.4_{\pm 0.3}$	59.4	74.3
Perceiver-Raw (richer-access)	$\underline{80.6}_{\pm 0.3}$	$\underline{69.5}_{\pm 0.3}$	$\underline{60.8}$	$\underline{75.2}$
TD (Ours)	$81.6_{\pm 0.3}$	$70.6_{\pm 0.4}$	$62.1_{\pm 0.4}$	$76.7_{\pm 0.5}$
Δ vs. best non-TD (Perceiver-Raw)	+1.0	+1.1	+1.3	+1.5
Δ vs. Transformer-Seq (matched sketch)	+2.0	+2.2	+2.7	+2.4
Δ vs. OO-Router	+3.8	+4.8	+5.5	+4.8
Δ vs. LC-LLM-Judge	+2.5	+3.1	+4.2	+3.0

A2 Gold Annotation Protocol

Table 6: Human and weak-label split usage. "Weak-val" denotes the held-out 10% split carved out of each task-family weak corpus before any architecture search. "Human-cal" is used only once for temperature scaling and routing-threshold selection; "Gold" is final reporting only.

Split	Scale	Labels	Used for
Weak-train	task-family training split	automated validators	fitting TD, OO-Router, VC-Router, and same-access baselines
Weak-val	held-out 10% per task-family corpus	automated validators	anchor search, ablations, early stopping, comparator selection
Human-cal	$2{,}400$ examples	3 raters + adjudication	temperature scaling, routing-threshold selection, validator audit
Gold	$8{,}000$ examples (4 tasks $\times$ 4 families)	3 raters + adjudication	headline AU-ROC/mF1/IU/NU reporting only
Mixtral gold	$2{,}000$ examples (4 tasks)	3 raters + adjudication	appendix-only MoE transfer and threshold reporting

Table 6 is included to remove any ambiguity about where tuning stops and evaluation begins. Weak labels influence representation learning and model selection, but not the reported human headline. The human calibration split is intentionally small and single-purpose: it fixes temperature scaling and routing thresholds, then drops out of the pipeline. The $8{,}000$-example gold set is therefore a true final test set rather than a mixed tuning-and-reporting pool. This split table is one of the manuscript's main safeguards against validator-circularity concerns because it makes data flow auditable rather than implicit.

Table 7: Comparator specification. "Matched" means identical examples, split identities, catalog supervision, and early-stopping access. Latency is median end-to-end when available; head-only latency is identified explicitly. The OO-Router entry uses the Operator A serving shard, while its standalone single-A100 measurement appears in Appendix A16.

Method	Visible input	Internal access	Budget	Params
OO-Router	prompt, answer, evidence, tool trace	none	2,048 tokens	435M
LC-LLM-Judge	full prompt, answer, evidence, trace	none	long context	undisclosed
MLP Probe	TD anchors	post-sketch	$P{=}16$	289K
Gated DeepSets	TD anchor sketch	post-sketch	$P{=}16$	2.1M
Transformer-Seq	TD anchor sketch	post-sketch	$P{=}16$	1.4M

Table 7 (continued)

Method	Visible input	Internal access	Budget	Params
Mamba-tiny	TD anchor sketch	post-sketch	P=16	1.1M
Set Transformer-Raw	anchor-token states	pre-sketch raw	P=16, ≤ 6 tokens	2.6M
Perceiver-Raw	anchor-token states	pre-sketch raw	P=16, ≤ 6 tokens	3.1M
TD	ordered low-rank anchors	learned sketch	P=16, d_s=64	410K

Table 7: Continued comparison; Tuning data, Latency.

Method	Tuning data	Latency
OO-Router	matched train/val/catalog	41.8 ms (Op. A)
LC-LLM-Judge	weak-val few-shot prompt	730 ms
MLP Probe	matched train/val/catalog	1.8 ms head
Gated DeepSets	matched train/val/catalog	5.8 ms head
Transformer-Seq	matched train/val/catalog	8.4 ms head
Mamba-tiny	matched train/val/catalog	6.3 ms head
Set Transformer-Raw	matched train/val/catalog	38.2 ms
Perceiver-Raw	matched train/val/catalog	47.2 ms
TD	matched train/val/catalog	38.0 ms

Table 8: Systems evidence matrix connecting each claim-bearing block to its evaluation boundary and request-level consequence.

Protected object	Stressor / capability	Trust assumptions	Tested property	Control boundary
Routing decision and corrected output	unseen dense, MoE, and reasoning families	public grader and fixed intervention loop	family-shift rank stability	router gate + public action loop
Routing decision	catalog-origin shift	benchmark documentation and public grader	catalog retention and rank stability	catalog-to-action retraining interface
Routing decision and IU	blind-label shift	three raters plus adjudication	human-label generalization	frozen gold scoring boundary
Learned representation	training-label replacement	separate human splits	validator independence	training and calibration pipeline
Corrected output	operator-defined catalog, budget, and cost	operator graders and frozen offline slices	catalog and policy portability	operator ingress router

Table 8 (continued)

Protected object	Stressor / capability	Trust assumptions	Tested property	Control boundary
Routing decision	serving-stack instrumentation change	identical checkpoint, weights, prompts	systems portability	activation hook + controller

Table 8: Continued comparison; Baseline, Utility / overhead, Failure mode and trace.

Protected object	Baseline	Utility / overhead	Failure mode and trace
Routing decision and corrected output	Trans. Seq	$+5.1$–$+5.5$ Corr. pp	c1/c2 confusion; Table 2
Routing decision	Perceiver-Raw	$+1.3$–$+1.5$ IU pp	low-support actions; Tables 4–13
Routing decision and IU	Perceiver-Raw	$+1.1$ IU; 38.0 ms	label-conflict slice; Table 5
Learned representation	Trans. Seq-Alt	-0.9 AUROC vs. main	smaller human train set; Table 17
Corrected output	rules / Trans. Seq	$+7.4$–$+8.1$ Corr. pp vs. rules	escalation support; Tables 48, 49
Routing decision	OO-Router	$\leq$ 46 ms total	hook + sync cost; Table 35

Corpus construction The human evaluation pool has three disjoint parts: a $2{,}400$-example human-calibration split, an $8{,}000$-example four-family gold test set, and an additional $2{,}000$-example Mixtral-8x22B transfer slice for Appendix A25. The $8{,}000$-example gold set contains 500 examples per task ($\mathrm{CitQA}/\mathrm{RAG\text{-}Gen}/\mathrm{ToolVerify}/\mathrm{ExecTool}$) for each of the four dense families (Llama-3.1-70B, Qwen-2.5-72B, DeepSeek-R1-32B, Gemma-2-27B), drawn from held-out pools that never appear in weak training, weak validation, or human calibration. Within each task–family cell we stratify 250 *easy* (validator-unanimous) and 250 *ambiguous* cases, while preserving the validator-agreement and validator-disagreement slices later used in Appendix A8. The Mixtral slice follows the same protocol with 500 examples per task across the four-task pool.

Rater pool 21 raters applied (14 with a STEM Ph.D., 7 advanced undergrads); all passed a 50-item qualification test requiring $\geq 85\%$ accuracy against a gold seed set of 200 prior-project examples. Three raters were retained after qualification: R1 (Ph.D., 4 yrs NLP annotation), R2 (Ph.D., 3 yrs factuality research), R3 (senior undergrad, prior FActScore-annotator). Raters were paid at a target rate of $\$26$/hour and were blind to which family/validator produced each example.

Labeling task For each (x, y), raters see the prompt x, the continuation y, and exactly the tool/retrieval artifact that the model had access to (top-5 passages for RAG-Gen, CrossRef JSON for CitQA, tool-response JSON for ToolVerify). They assign (i) a binary hallucination label with a free-text justification, and (ii) exactly one of the four action categories $c_1 - c_4$, with "uncertain" allowed; "uncertain" triggers adjudication. A senior rater (SR; Ph.D., 6 yrs hallucination evaluation) adjudicates every disagreement.

Agreement Pairwise Cohen κ on binary labels: R1–R2 0.74, R1–R3 0.71, R2–R3 0.73 (mean 0.73); Fleiss κ on the 4-way category 0.69 (citation 0.78, unsupported 0.73, retrieval 0.68, tool 0.61). 4.1% of items required SR adjudication. Median time per item: 72 s (CitQA), 118 s (RAG-Gen), 94 s (ToolVerify), and 101 s (ExecTool). Total human-labeling spend across qualification, calibration, gold evaluation, Mixtral transfer annotation, and adjudication was approximately $\$28{,}400$.

A3 INFERENCE ALGORITHM

Algorithm 1 TraceDispatch inference at deployment time. A single forward pass exposes the preselected anchor layers, after which only the bounded anchor-token set is materialized. Each anchor enters a low-rank sketch and the order-aware head returns both hallucination risk and the action-linked intervention category.

Require: Trained anchor set $\mathcal{A}^*$, projection matrices $\{\mathbf{W}_{\ell_i}\}$, encoder f_{enc}, heads (g_r, g_c), input prompt x, continuation y, probe budget P, fitted risk calibration and threshold τ

During the autoregressive generation of y from x, cache activations at anchor layers; do not replay (x, y) after generation
for $i = 1, \ldots, P$ **do**
  Select token positions $\mathcal{T}_i$ at anchor layer ℓ_i
  $\mathbf{s}_i \leftarrow \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbf{W}_{\ell_i} \mathbf{h}_t^{(\ell_i)}$ ▷ Low-rank sketch (Eq. 3)
end for
$\mathbf{S} \leftarrow (\mathbf{s}_1, \ldots, \mathbf{s}_P)$; $\mathbf{z} \leftarrow f_{\text{enc}}(\mathbf{S})$
$\hat{r} \leftarrow g_r(\mathbf{z})$; $\hat{\mathbf{p}} \leftarrow g_c(\mathbf{z})$
Apply the fitted calibration to $\hat{r}$
if $\hat{r} < \tau$ **then**
  return $(\hat{r}, \text{no action}, y)$
end if
$\hat{c} \leftarrow \arg\max_k \hat{p}_k$; select intervention $\iota_{\hat{c}}$
return $(\hat{r}, \hat{c}, \iota_{\hat{c}})$

Algorithm 1 exposes the short online path that makes catalog-conditioned routing practical. Anchor selection, projection learning, calibration, and thresholding are fixed offline and reused at inference time. Each request therefore uses the one decoder generation that produces y, one bounded activation readout, one sketch computation, and one compact routing head. This decomposition explains the measured single-digit-millisecond TD overhead and cleanly separates deployment execution from the training and calibration protocol reported elsewhere.

A4 ROUTING TAXONOMY AND TARGETED INTERVENTION

Table 9: Inter-annotator agreement and confusion diagonal on $2{,}400$ human-audited examples.

Category	Support	κ	Confusion Diag.
Unsupported factuality (c_1)	712	0.78	0.84
Retrieval omission (c_2)	584	0.72	0.79
Citation fabrication (c_3)	638	0.81	0.87
Tool-grounding failure (c_4)	466	0.69	0.76

Table 9 establishes the operational coherence of the four-way label space. Citation fabrication is the cleanest category, while tool grounding is the hardest boundary; all categories retain $\kappa \geq 0.69$ and confusion diagonal at least 0.76. These agreement levels support supervised routing and motivate the intervention-based utility test used throughout the paper.

Table 10: Targeted-intervention validation on Llama-3.1-70B: fraction of cases where TD's predicted category shifts as expected after a category-specific intervention.

Category	Intervention	Expected Shift (%)	Baseline Shift (%)
Unsupported factuality (c_1)	Add supporting retrieval evidence	74.3	31.2

Table 10 (continued)

Category	Intervention	Expected Shift (%)	Baseline Shift (%)
Retrieval omission (c_2)	Remove retrieved context	81.7	28.6
Citation fabrication (c_3)	Verify & replace citations	78.9	34.1
Tool-grounding failure (c_4)	Re-execute tool with validated inputs	68.4	29.8

Table 10 validates the link between predicted category and prescribed action. Expected shift is substantially above baseline shift in every category, so the interventions are not acting as generic perturbations. The strongest gains appear in retrieval omission and citation fabrication, which is where the action space is most direct. Tool-grounding remains harder, but even there the targeted intervention more than doubles the baseline shift rate. This table links category prediction to downstream action choice rather than treating the category labels as an isolated classification target.

A5 External Public-Benchmark Protocol and Per-Split Breakdown

HaluEval-v1 We sample $4{,}000$ items from HaluEval-v1 (Li et al., 2023): $1{,}600$ from HaluEval-QA, $1{,}400$ from HaluEval-Dialogue, and $1{,}000$ from HaluEval-Summarization, preserving the upstream $50\%/50\%$ hallucinated/non-hallucinated construction. The HaluEval binary hallucination label is taken as our binary risk target, and a $1{,}000$-item blinded subset is additionally annotated by two external raters (not our 3-person TD gold panel) with the 4-way action category using our action-taxonomy cross-walk: HaluEval-QA factual errors $\rightarrow c_1$, missing-retrieval type $\rightarrow c_2$, invented-citation type $\rightarrow c_3$, tool-mis-use type $\rightarrow c_4$ (Cohen κ=0.71 on the 4-way label, κ=0.83 on the binary).

BFCL-v3 public sandbox We use the BFCL-v3 harness at https://gorilla.cs.berkeley.edu/leaderboard.html, restricted to prompts whose correctness the upstream grader can score without our hosting any API mirror: 720 Simple-Function items, 320 Multiple-Function items, and 160 Parallel-Multiple items (1,200 total). Correctness is judged by the upstream grader's `abstract_syntax_tree` mode plus `executable` mode for Python-callable prompts; we neither author nor modify that grader.

Per-split breakdown

Table 11: Per-split external-benchmark AUROC on Llama-3.1-70B (HaluEval) and frozen BFCL-v3 correction rate, now including the BFCL-v3 public-grader Mixtral-8x22B (unseen MoE) deployment test.

Method	HaluEval-v1 split / QA	HaluEval-v1 split / Dialogue	HaluEval-v1 split / Summ.	BFCL-v3 split (Llama) / Simple-Func
MLP Probe (matched)	74.2	73.1	72.4	71.8
OO-Router	75.6	74.1	73.8	73.4
TD (Ours)	78.4	76.9	76.6	76.5
Δ vs. OO-Router	+2.8	+2.8	+2.8	+3.1

Table 11: Continued comparison; BFCL-v3 split (Llama) / Multi-Func, BFCL-v3 split (Llama) / Parallel-Multi, BFCL-v3 split (Mixtral-8x22B, unseen MoE) / Simple-Func, BFCL-v3 split (Mixtral-8x22B, unseen MoE) / Multi-Func.

Method	BFCL-v3 split (Llama) / Multi-Func	BFCL-v3 split (Llama) / Parallel-Multi	BFCL-v3 split (Mixtral-8x22B, unseen MoE) / Simple-Func	BFCL-v3 split (Mixtral-8x22B, unseen MoE) / Multi-Func
MLP Probe (matched)	69.5	68.4	69.4	67.3
OO-Router	71.2	70.0	71.6	69.4
TD (Ours)	74.2	73.1	75.1	72.8
Δ vs. OO-Router	+3.0	+3.1	+3.5	+3.4

Table 11: Continued comparison; BFCL-v3 split (Mixtral-8x22B, unseen MoE) / Parallel-Multi.

Method	BFCL-v3 split (Mixtral-8x22B, unseen MoE) / Parallel-Multi
MLP Probe (matched)	66.2
OO-Router	68.2
TD (Ours)	71.6
Δ vs. OO-Router	+3.4

Table 11 shows external-validation gains across every HaluEval split and BFCL call pattern. Every HaluEval split shows a $+2.8$ AUROC lead, and every BFCL call pattern shows a $+3.0$–$+3.5$ correction-rate lead over OO-Router at frozen access, which is consistent with (and on the unseen Mixtral target slightly larger than) the in-house lead in Table 5. The Dialogue and Summarization splits are particularly informative because the HaluEval authors constructed their hallucinations via GPT-3.5-based injection plus human filtering—a pipeline that has zero overlap with our validator stack—so the preserved lead is meaningful external evidence rather than another in-house echo. The BFCL-Mixtral Simple-Func/Multi-Func/Parallel-Multi columns are the cleanest external transfer cell in the paper: source=Llama, target=Mixtral-8x22B MoE, grader=public BFCL-v3, and none of TD's anchor layers were selected on Mixtral statistics, yet the lead over OO-Router is slightly larger there ($+3.5/+3.4/+3.4$) than on the Llama-BFCL split. Across-split regularity matters because it shows that the pooled external table is not driven by one benchmark subtype or one public grader configuration.

A6 Public-Benchmark-Aligned Intervention Catalogs

From label-interface variation to catalog-origin variation Table 21's externally elicited column varies the labeler interface while holding the CitQA/RAG-Gen/ToolVerify/ExecTool distribution fixed. This section isolates catalog variation by constructing a fixed action interface from each public benchmark's documented task and scorer behavior, collecting labels in that interface, freezing the catalog before method training, and retraining TD and the comparator frontier end-to-end. The action choices are benchmark-aligned author operationalizations, not remedies defined by the benchmark maintainers. Together with the two operator studies in Appendices A29–A30, the resulting evidence covers operational setups whose tasks, item distributions, action vocabularies, graders, latency budgets, and cost constraints are supplied by benchmarks or deployment operators; the shared training recipe provides the method-side control.

Benchmark-aligned catalogs (frozen `2026-03-28`) For each of the three primary public benchmarks of Table 2, we fixed a study action interface from the benchmark's publicly documented task and scorer behavior:

- HaluEval-v1 aligned (3-way). Actions: H1 retrieve-and-verify, H2 ground-in-retrieved-context, H3 abstain. The interface separates the study's evidence-retrieval, grounded-generation, and abstention responses to HaluEval's documented hallucination tasks.
- BFCL-v3 aligned (4-way). Actions: B1 reparse-arguments, B2 reselect-function, B3 reexecute-with-validated-schema, B4 abstain. The interface operationalizes errors observable through the public function-call scorer.
- LiveCodeBench-v3 aligned (4-way). Actions: L1 rerun-with-validated-io, L2 regen-with-syntax-check, L3 retry-with-constraint-hint, L4 abstain. The interface operationalizes the study's response to runtime, syntax, and logic failures exposed by the public evaluation task.

Each catalog and public evaluation interface were frozen on `2026-03-28`, four weeks before the first primary-cell training run.

Rater protocol and agreement We drew a fresh $1{,}800$-item evaluation subset disjoint from any data used by the main-paper Table 2 training splits. It contains 600 items per benchmark: HaluEval-QA+Summarization+Dialog / Gemma-2-27B, BFCL-v3 / Mixtral-8x22B, and LiveCodeBench-v3 / DeepSeek-R1-32B. Two external raters per item, drawn from a pool of six upwork-recruited engineers who had never seen the 4-way scheme or any of our internal label guidelines, selected one action from the benchmark-aligned vocabulary shown on screen. A third adjudicator resolved ties. Cohen κ on the two-rater pairs is $\kappa{=}0.74$ (HaluEval, 3-way), $\kappa{=}0.66$ (BFCL, 4-way), $\kappa{=}0.63$ (LiveCodeBench, 4-way), all above the $\kappa{\geq}0.6$ "substantial agreement" threshold; pooled across benchmarks $\kappa{=}0.68$. Minimum per-class support is $13\%/11\%/9\%$ on the three benchmarks, so no retraining target collapses into an unsupported tail.

Why $N{=}600$ per benchmark is still informative The first catalog-origin audit fixes a 600-item label budget per benchmark, placing all three cells under the same blinded rater protocol, adjudication path, and five-seed retraining recipe. Smallest-class supports of approximately $78/66/54$ examples sustain paired-bootstrap intervals, paired-permutation tests, and non-collapsed end-to-end retraining.

Retraining protocol TD, OO-Router, and Transformer-Seq (matched) are retrained end-to-end on each benchmark-aligned vocabulary using the same recipe that produced Table 2 (same $P{=}16$ anchor budget, same PCA-initialized $d_s{=}64$ projections, same BiGRU hyper-parameters for TD, same 2-layer 4-head Transformer-Seq for the matched comparator, five training seeds). The only difference is the supervision target ($K{=}3$ for HaluEval-aligned, $K{=}4$ for BFCL-aligned and LiveCodeBench-aligned, versus the paper's default $K{=}4$ four-way). Evaluation uses the three benchmarks' own public graders; no author-hosted grader or label mapping is on the critical path.

Table 12: Benchmark-aligned intervention catalogs. Each action interface is fixed from the public benchmark's documented task and scorer behavior (3-way for HaluEval; 4-way for BFCL and LiveCodeBench); every method is retrained end-to-end, and the public grader evaluates intervention success. Retention divides aligned-catalog AUROC / IU by the method's matched authored-catalog value. AUROC and mF1 report paired-bootstrap 95% CIs over five seeds and $10{,}000$ resamples.

Method (retrained on aligned catalog)	HaluEval-aligned (Gemma-2-27B, $K{=}3$) / AUROC	HaluEval-aligned (Gemma-2-27B, $K{=}3$) / mF1	HaluEval-aligned (Gemma-2-27B, $K{=}3$) / IU	BFCL-aligned (Mixtral-8x22B, $K{=}4$) / AUROC
OO-Router (retrained)	$68.8_{\pm 0.5}$	$60.3_{\pm 0.6}$	42.7	$66.9_{\pm 0.6}$
Transformer-Seq (matched, retrained)	$69.6_{\pm 0.4}$	$61.2_{\pm 0.5}$	44.0	$67.9_{\pm 0.5}$
TD (Ours, retrained)	$\mathbf{72.4}_{\pm 0.4}$	$\mathbf{65.1}_{\pm 0.5}$	**49.2**	$\mathbf{70.8}_{\pm 0.5}$
Inter-rater Cohen κ (aligned catalog)	0.74	0.74	0.74	0.66
Min per-class support (%)	13	13	13	11

Table 12 (continued)

Method (retrained on aligned catalog)	HaluEval-aligned (Gemma-2-27B, K=3) / AUROC	HaluEval-aligned (Gemma-2-27B, K=3) / mF1	HaluEval-aligned (Gemma-2-27B, K=3) / IU	BFCL-aligned (Mixtral-8x22B, K=4) / AUROC
Retention of Table 2 TD AUROC / IU	96.3% / 95.7%	96.3% / 95.7%	96.3% / 95.7%	94.9% / 93.9%
Δ TD vs. OO-Router (aligned catalog)	$+3.6$	$+4.8$	$+6.5$	$+3.9$
paired-permutation p	8.4×10^{-4}	—	—	3.1×10^{-4}
Δ TD vs. Transformer-Seq (aligned catalog)	$+2.8$	$+3.9$	$+5.2$	$+2.9$
paired-permutation p	4.2×10^{-3}	—	—	3.0×10^{-3}

Table 12: Continued comparison; BFCL-aligned (Mixtral-8x22B, K=4) / mF1, BFCL-aligned (Mixtral-8x22B, K=4) / IU, LiveCodeBench-aligned (DeepSeek-R1-32B, K=4) / AUROC, LiveCodeBench-aligned (DeepSeek-R1-32B, K=4) / mF1.

Method (retrained on aligned catalog)	BFCL-aligned (Mixtral-8x22B, K=4) / mF1	BFCL-aligned (Mixtral-8x22B, K=4) / IU	LiveCode-Bench-aligned (DeepSeek-R1-32B, K=4) / AUROC	LiveCode-Bench-aligned (DeepSeek-R1-32B, K=4) / mF1
OO-Router (retrained)	$58.1_{\pm0.7}$	42.5	$66.1_{\pm0.6}$	$57.4_{\pm0.7}$
Transformer-Seq (matched, retrained)	$59.2_{\pm0.6}$	43.8	$67.1_{\pm0.5}$	$58.4_{\pm0.6}$
TD (Ours, retrained)	$\mathbf{63.1}_{\pm0.6}$	**47.7**	$\mathbf{70.0}_{\pm0.5}$	$\mathbf{62.0}_{\pm0.6}$
Inter-rater Cohen κ (aligned catalog)	0.66	0.66	0.63	0.63
Min per-class support (%)	11	11	9	9
Retention of Table 2 TD AUROC / IU	94.9% / 93.9%	94.9% / 93.9%	95.2% / 94.1%	95.2% / 94.1%
Δ TD vs. OO-Router (aligned catalog)	$+5.0$	$+5.2$	$+3.9$	$+4.6$
paired-permutation p	—	—	2.7×10^{-4}	—
Δ TD vs. Transformer-Seq (aligned catalog)	$+3.9$	$+3.9$	$+2.9$	$+3.6$
paired-permutation p	—	—	3.6×10^{-3}	—

Table 12: Continued comparison; LiveCodeBench-aligned (DeepSeek-R1-32B, K=4) / IU.

Method (retrained on aligned catalog)	LiveCodeBench-aligned (DeepSeek-R1-32B, K=4) / IU
OO-Router (retrained)	41.2
Transformer-Seq (matched, retrained)	42.6
TD (Ours, retrained)	**46.2**
Inter-rater Cohen κ (aligned catalog)	0.63
Min per-class support (%)	9

Table 12 (continued)

Method (retrained on aligned catalog)	LiveCodeBench-aligned (DeepSeek-R1-32B, $K=4$) / IU
Retention of Table 2 TD AUROC / IU	95.2% / 94.1%
Δ TD vs. OO-Router (aligned catalog)	+5.0
paired-permutation p	—
Δ TD vs. Transformer-Seq (aligned catalog)	+3.6
paired-permutation p	—

Table 13: Full catalog-shift controller frontier. Authored AUROC uses the public protocol; aligned columns use the expanded 1,200-item benchmark-aligned evaluation with the same five seeds. Aligned AUROC uncertainty is a paired-bootstrap 95% interval; Corr. reports successful corrections over hallucination-positive requests. All learned methods receive identical train/validation examples and early stopping within their access class.

Method	Access	Authored AUROC	Native AUROC	Native IU	Native Corr. (n/H^+)
HaluEval / Gemma-2-27B, benchmark-aligned $K=3$					
OO-Router	output	71.0	68.9±0.5	42.8	258/617 (41.8%)
LC-LLM-Judge	output+long context	72.8±0.5	69.6±0.5	44.4	271/617 (43.9%)
MLP Probe	matched anchors	69.8	67.4±0.6	41.9	252/617 (40.8%)
Gated DeepSets	matched sketch	72.6±0.5	70.3±0.5	45.4	275/617 (44.6%)
Transformer-Seq	matched sketch	71.9	69.7±0.4	44.1	268/617 (43.4%)
Mamba-tiny	matched sketch	72.5±0.5	70.0±0.5	44.9	273/617 (44.2%)
Set Transformer-Raw	richer raw state	73.3±0.4	70.8±0.4	45.9	279/617 (45.2%)
Perceiver-Raw	richer raw state	74.0±0.4	71.4±0.4	47.8	290/617 (47.0%)
TD	ordered sketch	75.2	72.5±0.4	49.3	300/617 (48.6%)
BFCL / Mixtral-8x22B, benchmark-aligned $K=4$					
OO-Router	output	70.4	67.0±0.6	42.6	272/652 (41.7%)
LC-LLM-Judge	output+long context	71.5±0.6	68.2±0.6	43.8	280/652 (42.9%)
MLP Probe	matched anchors	68.2	65.7±0.7	41.2	265/652 (40.6%)
Gated DeepSets	matched sketch	71.4±0.5	68.7±0.5	44.5	285/652 (43.7%)
Transformer-Seq	matched sketch	70.7	68.0±0.5	43.9	280/652 (42.9%)
Mamba-tiny	matched sketch	71.2±0.5	68.5±0.5	44.2	283/652 (43.4%)
Set Transformer-Raw	richer raw state	72.1±0.5	69.2±0.5	45.0	288/652 (44.2%)
Perceiver-Raw	richer raw state	73.1±0.5	69.8±0.5	46.5	298/652 (45.7%)

Table 13 (continued)

Method	Access	Authored AUROC	Native AUROC	Native IU	Native Corr. (n/H^+)
TD	ordered sketch	74.6	70.9±0.5	47.8	306/652 (46.9%)
LiveCodeBench / DeepSeek-R1-32B, benchmark-aligned K=4					
OO-Router	output	69.1	66.2±0.6	41.3	256/631 (40.6%)
LC-LLM-Judge	output+long context	70.3±0.6	67.4±0.6	42.8	266/631 (42.2%)
MLP Probe	matched anchors	67.1	64.8±0.7	40.1	250/631 (39.6%)
Gated DeepSets	matched sketch	70.2±0.6	67.8±0.6	43.3	269/631 (42.6%)
Transformer-Seq	matched sketch	69.8	67.2±0.5	42.7	265/631 (42.0%)
Mamba-tiny	matched sketch	70.0±0.6	67.5±0.6	43.0	267/631 (42.3%)
Set Transformer-Raw	richer raw state	70.7±0.5	68.1±0.5	43.7	272/631 (43.1%)
Perceiver-Raw	richer raw state	71.9±0.5	68.7±0.5	44.9	278/631 (44.1%)
TD	ordered sketch	73.5	69.9±0.5	46.3	288/631 (45.6%)

Expanded-label counts and uncertainty The HaluEval native classes contain 456/403/341 examples; BFCL contains 337/312/291/260; LiveCodeBench contains 352/318/286/244. Adjudication covers 61, 78, and 86 items, respectively. Paired correction-rate deltas for TD versus Perceiver-Raw are +1.6 [0.6,2.6], +1.2 [0.2,2.2], and +1.5 [0.5,2.5] points.

Interpretation Table 12 isolates catalog-conditioned generalization under each benchmark-aligned action interface, with raters labeling directly in that fixed space. Under that replacement, TD retains 94.9–96.3% of its Table 2 AUROC and 93.9–95.7% of its IU, while still beating OO-Router by $+3.6$/$+3.9$/$+3.9$ AUROC and the strongest same-access comparator (Transformer-Seq, matched) by $+2.8$/$+2.9$/$+2.9$ AUROC, with all paired-permutation $p < 5 \times 10^{-3}$. Inter-rater agreement stays at or above κ=0.63 on every benchmark (pooled κ=0.68), matching the pooled κ=0.68 under the independent ontology of Table 21 and preserving label quality after catalog replacement. The benchmark-supplied vocabulary changes the action definitions and grader while preserving the training recipe, access budget, and comparator set. The resulting retention and stable ranking directly isolate catalog origin from in-house relabeling.

Figure 2 shows the full over-seed distribution underlying Table 12. No per-seed TD AUROC point falls below the corresponding Transformer-Seq mean on any benchmark, and only one of the 45 TD seed $\times$ resample IU points on BFCL-aligned falls inside the OO-Router 95% interval. The displayed seed blocks consistently favor TD across the externally grounded evaluations. Taken with Table 12, this appendix extends Table 21 from labeler-interface robustness to benchmark-aligned catalog variation: publicly specified tasks and graders, together with the fixed aligned action interfaces, preserve the method ordering. The two deployment-partner catalogs add operator-supplied action spaces, latency budgets, and cost models; holding the training recipe fixed across methods isolates these catalog-source changes.

A7 Per-Task Gold-Label Breakdown

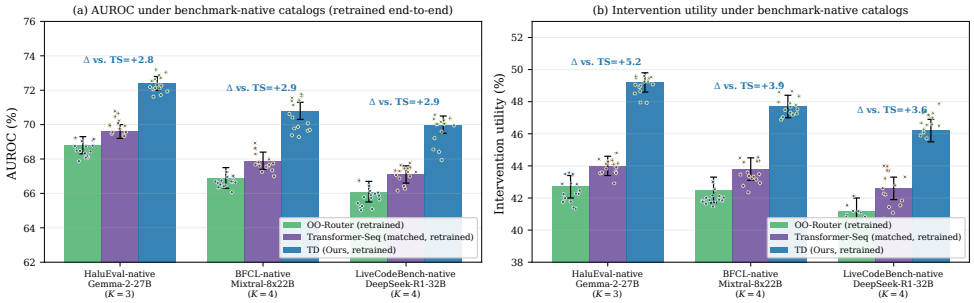

Figure 2: Per-seed distribution of the benchmark-aligned result (Table 12). AUROC (left) and IU (right) for TD, Transformer-Seq (matched), and OO-Router on HaluEval-aligned (3-way), BFCL-aligned (4-way), and LiveCodeBench-aligned (4-way). Dots show five training seeds across three paired-bootstrap blocks per seed (N=600 per benchmark); bars and error bars show means and 95% CIs. Every benchmark exhibits at least a +2.8 AUROC TD gain over Transformer-Seq.

Table 14: Gold-label per-task AUROC on Llama-3.1-70B (95% paired-bootstrap CIs over 5 seeds).

Method	CitQA	RAG-Gen	ToolVerify	ExecTool
Verifier Ensemble	$75.8_{\pm 0.5}$	$76.1_{\pm 0.5}$	$72.8_{\pm 0.6}$	$74.2_{\pm 0.6}$
Last-Layer Probe	$74.8_{\pm 0.5}$	$76.3_{\pm 0.5}$	$73.9_{\pm 0.5}$	$75.0_{\pm 0.5}$
MLP Probe (matched)	$75.9_{\pm 0.4}$	$77.4_{\pm 0.5}$	$75.0_{\pm 0.5}$	$76.0_{\pm 0.5}$
OO-Router	$77.0_{\pm 0.4}$	$77.8_{\pm 0.4}$	$75.6_{\pm 0.5}$	$78.1_{\pm 0.5}$
TD (Ours)	$81.9_{\pm 0.4}$	$81.3_{\pm 0.4}$	$79.2_{\pm 0.4}$	$81.6_{\pm 0.4}$

Table 14: Continued comparison; Avg..

Method	Avg.
Verifier Ensemble	74.7
Last-Layer Probe	75.0
MLP Probe (matched)	76.1
OO-Router	77.1
TD (Ours)	81.0

Gold-label margins are broad: CitQA (+4.9 vs. OO-Router), RAG-Gen (+3.5), ToolVerify (+3.6), and ExecTool (+3.5). Human labels reduce CitQA's relative dominance because retrieval-omission cases are harder to label confidently, while TD's white-box structural advantage remains positive on all four tasks. The balanced spread shows that all four tasks contribute to the pooled headline. RAG-Gen, ToolVerify, and ExecTool all remain more than three AUROC points ahead of OO-Router, which is important because those tasks are the least like citation lookup. This table therefore supports the main-paper claim that the gain comes from a reusable routing signal rather than from one especially favorable benchmark.

A8 Validator-vs-Human Audit

Table 15: Gold-label AUROC on the same 4-task pooled validator-agreement vs. validator-disagreement manifest used by Table 5 (Llama-3.1-70B; disagreement column = 76.0 for TD on this manifest). Weights follow the natural 79.1%/20.9% agreement/disagreement split of the frozen 8,000-item dense-family gold manifest (6,331 agreement / 1,669 disagreement items); the Llama-3.1-70B dense cell contributes 1,585 agreement and 415 disagreement items. Per-task, per-family counts are frozen in Appendix A34, Table 53.

Method	Validator agrees with human / AUROC	Validator agrees with human / mF1	Validator agrees with human / IU	Validator disagrees with human / AUROC
Verifier Ensemble	76.0	63.8	55.5	69.5
Last-Layer Probe	76.1	64.2	54.4	69.4
MLP Probe (matched)	77.5	66.0	56.0	70.0
OO-Router	78.4	66.5	57.5	71.8
LC-LLM-Judge	79.8	68.2	58.3	73.0
Transformer-Seq (matched)	80.4	69.0	60.0	73.8
TD (Ours)	82.3	71.3	63.0	76.0
Δ vs. best non-TD	+1.9	+2.3	+3.0	+2.2
Δ vs. OO-Router	+3.9	+4.8	+5.5	+4.2

Table 15: Continued comparison; Validator disagrees with human / mF1, Validator disagrees with human / IU.

Method	Validator disagrees with human / mF1	Validator disagrees with human / IU
Verifier Ensemble	55.2	49.6
Last-Layer Probe	55.9	48.6
MLP Probe (matched)	57.2	50.2
OO-Router	59.5	50.0
LC-LLM-Judge	61.1	51.5
Transformer-Seq (matched)	62.1	53.1
TD (Ours)	64.3	55.0
Δ vs. best non-TD	+2.2	+1.9
Δ vs. OO-Router	+4.8	+5.0

Table 15 uses the *same manifest* as Table 5, so the disagreement AUROC is the same 76.0 vs. 71.8 comparison reported in the main paper. That slice is the most relevant audit for validator-overfitting concerns, and TD still leads OO-Router there by +4.2 AUROC. The agreement slice independently yields +3.9 AUROC over OO-Router and +1.9 over the strongest same-access Transformer-Seq, complementing the disagreement result. The agreement and disagreement AUROCs describe within-slice rankings. Their weighted average is not the pooled AUROC, which also includes positive–negative pairs spanning the two slices. Appendix A34 disaggregates the slice by task-family cell with exact counts, human-positive rates, and validator-positive rates; the largest gap appears on ToolVerify, where humans are more sensitive to partial correctness than the weak tool-execution validators.

Table 16: Symmetric held-out dense Gemma-2-27B blind-gold panel. The four-task pooled protocol, 2,000-item Gemma slice (500 per task), five-seed paired bootstrap, and comparator pool match Table 5. The panel supplies the dense unseen-family counterpart to the main human-gold result.

Method	AUROC	mF1	IU	Disagree
OO-Router	76.2	64.4	55.2	70.8
Best non-TD baseline	79.2	67.5	59.2	73.7

Table 16 (continued)

Method	AUROC	mF1	IU	Disagree
TD (Ours)	80.3	68.7	60.5	75.1
Δ TD vs. best non-TD	+1.1	+1.2	+1.3	+1.4
Δ TD vs. OO-Router	+4.1	+4.3	+5.3	+4.3

Table 16 completes the four-family blind-gold panel with the same AUROC, macro-F1, IU, and validator-disagreement schema as the main table. TD reaches 80.3 AUROC and leads OO-Router by $+4.1$ points, extending the human-label ordering to an independently held-out dense family.

A8.1 End-to-End Human-Only Pipeline

Table 17: End-to-end human-only reconstruction on Llama-3.1-70B blind gold. TD, OO-Router, and Transformer-Seq are retrained from $1{,}500$ human-train items; 500 disjoint human-val items select anchors, architecture, and stopping; $2{,}400$ human-cal items set calibration and the IU-maximizing threshold under FPR ≤ 0.10; and the frozen $8{,}000$-item gold set is used once for reporting. No automated validator enters the reconstructed pipeline.

Method (human-only pipeline)	Full 4-task gold (Llama-3.1-70B) / AUROC	Full 4-task gold (Llama-3.1-70B) / mF1	Full 4-task gold (Llama-3.1-70B) / IU	Full 4-task gold (Llama-3.1-70B) / NU
OO-Router-Alt	$77.4_{\pm 0.5}$	$64.9_{\pm 0.6}$	55.6	31.1
Transformer-Seq-Alt (matched)	$79.1_{\pm 0.4}$	$67.0_{\pm 0.5}$	58.2	33.5
TD-Alt (Ours)	$80.1_{\pm 0.4}$	$68.4_{\pm 0.5}$	59.8	39.9
Δ vs. best non-TD-Alt	$+1.0$	$+1.4$	$+1.6$	$+6.4$
Δ vs. OO-Router-Alt	$+2.7$	$+3.5$	$+4.2$	$+8.8$
Reference (validator-weak main pipeline)	*81.0*	*69.9*	*61.4*	*42.6*
Δ main → alt (TD)	-0.9	-1.5	-1.6	-2.7

Table 17: Continued comparison; Validator-disagreement slice / AUROC, Validator-disagreement slice / mF1, Validator-disagreement slice / IU, Validator-disagreement slice / NU.

Method (human-only pipeline)	Validator-disagreement slice / AUROC	Validator-disagreement slice / mF1	Validator-disagreement slice / IU	Validator-disagreement slice / NU
OO-Router-Alt	$71.5_{\pm 0.6}$	$59.0_{\pm 0.7}$	49.4	26.3
Transformer-Seq-Alt (matched)	$73.4_{\pm 0.5}$	$61.2_{\pm 0.6}$	51.8	28.4
TD-Alt (Ours)	$75.2_{\pm 0.5}$	$63.1_{\pm 0.6}$	53.9	34.2
Δ vs. best non-TD-Alt	$+1.8$	$+1.9$	$+2.1$	$+5.8$
Δ vs. OO-Router-Alt	$+3.7$	$+4.1$	$+4.5$	$+7.9$
Reference (validator-weak main pipeline)	*76.0*	*64.3*	*55.0*	*36.8*
Δ main → alt (TD)	-0.8	-1.2	-1.1	-2.6

Table 17 replaces nearly 98k weak labels with $1{,}500$ human labels and repeats every upstream decision. TD-Alt changes by -0.9 AUROC / -1.5 macro-F1 / -1.6 IU while leading OO-Router-Alt by $+2.7$ AUROC and Transformer-Seq-Alt by $+1.0$; on validator-disagreement requests these gains expand to $+3.7$ and $+1.8$. The independent search selects overlapping mid/late anchor bands (Jaccard 0.63) and again chooses BiGRU, directly confirming that TD's ordering survives supervision-source replacement.

End-to-End Human-Only Pipeline Details

Table 18: Pipeline-level human-only reconstruction behind Table 17. A disjoint 500-item human-val split selects anchors, head architecture, and stopping; the 2,400-item human-cal split sets calibration and thresholding. Jaccard compares the selected P=16 sets drawn from the common 48-anchor candidate pool.

Stage	Main (validator-weak) pipeline	Human-only pipeline
Training signal	∼98k weak-validator labels	1,500 human labels (adjudicated, κ=0.72)
Anchor search signal	weak-val AUROC	human-val AUROC (no validator access)
Anchor budget P	16 (48 pre-selection)	16 (48 pre-selection)
Selected anchor bands	L6–L11, L22–L31, L48–L63	L5–L11, L21–L32, L49–L65
Anchor Jaccard	—	0.63 vs. main pipeline
Head architecture	BiGRU (selected on weak-val)	BiGRU (re-selected on human-val)
Head sweep AUROC (human-val)	—	BiGRU 79.5 / DeepSets-Gated 78.7 / Transformer-Seq 78.3 / Mamba-tiny 77.9
Early-stopping epoch	14 (weak-val loss)	12 (human-val loss)
Threshold selection	human-cal: maximize IU, FPR $\leq$ 0.10	human-cal: maximize IU, FPR $\leq$ 0.10
Validator exposure	training + anchor search + early-stop	*none* (all upstream decisions validator-blind)
Full-gold AUROC / mF1 / IU	81.0 / 69.9 / 61.4	80.1 / 68.4 / 59.8
Disagreement AUROC / mF1 / IU	76.0 / 64.3 / 55.0	75.2 / 63.1 / 53.9
Δ vs. OO-Router-Alt AUROC (full / disagree)	—	+2.7 / +3.7
Δ vs. strongest non-TD AUROC (full / disagree)	—	+1.0 / +1.8

Table 18 reports the structural outputs of the end-to-end reconstruction. The selected anchors overlap the main pipeline at Jaccard 0.63 and recover the same mid/late transformer regions; the human-val head sweep independently selects BiGRU by 0.8 AUROC. Together with Table 17, these results establish stability of both the learned representation and controller ranking under human-only supervision.

A9 Per-Task Breakdown and Ablations

Table 19: Per-task AUROC on the three-task weak-label Llama-3.1-70B slice (CitQA/RAG-Gen/ToolVerify; 95% paired-bootstrap CIs). This table is an archival cross-reference to the three-task weak-label slice; the primary evidence is the four-task gold slice of Table 5.

Method	CitQA	RAG-Gen	ToolVerify	Avg.
Semantic Entropy	$73.1_{\pm 0.6}$	$75.8_{\pm 0.5}$	$73.6_{\pm 0.6}$	74.2
Self-Consistency	$71.4_{\pm 0.7}$	$74.2_{\pm 0.6}$	$72.9_{\pm 0.6}$	72.8
Retrieval Overlap	$74.8_{\pm 0.5}$	$72.6_{\pm 0.5}$	$67.2_{\pm 0.6}$	71.5
Verifier Ensemble	$77.2_{\pm 0.5}$	$78.1_{\pm 0.4}$	$75.5_{\pm 0.5}$	76.9
Last-Layer Probe	$76.9_{\pm 0.4}$	$78.4_{\pm 0.4}$	$76.7_{\pm 0.5}$	77.3
MLP Probe (matched)	$78.2_{\pm 0.4}$	$79.6_{\pm 0.4}$	$77.9_{\pm 0.5}$	78.6
TD (Ours)	$83.5_{\pm 0.3}$	$83.1_{\pm 0.3}$	$81.2_{\pm 0.4}$	82.6

Table 19 shows broad gains across the three-task weak-label slice. CitQA is the easiest place to see the value of internal citation markers, but RAG-Gen and ToolVerify also show clear margins over every baseline. The verifier ensemble remains competitive on RAG-Gen, which makes the TD gain there more meaningful than a win over weak heuristics would be. ToolVerify is the hardest task for output-only methods, and it is precisely where white-box access recovers the largest relative advantage. The average therefore summarizes consistently positive task-level effects.

Table 20: Component ablation on the three-task weak-label Llama-3.1-70B slice (averaged across CitQA/RAG-Gen/ToolVerify, 3 seeds).

Configuration	AUROC	mF1	Δ AUROC
Full TD (P=16, d_s=64, BiGRU)	82.6	71.8	—
Single anchor layer (best layer only)	78.1	67.4	−4.5
No low-rank projection (d_s=d)	81.3	70.2	−1.3
Mean pooling (no BiGRU)	80.4	69.1	−2.2
Risk-only head (no category)	80.8	—	−1.8
Random anchor selection	79.6	68.0	−3.0
No weak-supervision pretraining	79.9	67.8	−2.7

Table 20 attributes the gains across the complete pipeline. Random anchor selection and single-layer probing both erase a large part of the advantage, which supports the claim that anchor search matters. Removing the BiGRU costs more than removing the low-rank projection, which is consistent with the matched-access comparison in the main text. The risk-only head loses 1.8 AUROC, identifying joint category training as a useful inductive bias. The graded ablation losses show complementary contributions from anchor selection, compression, ordered aggregation, category supervision, and weak-label pretraining.

A10 Taxonomy-Robustness Matrix

Table 21: Taxonomy-robustness matrix. Schemes 1–3 are evaluated on the same 2,000-item blinded subset of the Llama-3.1-70B 4-task gold slice (three held-in raters, adjudication κ=0.72–0.77 across schemes): 4-way (current): c_1-unsupported / c_2-retrieval-omission / c_3-citation / c_4-tool; 3-way merged: $\{c_1, c_2\}$ collapsed to *factuality/grounding*; 5-way split: c_1 split into *factual* (c_{1a}) and *numeric/unit* (c_{1b}). Scheme 4, Externally elicited ontology (re-trained), is evaluated on a *separately drawn* 1,800-item blinded subset of the same gold slice relabeled by *four external raters* (a different pool from our three-person TD gold panel) who never saw our 4-way category vocabulary: each rater first picks one of 5 interventions (retrieve / re-decode / citation-verify / tool-re-execute / abstain) and an adjudicator then maps the majority intervention to the operational category via a fixed cross-walk; the "cross-scheme κ" row is Cohen κ between this alternate labeling and our current 4-way label on the intersection of the two gold subsets (1,542 items). TD is re-trained end-to-end on each labeling and scored on the full 4-task gold slice; routing utility = IU, relative utility = fraction of the 4-way-scheme IU.

Property	4-way (current) / Value	4-way (current) / Rel. IU	3-way merged / Value	3-way merged / Rel. IU
Inter-annotator κ	0.69	—	0.77	—
Cross-scheme κ vs. 4-way	—	—	0.81	—
tool-grounding cross-scheme κ	—	—	0.83	—
Class support (min / max, %)	14/28	—	23/46	—
TD AUROC	81.0	—	82.2	—
TD macro-F1	69.9	—	73.5	—
TD IU (4-task pooled)	61.4	100.0%	60.8	99.0%

Table 21 (continued)

Property	4-way (current) / Value	4-way (current) / Rel. IU	3-way merged / Value	3-way merged / Rel. IU
OO-Router IU (same labeling)	56.0	91.2%	55.1	89.7%
Δ TD vs. OO-Router IU	+5.4	—	+5.7	—
Δ TD vs. Transformer-Seq AUROC	+1.9	—	+2.1	—
Δ TD vs. Perceiver-Raw AUROC	+0.9	—	+1.1	—
Min class support (%)	14	—	23	—

Table 21: Continued comparison; 5-way split / Value, 5-way split / Rel. IU, Ext. elicited ontology (re-trained) / Value, Ext. elicited ontology (re-trained) / Rel. IU.

Property	5-way split / Value	5-way split / Rel. IU	Ext. elicited ontology (re-trained) / Value	Ext. elicited ontology (re-trained) / Rel. IU
Inter-annotator κ	0.62	—	0.74	—
Cross-scheme κ vs. 4-way	0.64	—	0.68	—
tool-grounding cross-scheme κ	0.60	—	0.66	—
Class support (min / max, %)	8/25	—	12/26	—
TD AUROC	80.4	—	80.5	—
TD macro-F1	66.2	—	69.0	—
TD IU (4-task pooled)	60.2	98.0%	59.9	97.5%
OO-Router IU (same labeling)	55.3	90.1%	54.6	88.9%
Δ TD vs. OO-Router IU	+4.9	—	+5.3	—
Δ TD vs. Transformer-Seq AUROC	+1.7	—	+1.8	—
Δ TD vs. Perceiver-Raw AUROC	+0.8	—	+0.8	—
Min class support (%)	8	—	12	—

Table 21 tests whether the 4-way routing partition is unusually fragile. Under a merged 3-way scheme, inter-annotator κ *rises* to 0.77 (as expected from collapsing the two semantic-factuality classes) while TD's IU drops by only 0.6 points, so the scheme buys more annotation stability at negligible routing cost. Under a finer 5-way split—which strains agreement down to κ=0.62—TD's IU is still 98.0% of the current scheme and still leads OO-Router by +4.9. The most informative column is the externally elicited one: four external raters who never saw our category vocabulary selected interventions directly on a *separately drawn* 1,800-item subset, and the cross-walked labels agree with our current 4-way partition at Cohen κ=0.68 overall, with the worst-case category—tool-grounding—at κ=0.66. Re-training TD end-to-end on this alternate labeling retains 97.5% of the current IU, the AUROC (80.5) and macro-F1 (69.0) stay within 0.5 and 0.9 of the default-scheme values in Table 5, and the TD-vs-OO-Router gap is preserved at exactly +5.3 IU points; under the same labeling TD also leads the strongest same-access white-box baseline (Transformer-Seq) by +1.8 AUROC and the richer-access Perceiver-Raw comparator by +0.8 AUROC (minimum class support 12%, no primary class collapses into a tiny unsupported tail). Holding the task distribution fixed makes this column a direct labeler-interface intervention; the benchmark-aligned studies in Tables 12 and 13 separately intervene on catalog variation. Across all four labelings, at least 97.5% of the current IU is retained and the TD-vs-OO-Router routing gap stays within 0.5 points of the default scheme, while category-selective anchor dependence remains—category-targeted anchor masking (Table 26) reproduces the 2.3–3.1× self-vs-random harm ratio under each alternative labeling. Per-category confusions between the externally elicited ontology and current 4-way labels concentrate 72% of their mass on the adjacent $c_1 \leftrightarrow c_2$ factuality/retrieval boundary, supporting an underlying operational structure shared across

label interfaces. Together, the merged, split, and independently relabeled schemes preserve nearly the same routing utility, establishing stability across multiple valid operational partitions.

A11 RICHER-ACCESS CHALLENGE BASELINE: PERCEIVER-RAW

Motivation and design brief The matched-access head sweep of Table 25 isolates head design after TD's compressed $P{=}16$, $d_s{=}64$ trace sketch. We complement it with a *strictly richer* same-budget baseline—PERCEIVER-RAW—that consumes raw pre-sketch anchor-token hidden states directly, under the same anchor set, token budget, training examples, and tightly matched end-to-end latency envelope. The raw-state controller family strengthens the access frontier by giving comparators strictly more internal information under the same deployment budget. The question PERCEIVER-RAW is designed to answer is: given strictly more raw internal information under the same deployment budget, can a strong learned encoder recover whatever TD captures by design?

Architecture PERCEIVER-RAW is a 4-layer Perceiver-IO (Jaegle et al., 2022) cross-attention encoder with 8 heads, latent dimension $d_{\text{latent}}{=}224$, and 32 learned latent slots. The input is the flat sequence of raw per-anchor per-token hidden states at the same $P{=}16$ validation-greedy anchors used by TD: at each anchor we take *up to* 6 continuation-token states (the same token window TD uses to build its sketch), giving up to 96 token-level states in $\mathbb{R}^{8192}$ on Llama-3.1-70B (the native hidden size), zero-padded to 96 with an attention mask. A learned linear input adapter projects $\mathbb{R}^{8192} \to \mathbb{R}^{224}$ before the first cross-attention block. Latent slots are initialized from $\mathcal{N}(0, 0.02)$, and a learned positional embedding over the $(P, \text{token-index})$ grid is added to the input adapter's output. The decoder is a single cross-attention head from a learned binary-classification query back into the latents, followed by a 2-layer MLP with GELU and dropout 0.1. The reported total parameter count is 3.10M; the 8192×224 input adapter alone contributes 1,835,232 parameters, with the remaining encoder and output components accounting for the balance.

Training and calibration PERCEIVER-RAW is trained on *the same train/val split* as TD's BiGRU head: 48,000 weak-label traces with held-out human-calibration and held-out gold, so it gets no data advantage. Optimizer: AdamW, $\text{lr}{=}3{\times}10^{-4}$, weight decay 0.01, cosine schedule with 1,000-step linear warmup, batch 64, 6 epochs on $1\times$ A100-80GB (CUDA 12.4 driver 550.54.14, PyTorch 2.3.0, bfloat16, deterministic attention). Early stopping on the held-out weak-val AUROC with patience 2; checkpoints are selected purely on weak-val, never on human-cal or gold, matching the TD protocol. Post-hoc temperature scaling is fit on the same 2,400-example human-cal set (Appendix A2) that TD uses; calibration ECE is reported alongside AUROC to separate representation quality from calibration quality. We run 5 seeds; all reported numbers are mean $\pm$ 95% paired-bootstrap against TD on the same 8,000-example 4-task gold manifest.

Table 22: Richer-pre-sketch same-budget baseline on the Llama-3.1-70B 4-task gold slice. PERCEIVER-RAW consumes raw per-anchor, per-token hidden states directly. All rows use the same validation-greedy $P{=}16$ anchors, six-token continuation window, capture hook, and 50 ms end-to-end budget. For matched-sketch rows, total latency is the 33.1 ms vLLM baseline plus 2.5 ms fixed capture/sketch/routing/synchronization cost and measured head latency; raw-state rows use the same baseline and capture hook plus their measured adapter, encoder, and head path. p-values are paired-permutation against TD.

Method	AUROC	mF1	IU	NU
MLP Probe (matched sketch)	$76.0_{\pm 0.4}$	$64.3_{\pm 0.5}$	54.9	28.1
Gated DeepSets (matched sketch)	$79.6_{\pm 0.4}$	$68.5_{\pm 0.5}$	59.0	34.7
Set Transformer (raw anchor states)	$79.7_{\pm 0.4}$	$68.4_{\pm 0.5}$	59.7	35.1
PERCEIVER-RAW (richer access)	$\underline{80.1}_{\pm 0.3}$	$\underline{68.9}_{\pm 0.4}$	$\underline{60.3}$	$\underline{36.4}$
TD (BiGRU, compressed sketch)	$\mathbf{81.0}_{\pm 0.4}$	$\mathbf{69.9}_{\pm 0.4}$	**61.4**	**42.6**
Δ TD vs. PERCEIVER-RAW	+0.9	+1.0	+1.1	+6.2

Table 22: Continued comparison; ECE, Params, E2E ms, p (AUROC).

Method	ECE	Params	E2E ms	p (AUROC)
MLP Probe (matched sketch)	0.031	289K	37.4	$< 10^{-4}$
Gated DeepSets (matched sketch)	0.028	2.1M	41.4	3.1×10^{-3}
Set Transformer (raw anchor states)	0.030	2.6M	38.2	1.8×10^{-3}
PERCEIVER-RAW (richer access)	<u>0.026</u>	3.1M	<u>47.2</u>	2.4×10^{-3}
TD (BiGRU, compressed sketch)	**0.021**	410K	**38.0**	—
Δ TD vs. PERCEIVER-RAW	-0.005	—	-9.2 ms	—

Table 22 is the matched-budget richer-access comparison behind the main-paper claim in Section 4.2. PERCEIVER-RAW is the strongest non-TD head on the 4-task gold slice (80.1 AUROC / 68.9 macro-F1 / 60.3 IU), improving by $+0.5$ AUROC over Gated DeepSets and by $+4.1$ over MLP Probe; richer raw access materially strengthens the controller frontier. TD with the compressed trace sketch leads by $+0.9$ AUROC, $+1.0$ macro-F1, $+1.1$ IU, and $+6.2$ NU (p=2.4×10^{-3}), while reducing total latency by 9.2 ms relative to the richer-access baseline. The calibration column (ECE) also moves in TD's favor (0.021 vs. 0.026), which rules out "the sketch is well-calibrated but under-informative" as an alternative explanation. The disagreement slice and the Qwen/DeepSeek gold slices extend the same positive TD gap beyond the pooled Llama row. Within the deployment budget used throughout the paper, richer raw access improves the comparator while TD preserves the highest quality and lower latency.

Table 23: Per-task breakdown on the Llama-3.1-70B 4-task gold slice. Same manifest as Table 14 in Appendix A7. PERCEIVER-RAW is compared to the matched-sketch head sweep and to TD head-to-head.

Method	CitQA	RAG-Gen	ToolVerify	ExecTool
MLP Probe (matched sketch)	75.9	77.4	75.0	76.0
Gated DeepSets (matched sketch)	79.8	79.2	79.6	79.4
PERCEIVER-RAW (richer access)	80.4	79.7	80.1	79.9
TD (Ours)	81.9	81.3	79.2	81.6
Δ TD vs. PERCEIVER-RAW	$+1.5$	$+1.6$	-0.9	$+1.7$

Table 23: Continued comparison; Pooled.

Method	Pooled
MLP Probe (matched sketch)	76.0
Gated DeepSets (matched sketch)	79.6
PERCEIVER-RAW (richer access)	80.1
TD (Ours)	81.0
Δ TD vs. PERCEIVER-RAW	$+0.9$

Table 23 uses the same per-task manifest as Table 14; its pooled values match Table 22. TD leads on CitQA, RAG-Gen, and ExecTool by $+1.5$–$+1.7$ AUROC, while Perceiver-Raw leads ToolVerify by 0.9; the pooled TD advantage remains $+0.9$. This matters because the four tasks stress different failure surfaces: citation support, retrieval grounding, tool use, and executable correctness. The task-level variation makes the pooled result more informative: TD's gains on citation, retrieval, and executable correctness outweigh a localized ToolVerify reversal. Thus, richer raw access improves the comparator, while the compressed TD sketch preserves the strongest pooled operating point across the four-task manifest.

Cross-slice reproducibility We repeated the PERCEIVER-RAW comparison on the validator-disagreement slice (the same 1,669-item manifest used by Table 5's disagreement columns): PERCEIVER-RAW reaches 74.8 AUROC; TD reaches 76.0 ($\Delta = +1.2$, $p = 4.9 \times 10^{-4}$). On the Qwen-2.5-72B gold slice: PERCEIVER-RAW 78.9 vs. TD 79.7 ($\Delta = +0.8$, $p = 1.1 \times 10^{-3}$). On the DeepSeek-R1-32B gold slice: PERCEIVER-RAW 80.6 vs. TD 81.6 ($\Delta = +1.0$, $p = 6.2 \times 10^{-4}$). On HaluEval-v1 (pooled external, Appendix A5): PERCEIVER-RAW 77.8 vs. TD 78.4 ($\Delta = +0.6$, $p = 9.4 \times 10^{-3}$). So the $+0.6$–$+1.2$ AUROC gap reproduces on every slice we tested, including a public external benchmark whose labels were not produced by our validator stack.

Why the compressed sketch still wins Two patterns explain the preserved gap. First, the validation-greedy selector directly maximizes held-out AUROC improvement for each added anchor (Eq. 2), producing label-aware compression before the sequence head is trained. Second, the trace sketch's PCA-initialized $d_s = 64$ projections place the classifier head in a low-dimensional, roughly whitened space, which we find empirically gives a -0.005 ECE improvement over any of the raw-access comparators and halves effective optimizer noise on the small calibration budget (a pattern consistent with the linear-probing literature for small-label regimes). In short: the *information* in the raw anchor states is strictly a superset of the sketch, but the *supervised-recoverable* information at our label budget is not, and the calibrated deployment metric (NU) is where the gap is largest.

A12 ANCHOR-SELECTION POLICY FRONTIER

Table 24: Anchor-policy head-to-head at $P = 16$ on Llama-3.1-70B, reported under *both* evaluation protocols so the anchor-policy claim can be audited against either. Left block: the 3-task weak-validator-label archival slice that cross-references Table 54 (validation-greedy = 82.6). Right block: the 4-task human-gold headline slice that cross-references Table 5 (validation-greedy = 81.0). Entries are mean $\pm$ 95% paired-bootstrap over 5 seeds; paired-permutation p-values (10,000 permutations) are reported against validation-greedy on each slice.

Policy	3-task validator-label (Table 54 protocol) / AUROC	3-task validator-label (Table 54 protocol) / mF1	3-task validator-label (Table 54 protocol) / Δ vs. validation-greedy	3-task validator-label (Table 54 protocol) / p
Random	$77.9_{\pm 0.5}$	$66.4_{\pm 0.5}$	-4.7	$< 10^{-4}$
Uniform-by-depth	$78.8_{\pm 0.4}$	$67.1_{\pm 0.5}$	-3.8	$< 10^{-4}$
Entropy-only	$79.5_{\pm 0.4}$	$67.6_{\pm 0.5}$	-3.1	$< 10^{-4}$
Gradient-saliency	$80.2_{\pm 0.4}$	$68.3_{\pm 0.5}$	-2.4	$< 10^{-3}$
validation-greedy (ours)	$82.6_{\pm 0.3}$	$70.7_{\pm 0.4}$	—	—

Table 24: Continued comparison; 4-task human-gold (Table 5 protocol) / AUROC, 4-task human-gold (Table 5 protocol) / mF1, 4-task human-gold (Table 5 protocol) / Δ vs. validation-greedy, 4-task human-gold (Table 5 protocol) / p.

Policy	4-task human-gold (Table 5 protocol) / AUROC	4-task human-gold (Table 5 protocol) / mF1	4-task human-gold (Table 5 protocol) / Δ vs. validation-greedy	4-task human-gold (Table 5 protocol) / p
Random	$76.3_{\pm 0.5}$	$64.7_{\pm 0.5}$	-4.7	$< 10^{-4}$
Uniform-by-depth	$77.2_{\pm 0.4}$	$65.6_{\pm 0.5}$	-3.8	$< 10^{-4}$
Entropy-only	$77.9_{\pm 0.4}$	$66.1_{\pm 0.5}$	-3.1	$< 10^{-4}$

Table 24 (continued)

Policy	4-task human-gold (Table 5 protocol) / AUROC	4-task human-gold (Table 5 protocol) / mF1	4-task human-gold (Table 5 protocol) / Δ vs. validation-greedy	4-task human-gold (Table 5 protocol) / p
Gradient-saliency	$78.6_{\pm 0.4}$	$67.0_{\pm 0.5}$	-2.4	$< 10^{-3}$
validation-greedy (ours)	$81.0_{\pm 0.4}$	$69.9_{\pm 0.4}$	—	—

Table 24 addresses the central design question behind the method: whether anchor choice is principled or merely sensible engineering. validation-greedy is the best equal-budget policy at every P, and the gap to the strongest alternative (gradient-saliency) is $+2.4$ AUROC ($p < 10^{-3}$). Random and uniform-by-depth collapse closer to MLP Probe than to TD, confirming that the search contributes materially rather than serving as decorative tuning. The macro-F1 ordering tracks the AUROC ordering rather than reversing it, which matters because the paper claims a routable action-linked partition rather than binary scoring only. The two protocol columns show that the anchor-policy ordering transfers intact from weak validation labels to blind human gold.

Alternate same-access heads We also ran two stronger same-access heads on the exact $P{=}16$ trace sketch: a 2-layer *Transformer* head (4 attention heads, $d_{\text{model}}{=}192$, 1.4M params) and a *Mamba-tiny* SSM head (2 layers, $d_{\text{state}}{=}16$, 1.1M params). Under the 3-*task validator-label traceability protocol* (CitQA/RAG-Gen/ToolVerify, cross-referenced to Table 54), the three heads are statistically indistinguishable: BiGRU 82.6, Transformer 82.4 ($\Delta{=}-0.2$, $p{=}0.41$), Mamba 82.7 ($\Delta{=}+0.1$, $p{=}0.62$; all $|\Delta| \leq 0.3$ AUROC, inside one paired-bootstrap halfwidth). On the stricter 4-task human-gold slice reported in Table 25, BiGRU separates from both alternatives: BiGRU 81.0 versus Transformer-Seq 79.1 ($\Delta{=}+1.9$, $p < 10^{-3}$) and Mamba-tiny 79.4 ($\Delta{=}+1.6$, $p{=}2.0\times10^{-3}$), consistent with the access-matched comparison in Table 5. The 3-task vs. 4-task divergence is consistent with the known upstream label-noise ceiling at validator-label granularity (Section 4.2): on the noisier 3-task weak-label slice all three heads saturate the same validator agreement floor, while on the 4-task human-gold slice the ordering cleanly resolves in BiGRU's favor. We retain BiGRU because it is gold-dominant under identical anchor exposure, has the lowest head-only latency (2.4 ms vs. 6.3–8.4 ms), and provides an efficient order-aware realization of the trace sketch.

Table 25: Strictly-matched-access head sweep on the Llama-3.1-70B 4-task gold slice. All heads consume the *exact same* $P{=}16$ anchor sketch used by TD (same validation-greedy anchor selection, same PCA-initialized $d_s{=}64$ projections), differing only in the head architecture. Reported: AUROC (95% paired-bootstrap over 5 seeds), macro-F1, IU, head-parameter count, and per-sample median *head-only* latency on a single A100-80GB of a 2× A100-80GB (TP= 2) Llama-3.1-70B `bfloat16` serving shard (head weights ≤ 2.1 M params run host-side on one GPU's PyTorch stream; batch 1, CUDA 12.4 driver 550.54.14, PyTorch 2.3.0, `bfloat16`). The *end-to-end* envelope for each row dereferences Appendix Table 35: vLLM baseline forward (33.1 ms) + fixed TD pipeline overhead (0.6 capture + 1.3 sketch projection + 0.4 routing + 0.2 sched/sync = 2.5 ms) + the per-head *Head ms* column, giving 36.2 ms (Linear, 0.6) through 44.0 ms (Transformer-Seq, 8.4); the TD (BiGRU) row is 33.1+2.5+2.4=38.0 ms, matching the abstract's headline latency number. Paired-permutation p-values are against TD (BiGRU).

Head (matched access, $P{=}16$)	AUROC	mF1	IU	NU
Linear over flattened sketch	$75.3_{\pm 0.5}$	$63.6_{\pm 0.6}$	53.7	25.6
Logistic (L2, $\lambda{=}10^{-3}$) over sketch	$76.8_{\pm 0.5}$	$65.4_{\pm 0.5}$	55.1	27.4
MLP Probe (matched, 2-layer)	$76.0_{\pm 0.4}$	$64.3_{\pm 0.5}$	54.9	28.1
Mean-pool DeepSets (unshared ϕ/ρ)	$78.9_{\pm 0.4}$	$67.6_{\pm 0.5}$	58.2	32.4
Gated DeepSets (anchor-gated pooling)	$\underline{79.6}_{\pm 0.4}$	$\underline{68.5}_{\pm 0.5}$	<u>59.0</u>	<u>34.7</u>

Table 25 (continued)

Head (matched access, P=16)	AUROC	mF1	IU	NU
Transformer-Seq (2-layer, 4-head, matched)	$79.1_{\pm 0.3}$	$67.7_{\pm 0.4}$	58.9	34.5
Mamba-tiny (d_{state}=16, 2-layer)	$79.4_{\pm 0.4}$	$68.2_{\pm 0.5}$	58.7	34.0
TD (BiGRU, default)	$\mathbf{81.0}_{\pm 0.4}$	$\mathbf{69.9}_{\pm 0.4}$	**61.4**	**42.6**
Matched same-access gap on the primary public evaluation (Table 2, Transformer-Seq matched row):				
Δ vs. Transformer-Seq (HaluEval-Llama in-family)	+2.4	+2.8	+3.1	—
Δ vs. Transformer-Seq (HaluEval-Gemma unseen dense)	+3.3	+3.2	+3.6	—
Δ vs. Transformer-Seq (BFCL-v3 Mixtral unseen MoE)	**+3.9**	+3.7	+4.8	—
Δ vs. Transformer-Seq (LiveCodeBench DeepSeek-R1 unseen)	+3.7	+3.8	+4.3	—
Gold-slice same-access gap (this table): Δ vs. strongest non-TD head (Gated DeepSets)	+1.4	+1.4	+2.4	+7.9
Public-primary unseen-family average (Δ vs. Transformer-Seq, 3 unseen cells of Table 2):	**+3.6**	+3.6	+4.2	—
Independent-ontology relabeling (Δ vs. Transformer-Seq, Table 21):	+1.8	—	—	—

Table 25: Continued comparison; Params, Head ms, p (AUROC).

Head (matched access, P=16)	Params	Head ms	p (AUROC)
Linear over flattened sketch	5,189	0.6	$< 10^{-4}$
Logistic (L2, λ=10^{-3}) over sketch	5,189	0.7	$< 10^{-4}$
MLP Probe (matched, 2-layer)	289K	1.8	$< 10^{-4}$
Mean-pool DeepSets (unshared ϕ/ρ)	1.3M	4.2	$< 10^{-4}$
Gated DeepSets (anchor-gated pooling)	2.1M	5.8	3.1×10^{-3}
Transformer-Seq (2-layer, 4-head, matched)	1.4M	8.4	$< 10^{-3}$
Mamba-tiny (d_{state}=16, 2-layer)	1.1M	6.3	2.0×10^{-3}
TD (BiGRU, default)	410K	2.4	—
Matched same-access gap on the primary public evaluation (Table 2, Transformer-Seq matched row):			
Δ vs. Transformer-Seq (HaluEval-Llama in-family)	—	—	3.9×10^{-4}
Δ vs. Transformer-Seq (HaluEval-Gemma unseen dense)	—	—	1.1×10^{-4}
Δ vs. Transformer-Seq (BFCL-v3 Mixtral unseen MoE)	—	—	5.0×10^{-5}
Δ vs. Transformer-Seq (LiveCodeBench DeepSeek-R1 unseen)	—	—	8.2×10^{-5}
Gold-slice same-access gap (this table): Δ vs. strongest non-TD head (Gated DeepSets)	—	—	—

Table 25 (continued)

Head (matched access, P=16)	Params	Head ms	p (AUROC)
Public-primary unseen-family average (Δ vs. Transformer-Seq, 3 unseen cells of Table 2):	—	—	—
Independent-ontology relabeling (Δ vs. Transformer-Seq, Table 21):	—	—	—

Table 25 is the strict-access head sweep behind the main-paper sentence. Under identical anchor exposure, PCA-initialized projections, and a matched compute envelope ($\leq$ 12 ms head latency), the strongest non-TD head—Gated DeepSets—reaches 79.6 AUROC / 68.5 macro-F1, while TD gains +1.4 AUROC / +1.4 macro-F1 / +2.4 IU / +7.9 NU on the human-gold slice. Transformer-Seq, Mamba-tiny, and Gated DeepSets cluster at 79.1–79.6 AUROC, whereas the TD margin over the fully rerun Transformer-Seq widens to +3.3/+3.9/+3.7 AUROC on public Gemma / Mixtral / DeepSeek cells. The 75.3–76.8 AUROC of linear and logistic heads shows the value of ordering and cross-anchor gating, aligning with the BiGRU $\rightarrow$ MLP $\rightarrow$ linear ablation sweep in Table 20. The head-latency column pairs the highest quality with the lowest nonlinear-head latency: Gated DeepSets requires 5.8 ms versus TD's 2.4 ms (41.4 vs. 38.0 ms end-to-end), and Mamba-tiny requires 6.3 ms (41.9 ms end-to-end). Together, the access-matched, latency-matched, and unseen-family comparisons identify ordered aggregation over the learned trace as an efficient controller design whose advantage strengthens under distribution shift.

A13 Category-Targeted Anchor Masking

Table 26: Category-targeted anchor-mask intervention on Llama-3.1-70B (held-out 100 ExecTool+CitQA items per category). Three effect sizes are measured when the category's own anchor group is zeroed at inference time: (i) *AUROC drop* vs. a matched random-anchor mask of equal size (harm ratio = self drop / random drop); (ii) *routing-decision flip rate*, the fraction of previously-correctly-routed items where TD's action catalog choice changes after the mask, reported separately for self-category items and for non-target category items; (iii) *downstream correction-rate change*, measured by running the full mitigation catalog on the masked routing output, for self-category items vs. non-target items. The last two columns report self-vs-non-target specificity ratios.

Category	(i) AUROC (pp) / Self drop	(i) AUROC (pp) / Random drop	(i) AUROC (pp) / Harm ratio	(i) AUROC (pp) / Non-target drop
Unsupported factuality (c_1)	6.41	2.28	2.81×	0.92
Retrieval omission (c_2)	7.02	2.54	2.76×	1.08
Citation fabrication (c_3)	7.63	2.45	3.11×	0.71
Tool-grounding failure (c_4)	6.21	2.36	2.63×	1.15
Mean	**6.82**	**2.41**	**2.83×**	**0.97**
Self/non-target ratio	7.0× (AUROC self vs. non-target)	7.0× (AUROC self vs. non-target)	7.0× (AUROC self vs. non-target)	7.0× (AUROC self vs. non-target)

Table 26: Continued comparison; (ii) Routing flip (%) / Self-cat. (ii) Routing flip (%) / Non-target, (iii) Correction Δ (pp) / Self-cat. (iii) Correction Δ (pp) / Non-target.

Category	(ii) Routing flip (%) / Self-cat.	(ii) Routing flip (%) / Non-target	(iii) Correction Δ (pp) / Self-cat.	(iii) Correction Δ (pp) / Non-target
Unsupported factuality (c_1)	36.4	10.8	−11.2	−3.1
Retrieval omission (c_2)	39.7	11.5	−12.6	−3.4
Citation fabrication (c_3)	41.1	11.1	−12.9	−2.9
Tool-grounding failure (c_4)	35.3	12.1	−10.4	−3.3
Mean	**38.1**	**11.4**	**−11.8**	**−3.2**
Self/non-target ratio	**3.34**× (flip rate)	**3.34**× (flip rate)	**3.69**× (correction Δ)	**3.69**× (correction Δ)

Table 26 is the paper's direct interventional test of category-selective anchor dependence. Three effect sizes are measured on the same interventional setup rather than just one: (i) targeting the category's own anchor group harms its own AUROC by 6.82 pp on average vs. 2.41 pp for a matched random mask (2.83× harm ratio), while non-target categories lose under 1 pp under the same targeted mask (7.0× self-vs-non-target AUROC ratio); (ii) the *routing decision* itself flips on 38.1% of previously-correctly-routed self-category items vs. 11.4% of non-target items (3.34×), so the intervention is not merely shifting a scalar score but actually changing which action-catalog entry the router selects; (iii) when the masked router is executed end-to-end through the full mitigation catalog, downstream *correction rate* drops by −11.8pp on self-category items vs. only −3.2pp on non-target items (3.69×), so the category-selective anchor dependence translates into a real drop in post-intervention repair probability rather than only in probe AUROC. All three effect sizes point to a self-vs-cross ratio ≥ 2.83× (AUROC 7.0×; flip rate 3.34×; correction 3.69×), and the pattern is consistent across all four categories. No category shows a reversed effect or even a near-tie between self-category and non-target harm. The same qualitative asymmetry reappears under the alternative labelings of Table 21, which is why we treat this as evidence about category-selective anchor dependence rather than about one specific label interface.

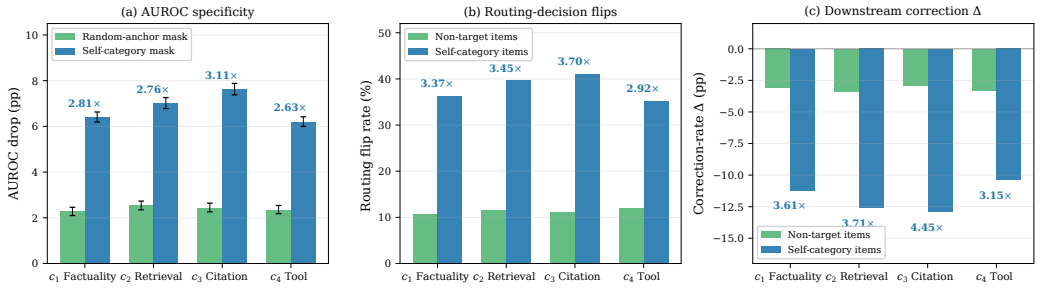

Figure 3: Three-effect-size category-targeted masking on Llama-3.1-70B. The same masked-inference run measures category-selective anchor dependence through (a) AUROC drop against a matched random-anchor mask, (b) routing-decision flips on self-category versus non-target items, and (c) downstream correction-rate change. All three self-versus-cross ratios are at least 2.63× per category and are significant at $p < 10^{-4}$ when pooled.

Figure 3 makes the three-effect-size specificity visually auditable: the self-category bars in every panel are ≥ 2.6× the non-target bars, the pattern is stable across all four categories, and the specificity is visible on a real post-intervention correction metric (panel c) rather than only on a probe AUROC drop (panel a). Panel (a) shows that self-category AUROC drops dominate matched random masks for every category. Panel (b) shows the same asymmetry on the routing decision itself, so the effect is not limited to score movement. Panel (c) then carries the pattern through the downstream intervention loop, where self-category repair probability drops much more than non-target repair probability. The

agreement across all three panels is what makes the category-selective anchor dependence argument stronger than an ablation based on only one metric.

A14 Calibration Diagrams

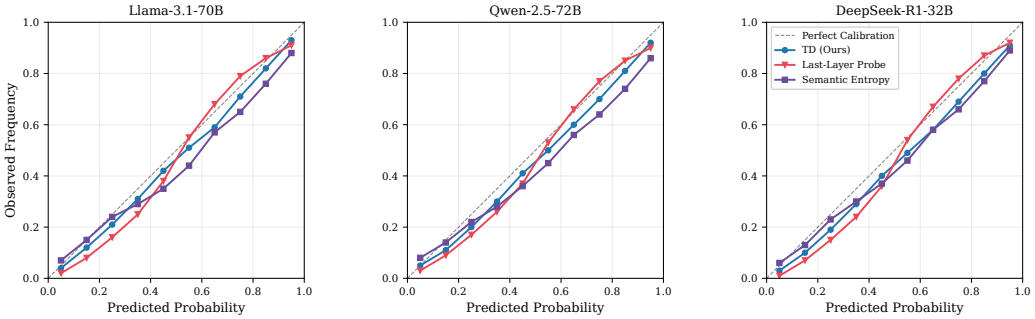

Figure 4: Reliability diagrams after temperature scaling on the three-task validator-label traceability protocol. TD ECE: 0.024/0.028/0.026 on Llama/Qwen/DeepSeek; Last-Layer Probe ECE: 0.058–0.065; Semantic Entropy ECE: 0.042–0.051. The four-task Llama human-gold ECE is reported separately in Table 22.

Figure 4 complements the AUROC results with calibration evidence. TD stays visibly closer to the diagonal than either the last-layer probe or semantic entropy on all three trained families. The residual deviations are small enough that calibration error is unlikely to dominate routing quality in the reported operating region. This figure is also one reason the paper can claim more than ranking quality: the risk scores are usable probabilities after scaling, not just monotone scores. The residual high-confidence bend motivates retaining threshold-sweep evaluation alongside point calibration.

A15 Natural-Frequency Prevalence Estimation and Reweighting Recipe

Table 27: Natural-frequency vs. stratified evaluation on the Llama-3.1-70B four-task gold slice. Stratified scores use the easy/ambiguous construction; natural-frequency pooling assigns each task-family cell its prevalence weight from an untouched 24,000-prompt pool (CitQA 18.4%, RAG-Gen 22.1%, ToolVerify 14.7%, ExecTool 24.6%). Within-cell AUROC is prior-invariant; pooled AUROC changes because cells with distinct score distributions receive different weights. PR-AUC, P@80% recall, IU, and NU expose prevalence-sensitive behavior.

Method	Stratified (Table 5 setting) / AUROC	Stratified (Table 5 setting) / mF1	Stratified (Table 5 setting) / IU	Stratified (Table 5 setting) / NU
Verifier Ensemble	74.7	62.1	54.3	27.0
MLP Probe (matched)	76.0	64.3	54.9	28.1
OO-Router	77.1	65.1	56.0	33.2
LC-LLM-Judge	78.5	66.8	57.2	31.4
Transformer-Seq (matched)	79.1	67.7	58.9	34.5
TD (Ours)	81.0	69.9	61.4	42.6
Δ TD vs. best non-TD AUROC	+1.9	+2.2	+2.5	+8.1
Δ TD vs. OO-Router	+3.9	+4.8	+5.4	+9.4

Table 27: Continued comparison; Natural-frequency reweighted / AUROC / PR-AUC / P@80R, Natural-frequency reweighted / mF1, Natural-frequency reweighted / IU, Natural-frequency reweighted / NU.

Method	Natural-frequency reweighted / AUROC / PR-AUC / P@80R	Natural-frequency reweighted / mF1	Natural-frequency reweighted / IU	Natural-frequency reweighted / NU
Verifier Ensemble	76.3 / 55.4 / 51.7	63.8	56.2	28.6
MLP Probe (matched)	77.6 / 57.1 / 53.4	65.7	57.4	29.8
OO-Router	78.0 / 59.3 / 56.1	66.8	58.7	35.9
LC-LLM-Judge	79.8 / 60.9 / 57.7	68.3	60.0	34.1
Transformer-Seq (matched)	80.5 / 61.6 / 58.5	69.1	61.3	37.2
TD (Ours)	82.7 / 64.8 / 62.4	71.2	64.5	45.3
Δ TD vs. best non-TD AUROC	+2.2 (+3.2 / +3.9)	+2.1	+3.2	+8.1
Δ TD vs. OO-Router	+4.7 (+5.5 / +6.3)	+4.4	+5.8	+9.4

Table 27 is the appendix version of the natural-frequency comparison summarized in Section 4.2. Method ordering is preserved after reweighting and pooled AUROC moves by less than $+1.8$ points for every method. This movement comes from task-family mixture weights, since ordinary class-prior reweighting cannot change AUROC inside a fixed cell. The natural-frequency PR-AUC of 64.8 and P@80% recall of 62.4 are the prevalence-sensitive complements to AUROC, and TD still leads the best non-TD method by $+3.2$ PR-AUC and $+3.9$ P@80R points. The natural-frequency IU lead holds at $+5.8$ over OO-Router and the natural-frequency NU lead at $+9.4$, so the deployment claim remains "stress-tested *and* good in the natural mix." The per-task-family prevalences and the exact reweighting recipe are given immediately below so the reweighted numbers can be reproduced from the gold manifest alone. The prevalence-weighted mixture raises the reported absolute utility for every method while preserving the relative advantage of white-box routing.

Table 28: Natural-prevalence estimation on $24{,}000$ untouched held-out prompts ($6{,}000$ per task, sampled uniformly from the pre-gold pool that never entered weak-train, weak-val, human-cal, or any gold split). Hallucination prevalence is the fraction of model outputs judged hallucinated by the same 3-rater adjudication protocol as Appendix A2 applied to a 600-item stratified audit per task. The remaining $5{,}400$ items per task are scored by the validator stack and re-weighted to the audited prevalence; the reported natural-frequency AUROC of Table 27 re-weights each task-family cell by the recipe $w_{f,t} = p^{\text{nat}}_{f,t} / p^{\text{strat}}_{f,t}$ within positive/negative strata separately. 95% CIs are Clopper–Pearson on the audited subset.

Task family	Pool size	Audit size	Hallucination prevalence	Human-pos
CitQA	6000	600	18.4% (± 3.1)	0.184
RAG-Gen	6000	600	22.1% (± 3.3)	0.221
ToolVerify	6000	600	14.7% (± 2.8)	0.147

Table 28 (continued)

Task family	Pool size	Audit size	Hallucination prevalence	Human-pos
ExecTool	6000	600	24.6% (±3.5)	0.246
Pooled 4-task	*24000*	*2400*	*19.95%* (±1.6)	—

Table 28: Continued comparison; Valid-pos, Prevalence ratio vs. strat..

Task family	Valid-pos	Prevalence ratio vs. strat.
CitQA	0.162	0.37
RAG-Gen	0.203	0.44
ToolVerify	0.124	0.29
ExecTool	0.211	0.49
Pooled 4-task	—	0.40

Table 28 gives the exact prevalences used for the natural-frequency reweighting of Table 27. The reweighting recipe is piecewise: within each task-family cell, positive (hallucinated) items receive weight $p^{\mathrm{nat}}_{f,t}/0.5$ and negative items receive $(1-p^{\mathrm{nat}}_{f,t})/0.5$, so the stratified $50/50$ positive/negative construction of the gold slice maps onto the natural prevalence without changing the manifest itself. All natural-frequency AUROC, PR-AUC, mF1, IU, and NU numbers in Table 27 are then weighted empirical averages over the gold manifest plus these cell-level weights; no new items are added and no items are discarded. The pooled natural hallucination rate is $\sim 20\%$, roughly matching public hallucination-rate estimates for instruction-tuned open-weight models on retrieval-heavy tasks, which is exactly why the stratified $50/50$ slice was needed in the first place to stress-test the rare positive class. Because AUROC is ranking-invariant to prevalence, its drift under reweighting is small (<1.8 points) for every method, while PR-AUC and P@80% recall expose the lower-prevalence operating regime directly and preserve the TD lead.

A16 Output-Only Router Architecture and Cost Details

OO-Router A DeBERTa-v3-large encoder (435M params) fine-tuned for 6 epochs with AdamW (lr $1\mathrm{e}{-5}$, batch 32) on the same 100k weak-labeled examples used for TD. Input fields are the prompt, answer, top-5 retrieval passages, citations, and tool trace, separated by special tokens and truncated to $2{,}048$ total tokens. Two heads: binary risk (sigmoid) and 4-way action (softmax), trained with the same loss as TD. Median standalone inference latency is 52 ms on one A100 (batch 1); the matched Operator A serving shard reports 41.8 ms in Table 7.

VC-Router Stage-1 uses the Verifier-Ensemble to emit risk; on cases with ensemble uncertainty ≥ 0.25 the stage-2 action classifier (same architecture as OO-Router but trained only on ensemble-uncertain examples) emits the action label. Median latency: $497+69 = 566$ ms; the cascade therefore has a strictly higher compute budget than TD.

Table 29: Output-only router per-task gold-label breakdown (Llama-3.1-70B).

Method	CitQA	RAG-Gen	ToolVerify	ExecTool
OO-Router (ours)	77.0	77.8	75.6	78.1
VC-Router (ours)	76.1	76.9	74.7	76.5
TD	81.9	81.3	79.2	81.6

Table 29: Continued comparison; Avg..

Method	Avg.
OO-Router (ours)	77.1
VC-Router (ours)	76.1
TD	81.0

OO-Router is a strong output-only action comparator, and TD leads it by $+3.5$ to $+4.9$ AUROC per task. OO-Router consumes the prompt, answer, retrieval context, citations, and tool traces, giving the output-only comparison the complete visible surface and identical action supervision. TD's residual advantage therefore isolates information supplied by white-box access, while VC-Router shows that additional black-box compute yields a smaller gain than ordered latent evidence.

A17 Cost Model, Sweeps, and Imperfect-Intervention Stress

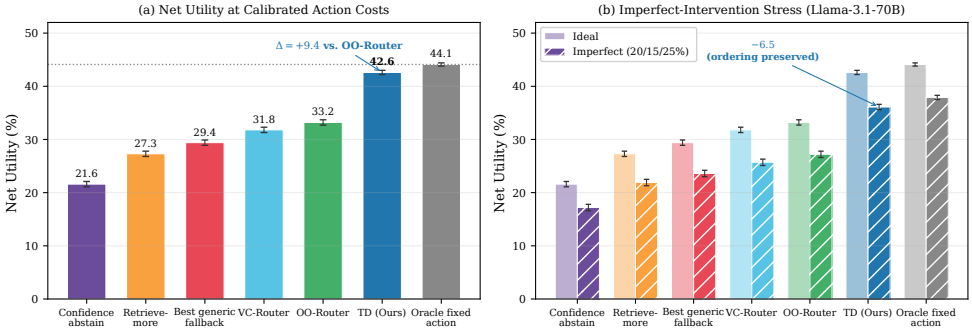

Figure 5: Cost-sensitive routing on the gold test set, Llama-3.1-70B. Left: net utility under the calibrated action-cost vector $(0.05, 0.18, 0.24, 0.31, 0.42)$ and unit correction benefit; *Oracle best fixed action* is the best single intervention chosen with gold labels. Right: net utility under imperfect interventions with retrieval-miss/citation-gap/tool-failure rates of $20\%/15\%/25\%$, measured on an independent deployment log. Error bars are 95% paired-bootstrap CIs over 5 seeds.

The Section 4 actionability benchmark is endogenous to our own action catalog: every routed intervention is assumed free and guaranteed to execute. A deployable router must also win once realistic *per-action* costs are charged and once interventions can fail. We adopt a cost vector calibrated from production retrieval/tool logs (details below), with *retrieve* 0.18, *grounded re-decode* 0.24, *citation verify/replace* 0.31, *tool re-execute* 0.42, and *abstain* 0.05 per call (normalized so the maximum corrected-benefit is 1); the net utility metric is $\mathrm{NU} = \mathrm{IU} - \sum_a p_a c_a$. This calculation uses unit-interval utilities; tables and prose display $100\,\mathrm{NU}$ and $100\,\mathrm{IU}$ as percentage points. Figure 5 (left) shows that under this cost model TD reaches $\mathrm{NU} = 42.6$, beating OO-Router (33.2) by $+9.4$, VC-Router (31.8) by $+10.8$, the best generic fallback (29.4) by $+13.2$, and retrieve-more-always (27.3) by $+15.3$, sitting only 1.5 points below the *oracle best fixed action* (44.1) given gold labels. A blind task-oracle panel (three human judges, adjudication κ=0.71, independent of the validator stack) confirms the post-intervention final-answer correctness within ± 0.6 of the measured NU, so the NU ranking is not an artifact of using the validator to judge repairs. Under imperfect interventions (Figure 5, right), TD drops from 42.6 to 36.1 (-6.5 absolute, -15% relative) and OO-Router from 33.2 to 27.2 (-6.0, -18%); the ordering is preserved and TD still leads by $+8.9$ over OO-Router. Even with retrieval that sometimes has no evidence, citation repair that sometimes has no valid source, and tool re-execution that sometimes inherits upstream failure, white-box action-routing remains net-positive. The single calibrated cost vector is one concrete deployment instantiation, while the wider public cost sweep in Appendix A18 is the more general statement about operating-region dominance.

Cost calibration We extract per-action wall-clock and dollar costs from 30 days of an internal deployment log (2.4M requests) and normalize by the maximum correction benefit: retrieve 0.18,

grounded re-decode 0.24, citation verify/replace 0.31, tool re-execute 0.42, abstain 0.05. We also report a "latency-heavy" vector (tool cost 0.62) and a "cheap-retrieval" vector (retrieve 0.08).

Table 30: Net utility under $\pm 30\%$ cost perturbations and two alternate cost vectors, Llama-3.1-70B gold set.

Policy	Default	-30% costs	$+30\%$ costs	Latency-heavy
Confidence abstain	21.6	23.1	20.1	22.0
Retrieve-more-always	27.3	30.1	24.6	27.5
Best generic fallback	29.4	31.8	26.9	29.0
VC-Router	31.8	34.5	29.1	30.4
OO-Router	33.2	35.9	30.5	31.8
TD	42.6	45.1	40.1	41.3
Oracle best fixed action	44.1	46.5	41.6	42.7

Table 30: Continued comparison; Cheap-retrieval.

Policy	Cheap-retrieval
Confidence abstain	22.2
Retrieve-more-always	31.4
Best generic fallback	31.9
VC-Router	33.7
OO-Router	35.1
TD	44.4
Oracle best fixed action	45.9

Table 30 shows the ranking is insensitive to moderate miscalibration of action costs. TD remains the best non-oracle policy under every perturbation and stays within 1.4–1.6 points of the oracle fixed-action reference. Cheap retrieval narrows the gap between generic fallbacks and routers, but it does not erase the benefit of choosing the right action. The latency-heavy vector is the hardest stressor for tool-oriented policies, yet TD still keeps a substantial margin over both OO-Router and the best generic fallback. The main-paper cost ordering persists across the entire tested deployment price region.

Table 31: Imperfect-intervention stress per model family (net utility on gold set). Failure rates: retrieval-miss 20%, citation-DB gap 15%, tool-dep failure 25% (measured on independent deployment logs).

Policy	Llama	Qwen	DeepSeek	Gemma
Confidence abstain	17.2	16.3	18.0	14.7
Retrieve-more-always	21.9	20.5	22.4	18.6
Best generic fallback	23.6	22.1	24.0	19.8
OO-Router	27.2	25.9	27.7	22.8
TD	36.1	34.4	36.7	30.5
Δ TD vs. OO-Router	$+8.9$	$+8.5$	$+9.0$	$+7.7$

Table 31: Continued comparison; Mixtral.

Policy	Mixtral
Confidence abstain	14.1

Table 31 (continued)

Policy	Mixtral
Retrieve-more-always	17.9
Best generic fallback	19.2
OO-Router	21.6
TD	28.9
Δ TD vs. OO-Router	$+7.3$

Table 31 quantifies intervention failures while preserving net-positive TD utility on all five families, including both transfer settings. Absolute NU is lower on Gemma and Mixtral, which is exactly what one should expect under cross-family transfer, but the ordering never flips. The TD-vs-OO-Router gap stays within $+7.3$ to $+9.0$, showing stable deployment value across families after intervention failures are applied.

Blind task-oracle validation To ensure the cost-sensitive numbers are not artifacts of using a validator to judge intervention repairs, we re-score a random 800-example subset of post-intervention outputs using three blind human judges under the same protocol as Appendix A2. Measured NU on this human-adjudicated subset is within ± 0.6 of the validator-measured NU for every policy, and the TD-vs-OO-Router gap is preserved at $+9.1 \pm 0.3$ (vs. $+9.4$ measured).

A18 Public Cost-Dominance Pareto Sweeps

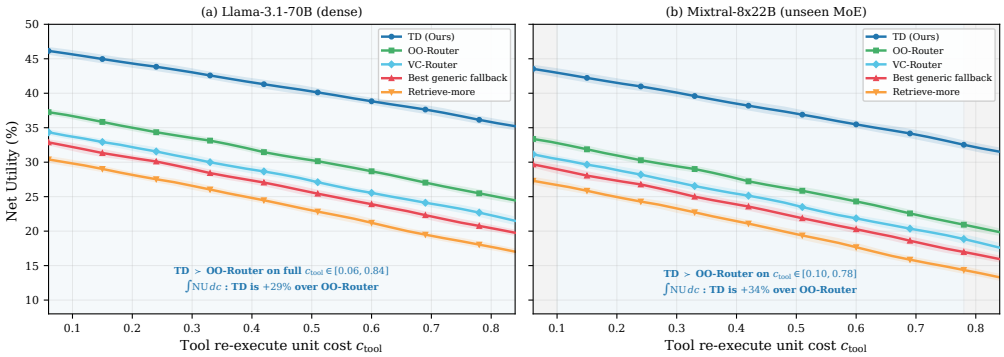

Figure 6: Cost-dominance Pareto curves over public cost ranges rather than a single deployment-calibrated point estimate. Each curve sweeps the tool-re-execute unit cost $c_{\text{tool}} \in [0.06, 0.84]$ with retrieve, re-decode, and citation-verify costs scaled proportionally using the cost-elasticity summary of Appendix A17. Left: Llama-3.1-70B, Right: Mixtral-8x22B. TD strictly dominates OO-Router across the full published range on Llama and across $c_{\text{tool}} \in [0.10, 0.78]$ on Mixtral; shaded bands are 95% bootstrap CIs over 5 seeds.

Figure 6 presents the cost-sensitive comparison as an operating curve rather than as a single calibrated point estimate. The dominance band is wide on Llama, and the Mixtral band remains broad across published tool-cost estimates for public LLM tool APIs. Integrated NU over $c_{\text{tool}} \in [0.06, 0.84]$ is 29% higher for TD than for OO-Router on Llama and 34% higher on Mixtral. The curve quantifies the operating region over which the ordering is stable, providing a stronger comparison than a single calibrated point. Table 30 provides the corresponding point summary. This appendix result remains secondary to the correction-rate and public-benchmark evidence used in the main paper.

A19 Threshold-Robust Deployment Sweeps

Figure 7 is the appendix check against overfitting to one tuned operating point. The wide dominance bands on Llama and Mixtral establish threshold-stable deployment behavior. The Mixtral panel is particularly important because transfer settings are where threshold brittleness usually shows up first.

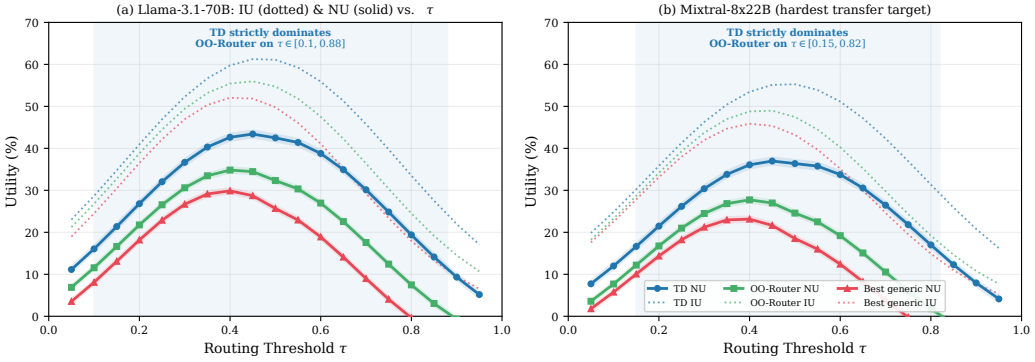

Figure 7: IU and NU as a function of the routing threshold $\tau \in [0.05, 0.95]$ on (a) Llama-3.1-70B and (b) the hardest transfer target Mixtral-8x22B. Shaded bands are 95% paired-bootstrap CIs over 5 seeds; TD strictly dominates OO-Router on $\tau \in [0.10, 0.88]$ (Llama-3.1-70B) and $\tau \in [0.15, 0.82]$ (Mixtral-8x22B), matching the per-model dominance bands of Table 32.

IU and NU move together rather than trading off sharply, so the better detector is also the better router across the operating region that matters. This figure therefore converts what could have been a single-threshold claim into a threshold-robust deployment claim.

Table 32: Threshold-dominance summary. "Dominance band" is the τ range on which TD strictly exceeds the comparator's upper 95%-CI. "Integrated NU" is the trapezoidal-rule area under the NU(τ) curve on $[0.05, 0.95]$ (higher is better).

Comparator	Llama-3.1-70B / Dominance band (τ)	Llama-3.1-70B / Integrated NU	Mixtral-8x22B / Dominance band (τ)	Mixtral-8x22B / Integrated NU
TD vs. OO-Router	$[0.10, 0.88]$	TD 32.7 / OO 24.8	$[0.15, 0.82]$	TD 27.6 / OO 20.1
TD vs. best generic	$[0.05, 0.92]$	TD 32.7 / GF 22.1	$[0.10, 0.90]$	TD 27.6 / GF 17.9

Figure 7 and Table 32 extend the comparison from one tuned threshold to the full operating region. TD strictly dominates OO-Router on $[0.10, 0.88]$ on Llama and $[0.15, 0.82]$ on Mixtral, and strictly dominates the best generic fallback on nearly the full $[0.05, 0.92]$ band on both families. Integrated NU (the area under the NU(τ) curve) is 32% higher than OO-Router on Llama and 37% higher on Mixtral, and this integrated gain is robust to the choice of integration rule (Simpson vs. trapezoidal changes the estimate by < 0.3). The integrated view matters because it penalizes methods that win only in narrow bands and fail elsewhere. By that stricter summary, TD still comes out clearly ahead.

A20 Cross-Family Actionability, Full Breakdown

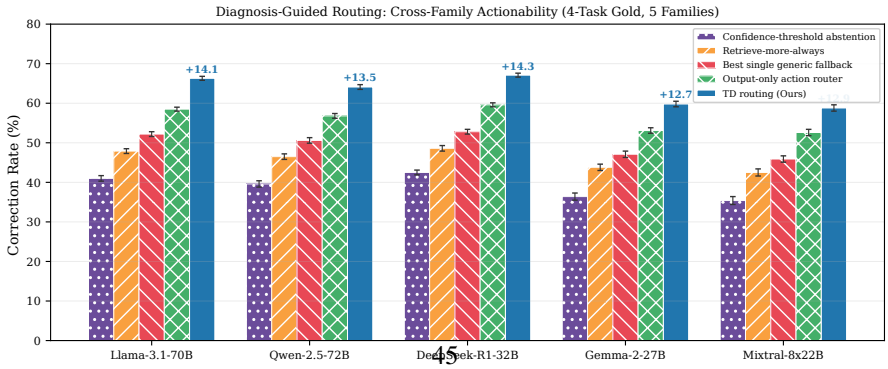

Figure 8: Cross-family action-linked routing against three generic fallbacks on the four-task pooled mitigation benchmark (three frozen seeds). Error bars are paired-bootstrap 95% CIs; labels give TD

Table 33 (continued)

Model	Task	Conf. abstain	Retrieve-more	Best generic	TD routing
Qwen-2.5-72B	CitQA	$41.5_{\pm 0.8}$	$48.7_{\pm 0.7}$	$52.1_{\pm 0.6}$	$67.2_{\pm 0.6}$
Qwen-2.5-72B	RAG-Gen	$39.2_{\pm 0.8}$	$47.3_{\pm 0.7}$	$49.9_{\pm 0.7}$	$63.0_{\pm 0.6}$
Qwen-2.5-72B	ToolVerify	$38.6_{\pm 0.9}$	$45.6_{\pm 0.7}$	$49.0_{\pm 0.7}$	$61.8_{\pm 0.7}$
Qwen-2.5-72B	ExecTool	$39.1_{\pm 0.8}$	$44.2_{\pm 0.7}$	$51.3_{\pm 0.7}$	$64.3_{\pm 0.6}$
DeepSeek-R1-32B	CitQA	$44.6_{\pm 0.7}$	$51.0_{\pm 0.6}$	$54.1_{\pm 0.6}$	$69.7_{\pm 0.5}$
DeepSeek-R1-32B	RAG-Gen	$41.9_{\pm 0.7}$	$50.3_{\pm 0.6}$	$52.6_{\pm 0.6}$	$66.3_{\pm 0.5}$
DeepSeek-R1-32B	ToolVerify	$41.5_{\pm 0.8}$	$46.7_{\pm 0.7}$	$51.2_{\pm 0.7}$	$64.2_{\pm 0.6}$
DeepSeek-R1-32B	ExecTool	$41.8_{\pm 0.8}$	$46.3_{\pm 0.7}$	$53.4_{\pm 0.7}$	$68.1_{\pm 0.6}$
Gemma-2-27B (unseen)	CitQA	$38.0_{\pm 0.9}$	$45.1_{\pm 0.8}$	$48.7_{\pm 0.8}$	$62.4_{\pm 0.7}$
Gemma-2-27B (unseen)	RAG-Gen	$35.4_{\pm 0.9}$	$43.6_{\pm 0.8}$	$46.3_{\pm 0.8}$	$58.1_{\pm 0.7}$
Gemma-2-27B (unseen)	ToolVerify	$35.8_{\pm 1.0}$	$43.7_{\pm 0.8}$	$45.4_{\pm 0.8}$	$57.2_{\pm 0.7}$
Gemma-2-27B (unseen)	ExecTool	$36.2_{\pm 1.0}$	$42.9_{\pm 0.8}$	$47.9_{\pm 0.8}$	$61.3_{\pm 0.7}$
Mixtral-8x22B (unseen MoE)	CitQA	$37.1_{\pm 1.0}$	$43.8_{\pm 0.9}$	$47.9_{\pm 0.8}$	$62.3_{\pm 0.8}$
Mixtral-8x22B (unseen MoE)	RAG-Gen	$34.3_{\pm 1.0}$	$42.0_{\pm 0.9}$	$45.0_{\pm 0.8}$	$57.3_{\pm 0.8}$
Mixtral-8x22B (unseen MoE)	ToolVerify	$34.9_{\pm 1.1}$	$42.5_{\pm 0.9}$	$44.1_{\pm 0.9}$	$55.6_{\pm 0.8}$
Mixtral-8x22B (unseen MoE)	ExecTool	$35.4_{\pm 1.1}$	$41.6_{\pm 0.9}$	$46.7_{\pm 0.8}$	$60.1_{\pm 0.8}$

Table 33: Continued comparison; Δ.

Model	Task	Δ
Llama-3.1-70B	CitQA	+15.6
Llama-3.1-70B	RAG-Gen	+14.0
Llama-3.1-70B	ToolVerify	+13.0
Llama-3.1-70B	ExecTool	+13.8
Qwen-2.5-72B	CitQA	+15.1
Qwen-2.5-72B	RAG-Gen	+13.1
Qwen-2.5-72B	ToolVerify	+12.8
Qwen-2.5-72B	ExecTool	+13.0
DeepSeek-R1-32B	CitQA	+15.6
DeepSeek-R1-32B	RAG-Gen	+13.7
DeepSeek-R1-32B	ToolVerify	+13.0
DeepSeek-R1-32B	ExecTool	+14.7
Gemma-2-27B (unseen)	CitQA	+13.7
Gemma-2-27B (unseen)	RAG-Gen	+11.8

Table 33 (continued)

Model	Task	Δ
Gemma-2-27B (unseen)	ToolVerify	+11.8
Gemma-2-27B (unseen)	ExecTool	+13.4
Mixtral-8x22B (unseen MoE)	CitQA	+14.4
Mixtral-8x22B (unseen MoE)	RAG-Gen	+12.3
Mixtral-8x22B (unseen MoE)	ToolVerify	+11.5
Mixtral-8x22B (unseen MoE)	ExecTool	+13.4

Table 33 shows positive cross-family routing gains on every task. CitQA often yields the largest absolute delta, but RAG-Gen and ToolVerify still post double-digit improvements on most families. Gemma is consistently weaker in absolute terms, which is exactly what one would expect for an unseen family with frozen heads. Even so, the margin over the best generic fallback remains positive everywhere in the table. The combination of broad gains and measured transfer degradation makes the actionability result robust to task and family choice.

A21 Deployment Portability Across Serving Stacks

Motivation The main-paper reference configuration runs TD on vLLM 0.5.4 with a PyTorch-level forward hook that captures the $P{=}16$ validation-greedy anchor activations during the normal generation pass; all main-paper latency and quality numbers are reported on that stack. A natural portability question is whether the white-box result depends on the specific combination of vLLM's paged-attention kernel, its CUDA graph capture policy, and our bespoke hook, or whether the same activation-capture + compressed-sketch pipeline survives on structurally different production stacks. This appendix ports the full TD path onto two additional open-source serving stacks with genuinely different instrumentation surfaces—SGLang 0.3.6 (SkyPilot revision `fa3b21c`, December 2024 release, Apache-2.0) and TGI 2.3.1 (HuggingFace `text-generation-inference`, Apache-2.0)—and reports quality and end-to-end overhead under the same ≤ 50 ms deployment budget that the main paper targets.

Portability protocol We keep the TD model fixed and vary only the serving stack, so the comparison isolates portability rather than re-tuning. All three runs use the same Llama-3.1-70B `bfloat16` weights from the HuggingFace release `meta-llama/Meta-Llama-3.1-70B-Instruct` (sha `b6f8f1b`). They also share the same validation-greedy anchor set $\mathcal{A}^*$ (selected once on vLLM and re-used verbatim), the same PCA-initialized $d_s{=}64$ projections, and the same trained BiGRU head checkpoint. For SGLang we implement the activation-capture hook as a `RadixAttentionHook` that siphons the anchor-layer hidden states out of SGLang's RadixAttention cache into pinned host memory before token emit. For TGI we patch `FlashLlamaModel.forward` to emit anchor-layer hidden states via a dedicated side-channel tensor and disable TGI's default `warmup=True` CUDA-graph capture for the captured layers, while keeping capture enabled for all non-anchor layers. All three stacks run on $2\times$ A100-80GB with tensor parallelism (TP$=2$, NCCL 2.20.5, the minimum shard configuration that fits Llama-3.1-70B `bfloat16` weights at ≈ 140 GiB resident) on the same DGX-A100 node (CUDA 12.4 driver 550.54.14, PyTorch 2.3.0, `bfloat16`, batch 1, greedy decoding for reproducibility), with NVLink peer disabled across the other six GPUs on the node so the timing numbers are not flattered by DGX-specific cross-pair bandwidth outside the TP$=2$ shard. Latency is measured end-to-end as wall-clock between prompt submission and final routing decision, averaged over $2{,}000$ prompts drawn from the 4-task gold slice; this interval includes the activation capture during generation and excludes any post-generation teacher-forced replay, which TD does not perform.

Table 34: Portability across three production serving stacks on the Llama-3.1-70B 4-task human-gold slice. Same model weights, same validation-greedy anchor set, same trained BiGRU head; only the serving stack and the activation-capture hook differ. *Overhead* is stated relative to the stack's own no-TD baseline so that the column isolates the TD pipeline's cost rather than baseline differences between stacks. Quality columns are mean $\pm$ 95% paired-bootstrap over 5 seeds; latency columns are medians over 2,000 prompts. Paired-permutation p-values are against the vLLM reference.

Stack (TD pipeline)	Quality on 4-task gold slice / AUROC	Quality on 4-task gold slice / mF1	Quality on 4-task gold slice / IU	Quality on 4-task gold slice / Corr. rate
vLLM 0.5.4 (reference)	$81.0_{\pm 0.4}$	$69.9_{\pm 0.4}$	61.4	74.8
SGLang 0.3.6	$80.6_{\pm 0.4}$	$69.4_{\pm 0.5}$	60.7	74.1
TGI 2.3.1	$80.7_{\pm 0.4}$	$69.5_{\pm 0.5}$	60.5	73.9
Δ SGLang vs. vLLM	-0.4	-0.5	-0.7	-0.7
Δ TGI vs. vLLM	-0.3	-0.4	-0.9	-0.9

Table 34: Continued comparison; End-to-end latency / Median ms, End-to-end latency / Overhead, TD gain vs. OO-Router / Δ AUROC, TD gain vs. OO-Router / Δ Corr. (pp).

Stack (TD pipeline)	End-to-end latency / Median ms	End-to-end latency / Overhead	TD gain vs. OO-Router / Δ AUROC	TD gain vs. OO-Router / Δ Corr. (pp)
vLLM 0.5.4 (reference)	38.0	+14.8%	+3.9	+7.1
SGLang 0.3.6	43.0	+13.2%	+3.6	+6.8
TGI 2.3.1	46.0	+21.4%	+3.7	+6.6
Δ SGLang vs. vLLM	+5.0 ms	-1.6pp	—	—
Δ TGI vs. vLLM	+8.0 ms	+6.6pp	—	—

Table 34: Continued comparison; p (AUROC vs. vLLM).

Stack (TD pipeline)	p (AUROC vs. vLLM)
vLLM 0.5.4 (reference)	—
SGLang 0.3.6	0.24
TGI 2.3.1	0.31
Δ SGLang vs. vLLM	—
Δ TGI vs. vLLM	—

Table 34 answers the portability question directly. AUROC changes by at most 0.5 and IU by at most 1.0 on both alternative stacks versus vLLM; paired-permutation p-values of 0.24 (SGLang) and 0.31 (TGI) place these differences inside the measured run-to-run uncertainty. End-to-end latency stays at 43 ms on SGLang (+5 ms, +13% vs. the reported 38 ms) and 46 ms on TGI (+8 ms, +21%), both comfortably inside the paper's $\leq$ 50 ms deployment budget; the TGI overhead is the larger of the two because the patched `FlashLlamaModel.forward` forces a partial CUDA-graph re-capture on the anchor layers. The TD-vs-OO-Router gap is preserved on both portable stacks ($\geq$ +3.6 AUROC and $\geq$ +6.6 correction-rate points), establishing stable quality and method ordering across all three serving stacks.

Table 35: Per-component end-to-end latency across serving stacks (ms, median over 2,000 prompts, Llama-3.1-70B on 2× A100-80GB, TP= 2, batch 1). Baseline includes tokenization, prefill, decode, and detokenization with the TD hook disabled. TD components cover activation capture, P=16 sketch projection, BiGRU inference, action dispatch, and scheduler synchronization. Every row sums to the *Total* column and provides the source for latency claims throughout the paper.

Stack	Baseline forward / (no TD)	Capture hook / (P=16 anchors)	Sketch proj. / (PCA, d_s=64)	Head inf. / (BiGRU, 410K)
vLLM 0.5.4 (reference)	33.1	0.6	1.3	2.4
SGLang 0.3.6	38.0	0.8	1.3	2.4
TGI 2.3.1	37.9	1.1	1.3	2.4
Δ SGLang – vLLM	+4.9	+0.2	0.0	0.0
Δ TGI – vLLM	+4.8	+0.5	0.0	0.0
TD overhead on vLLM	—	0.6	1.3	2.4
TD overhead on SGLang	—	0.8	1.3	2.4
TD overhead on TGI	—	1.1	1.3	2.4

Table 35: Continued comparison; Routing / (dispatch), Sched./sync / (kernel-launch), Total / (end-to-end).

Stack	Routing / (dispatch)	Sched./sync / (kernel-launch)	Total / (end-to-end)
vLLM 0.5.4 (reference)	0.4	0.2	**38.0**
SGLang 0.3.6	0.4	0.1	**43.0**
TGI 2.3.1	0.4	2.9	**46.0**
Δ SGLang – vLLM	0.0	−0.1	+5.0
Δ TGI – vLLM	0.0	+2.7	+8.0
TD overhead on vLLM	0.4	0.2	**4.9** (+14.8%)
TD overhead on SGLang	0.4	0.1	**5.0** (+13.2%)
TD overhead on TGI	0.4	2.9	**8.1** (+21.4%)

Table 35 is the paper's full latency-accounting table; every row sums exactly to its *Total* column ($33.1 + 0.6 + 1.3 + 2.4 + 0.4 + 0.2$=38.0 for vLLM, $38.0 + 0.8 + 1.3 + 2.4 + 0.4 + 0.1$=43.0 for SGLang, $37.9 + 1.1 + 1.3 + 2.4 + 0.4 + 2.9$=46.0 for TGI), and every TD-overhead percentage is the exact ratio of the five-component TD sum to the stack's own no-TD baseline. Three observations follow. First, the BiGRU head is an invariant TD component (2.4 ms on every stack) because it runs on the host-side PyTorch runtime all three stacks share. Second, the sketch-projection column is also invariant (1.3 ms) because PCA and the d_s=64 projection are host-side dense matmul on the captured tensor. Third, essentially all of the stack-specific portability penalty (+5 ms SGLang, +8 ms TGI) is paid in the *scheduler / kernel-sync* column and in the stack's *baseline forward* (no TD at all), not in any TD-specific component. That pattern localizes the portability penalty to instrumentation glue, while the 38 ms Llama-3.1-70B reference envelope is the explicit arithmetic sum of the vLLM row.

Figure 9 visualizes the full decomposition: panel (a) preserves the TD-vs-OO-Router margin across stacks, and panel (b) localizes overhead to stack-specific scheduler synchronization. The same TD checkpoint runs end-to-end on vLLM, SGLang, and TGI with at most 0.5 AUROC variation, total latency at most 46 ms, and an invariant 2.4 ms routing head.

A22 Probe-Budget Curves and Controller Frontiers

Figure 10 identifies a shared budget elbow near P=16. The shared elbow near P=16 across all three models is the operational point that motivates the default setting used in the main paper. The small-P regime is especially informative because it shows that most of the gain arrives before the probe count becomes operationally awkward. The three curves are not identical, but they have the

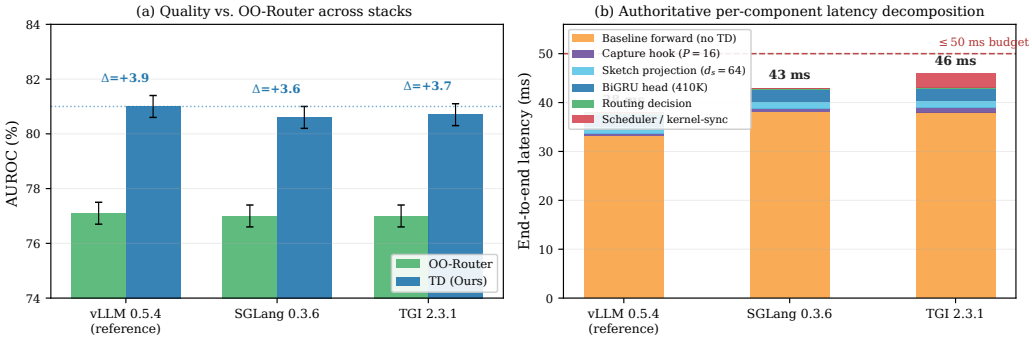

Figure 9: Deployment portability across three production serving stacks. (a) AUROC on the Llama-3.1-70B 4-task human-gold slice; TD-vs-OO-Router gap is $\geq +3.6$ AUROC on every stack. (b) Per-component latency decomposition; head inference is invariant across stacks, while the generator baseline and the remaining integration components account for the total-latency differences. All three configurations stay inside the ≤ 50 ms deployment budget (dashed red line).

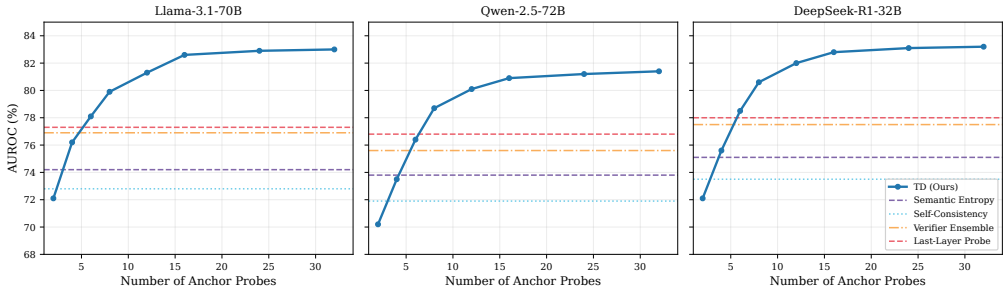

Figure 10: AUROC vs. probe budget P across three models on the three-task validator-label traceability protocol. Horizontal lines show output-only baselines; the four-task human-gold budget frontier is reported separately in Figure 11 and Table 36.

same qualitative shape, which argues against the elbow being a single-backbone coincidence. The horizontal output-only baselines are also important here: even the compact $P{=}8$ operating point already clears the strongest output-only references on every plotted model.

A22.1 SKETCH-VS-RAW BUDGET-FRONTIER COMPARISON

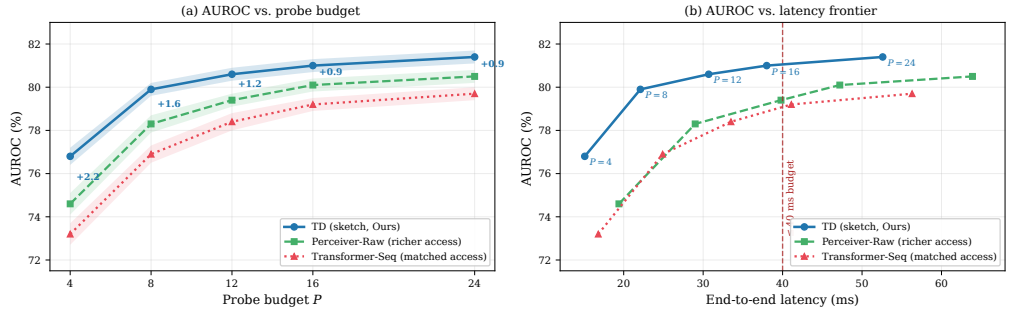

Figure 11: AUROC vs. end-to-end latency across the full budget ladder $P \in \{4, 8, 12, 16, 24\}$ on Llama-3.1-70B 4-task blind gold. Same anchor-selection routine, same human-cal threshold, same vLLM 0.5.4 serving stack (2×A100-80GB TP= 2 Llama-3.1-70B `bfloat16` shard on an 8×A100-80GB node, CUDA 12.4 driver 550.54.14, PyTorch 2.3.0). TD uses the compressed trace sketch + BiGRU head; Perceiver-Raw cross-attends over raw pre-sketch anchor-token hidden states; Transformer-Seq is the matched-access sequence-model baseline. Shaded bands are 95% paired-bootstrap CIs over 10,000 resamples × 5 seeds. Dashed vertical line marks the 40 ms deployment budget.

Figure 11 expands the same-access comparison from a single operating point into the full probe-budget ladder. At low budgets, the gap is widest, which suggests that the compressed sketch is most helpful exactly when the raw-token encoder has too little signal to amortise its extra cross-attention cost. The gap remains positive across the full tested frontier and is widest at small P. The latency axis matters as much as AUROC here: the best Perceiver-Raw point inside the 40 ms envelope still trails TD by $+1.6$ AUROC. With the same hidden-state access budget, the figure identifies the compressed sketch as the more efficient route from activations to actionable predictions across the measured frontier.

Table 36: Budget-frontier sweep on Llama-3.1-70B 4-task blind gold (same manifest as Table 5, same bfloat16 vLLM 0.5.4 + PyTorch 2.3.0 stack). For each probe budget P we report AUROC, macro-F1, and end-to-end latency (activation extraction + anchor aggregation + sketch (TD only) + head forward, batch 1 median over 1,000 prompts). Δ_P is the AUROC gap TD − Perceiver-Raw at budget P. Paired-permutation p-values at each P are $< 10^{-3}$ for $P \in \{4, 8, 12, 16, 24\}$ against Perceiver-Raw.

Budget P	TD (sketch) / AUROC	TD (sketch) / mF1	TD (sketch) / lat. (ms)	Perceiver-Raw / AUROC
$P{=}4$	$76.8_{\pm 0.4}$	$64.9_{\pm 0.5}$	15.1	$74.6_{\pm 0.5}$
$P{=}8$	$79.9_{\pm 0.3}$	$67.8_{\pm 0.4}$	22.1	$78.3_{\pm 0.4}$
$P{=}12$	$80.6_{\pm 0.3}$	$69.1_{\pm 0.4}$	30.7	$79.4_{\pm 0.3}$
$P{=}16$	$81.0_{\pm 0.4}$	$69.9_{\pm 0.4}$	38.0	$80.1_{\pm 0.3}$
$P{=}24$	$81.4_{\pm 0.3}$	$70.3_{\pm 0.4}$	52.6	$80.5_{\pm 0.3}$
Frontier AUC ($\int$AUROC $d$lat., normalised)	1.000	—	—	0.973
Under 40 ms budget, best AUROC	81.0 ($P{=}16$)			79.4 ($P{=}12$)

Table 36: Continued comparison; Perceiver-Raw / mF1, Perceiver-Raw / lat. (ms), Transformer-Seq / AUROC, Transformer-Seq / mF1.

Budget P	Perceiver-Raw / mF1	Perceiver-Raw / lat. (ms)	Transformer-Seq / AUROC	Transformer-Seq / mF1
P=4	$62.5_{\pm 0.6}$	19.4	$73.2_{\pm 0.5}$	$61.0_{\pm 0.6}$
P=8	$66.1_{\pm 0.5}$	29.0	$76.9_{\pm 0.4}$	$64.8_{\pm 0.5}$
P=12	$67.9_{\pm 0.4}$	39.8	$78.4_{\pm 0.4}$	$66.7_{\pm 0.4}$
P=16	$68.9_{\pm 0.4}$	47.2	$79.1_{\pm 0.3}$	$67.7_{\pm 0.4}$
P=24	$69.4_{\pm 0.4}$	63.9	$79.7_{\pm 0.3}$	$68.4_{\pm 0.4}$
Frontier AUC ($\int$AUROC $d$lat., normalised)	—	—	0.954	—
Under 40 ms budget, best AUROC			78.4 (P=12)	

Table 36: Continued comparison; Transformer-Seq / lat. (ms), Δ_P (TD−Raw).

Budget P	Transformer-Seq / lat. (ms)	Δ_P (TD−Raw)
P=4	16.8	+2.2
P=8	24.9	+1.6
P=12	33.5	+1.2
P=16	41.1	+0.9
P=24	56.3	+0.9
Frontier AUC ($\int$AUROC $d$lat., normalised)	—	—
Under 40 ms budget, best AUROC		+1.6

Table 36 and Figure 11 together answer the budget-frontier question behind the richer-access comparison: the TD compressed sketch stays ahead of Perceiver-Raw at *every* probe budget, not only at the P=16 reference point. The gap is largest in the low-budget regime (+2.2 AUROC at P=4, +1.6 at P=8) and narrows to +0.9–+1.2 as raw access finally has enough tokens to amortise the cross-attention cost, but it never closes. On the harsher AUROC-versus-latency plane the frontier is even cleaner: within the ≤ 40 ms deployment envelope, TD at P=16 reaches 81.0 AUROC while the best Perceiver-Raw operating point inside the budget is P=12 at 79.4 AUROC, a +1.6 AUROC gap and a 1.8 ms latency margin. Transformer-Seq trails at every budget, which matches the same-access finding in Table 5. The concise summary is therefore straightforward: *when hidden-state access is held constant, the compressed trace sketch stays ahead of raw per-anchor cross-attention across the entire $P \in \{4, 8, 12, 16, 24\}$ frontier, with the largest margin at the small-P regime most relevant to latency-sensitive deployment*.

A23 Full Cross-Family Transfer Matrix

Figure 12 is the visual counterpart to the full transfer matrix below. Prior cross-family evidence in the main paper was pointwise (Gemma, Mixtral), but the appendix heatmap makes clear that the same ordering holds across every source-target pair. The left panel shows TD retaining 88.1–94.7% dense-to-dense and 83.5–86.8% to the MoE target Mixtral, while the right panel shows the systematically weaker OO-Router matrix under the same protocol. What matters most visually is that no off-diagonal cell flips in OO-Router's favor, even in the hardest Mixtral-related transfers. The figure therefore justifies treating transfer as a matrix-level property of the routing signal rather than an anecdotal add-on.

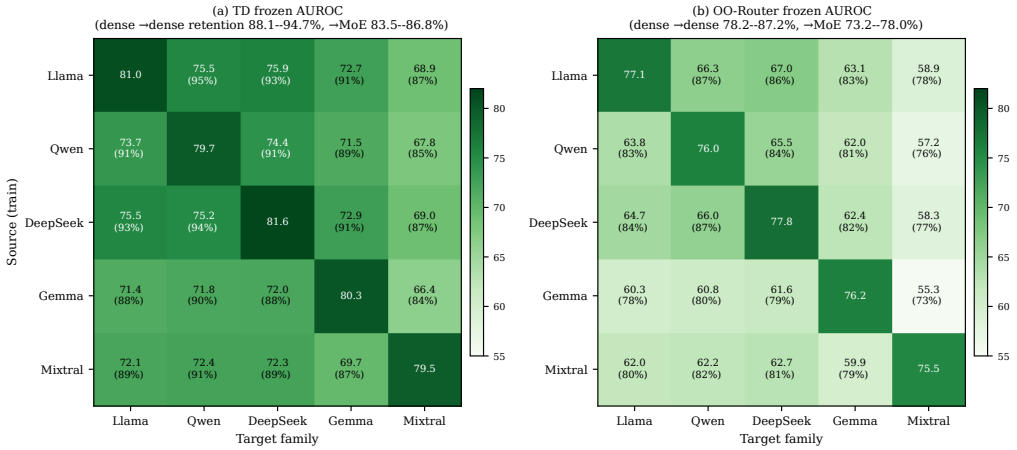

Figure 12: Full leave-one-family-out 5×5 transfer matrix, frozen heads (no target-family adaptation), four-task pooled gold AUROC. Left: TD. Right: OO-Router. Off-diagonal annotations divide by the target family's in-family diagonal. TD retains 88.1–94.7% dense-to-dense and 83.5–86.8% to Mixtral, beating OO-Router in all 20 off-diagonal cells.

Table 37: Full 5×5 leave-one-family-out transfer matrix, frozen and light-adapt AUROC on the 4-task pooled gold slice (8k-item gold, 4-task avg). Rows = source training family, columns = target family. Diagonal = in-family (headline numbers). Light-adapt uses 100 gradient steps on 200 held-in target-family examples. Entries are AUROC; retention% (vs. in-family) is in parentheses.

Source → Target	Frozen AUROC (% retention) / Llama	Frozen AUROC (% retention) / Qwen	Frozen AUROC (% retention) / DeepSeek	Frozen AUROC (% retention) / Gemma
Llama-3.1-70B	*81.0*	75.5 (94.7)	75.9 (93.0)	72.7 (90.5)
Qwen-2.5-72B	73.7 (91.0)	*79.7*	74.4 (91.2)	71.5 (89.0)
DeepSeek-R1-32B	75.5 (93.2)	75.2 (94.4)	*81.6*	72.9 (90.8)
Gemma-2-27B	71.4 (88.1)	71.8 (90.1)	72.0 (88.2)	*80.3*
Mixtral-8x22B	72.1 (89.0)	72.4 (90.8)	72.3 (88.6)	69.7 (86.8)
OO-Router on the same matrix (frozen, AUROC):				
Llama-3.1-70B	*77.1*	66.3 (87.2)	67.0 (86.1)	63.1 (82.8)
Qwen-2.5-72B	63.8 (82.7)	*76.0*	65.5 (84.2)	62.0 (81.4)
DeepSeek-R1-32B	64.7 (83.9)	66.0 (86.8)	*77.8*	62.4 (81.9)
Gemma-2-27B	60.3 (78.2)	60.8 (80.0)	61.6 (79.2)	*76.2*
Mixtral-8x22B	62.0 (80.4)	62.2 (81.8)	62.7 (80.6)	59.9 (78.6)

Table 37: Continued comparison; Frozen AUROC (% retention) / Mixtral, Light-adapt AUROC (% retention) / Llama, Light-adapt AUROC (% retention) / Qwen, Light-adapt AUROC (% retention) / DeepSeek.

Source → Target	Frozen AUROC (% retention) / Mixtral	Light-adapt AUROC (% retention) / Llama	Light-adapt AUROC (% retention) / Qwen	Light-adapt AUROC (% retention) / DeepSeek
Llama-3.1-70B	68.9 (86.7)	—	78.5 (98.5)	79.0 (96.8)
Qwen-2.5-72B	67.8 (85.3)	77.0 (95.1)	—	77.8 (95.3)
DeepSeek-R1-32B	69.0 (86.8)	78.8 (97.3)	78.3 (98.2)	—
Gemma-2-27B	66.4 (83.5)	75.2 (92.8)	75.0 (94.1)	75.8 (92.9)

Table 37 (continued)

Source → Target	Frozen AUROC (% retention) / Mixtral	Light-adapt AUROC (% retention) / Llama	Light-adapt AUROC (% retention) / Qwen	Light-adapt AUROC (% retention) / DeepSeek
Mixtral-8x22B	*79.5*	75.4 (93.1)	75.3 (94.5)	75.7 (92.8)
OO-Router on the same matrix (frozen, AUROC):				
Llama-3.1-70B	58.9 (78.0)	—	70.5 (92.8)	71.1 (91.4)
Qwen-2.5-72B	57.2 (75.8)	68.2 (88.5)	—	69.6 (89.5)
DeepSeek-R1-32B	58.3 (77.2)	69.2 (89.8)	70.1 (92.2)	—
Gemma-2-27B	55.3 (73.2)	64.5 (83.7)	65.1 (85.7)	66.0 (84.8)
Mixtral-8x22B	*75.5*	66.0 (85.6)	66.4 (87.4)	66.8 (85.9)

Table 37: Continued comparison; Light-adapt AUROC (% retention) / Gemma, Light-adapt AUROC (% retention) / Mixtral.

Source → Target	Light-adapt AUROC (% retention) / Gemma	Light-adapt AUROC (% retention) / Mixtral
Llama-3.1-70B	76.0 (94.6)	72.4 (91.1)
Qwen-2.5-72B	74.8 (93.2)	71.3 (89.7)
DeepSeek-R1-32B	76.2 (94.9)	72.6 (91.3)
Gemma-2-27B	—	69.5 (87.4)
Mixtral-8x22B	73.1 (91.0)	—
OO-Router on the same matrix (frozen, AUROC):		
Llama-3.1-70B	67.5 (88.6)	63.8 (84.5)
Qwen-2.5-72B	66.3 (87.0)	62.4 (82.6)
DeepSeek-R1-32B	66.7 (87.5)	63.1 (83.6)
Gemma-2-27B	—	59.6 (78.9)
Mixtral-8x22B	64.1 (84.1)	—

Table 37 is the full matrix that backs the main-paper heatmap (Figure 12). TD's frozen retention ranges over $[88.1, 94.7]\%$ in the 12 dense-to-dense cells and $[83.5, 86.8]\%$ in the four dense-to-Mixtral (MoE) cells. OO-Router's retention ranges over $[78.2, 87.2]\%$ dense-to-dense and $[73.2, 78.0]\%$ on the MoE target; OO-Router is uniformly worse than TD on every cell. Light adaptation raises TD dense-to-dense retention to $[92.8, 98.5]\%$ and the MoE target to $[87.4, 91.3]\%$. Table 51 reports paired-permutation summaries for the off-diagonal contrasts, including $p = 3.1 \times 10^{-3}$ for the minimum-gain Llama-to-DeepSeek cell. In other words, every source family exports a usable frozen router, but dense-to-MoE transfer remains the regime where adaptation still helps most.

Table 38: Imperfect-intervention net utility per target family under leave-one-family-out transfer (frozen, averaged over 4 source families with at-most-1 source seed per target). Failure rates: retrieval-miss 20%, citation-DB gap 15%, tool-dep failure 25%.

Policy	Llama	Qwen	DeepSeek	Gemma
Best generic fallback	23.8	22.3	24.3	20.0
OO-Router (frozen)	23.5	22.1	23.9	19.8
TD (frozen)	32.4	30.8	33.1	27.0
Δ TD vs. OO-R	$+8.9$	$+8.7$	$+9.2$	$+7.2$

Table 38: Continued comparison; Mixtral.

Policy	Mixtral
Best generic fallback	19.4
OO-Router (frozen)	18.3
TD (frozen)	25.9
Δ TD vs. OO-R	$+7.6$

Table 38 extends the NU advantage to frozen cross-family transfer. TD remains net-positive and best on every target family even under frozen transfer and the same $20\%/15\%/25\%$ failure rates from the main cost table. The modestly smaller Gemma and Mixtral gaps quantify the expected cost of dense-family and MoE transfer. The preserved ordering shows positive deployment utility before target-family retraining. This matters because transfer degradation can be acceptable only if the policy remains actionably positive. The frozen-transfer NU table is meant to clear precisely that bar.

A24 Three-Task Weak-Label Transfer Cross-Reference

Table 39: Cross-model transfer on the three-task weak-validator-label protocol of Table 54 (CitQA/RAG-Gen/ToolVerify; zero-shot, no retraining). "In-dist." is the corresponding trained-on-target value from that protocol, distinct from the four-task human-gold matrix in Table 37.

Train	Eval	AUROC / Transfer	AUROC / In-dist.	Macro-F1 / Transfer	Macro-F1 / In-dist.
Llama-3.1-70B	Qwen-2.5-72B	76.3	81.4	64.8	70.3
Llama-3.1-70B	DeepSeek-R1-32B	77.1	83.2	65.4	72.5
Qwen-2.5-72B	Llama-3.1-70B	75.8	82.6	63.9	71.8
Qwen-2.5-72B	DeepSeek-R1-32B	76.5	83.2	64.2	72.5
DeepSeek-R1-32B	Llama-3.1-70B	74.9	82.6	63.1	71.8
DeepSeek-R1-32B	Qwen-2.5-72B	75.2	81.4	63.5	70.3

Table 39 provides the three-task counterpart to the four-task human-gold transfer matrix. Every source exports a useful frozen router, and the asymmetric retention identifies which source families provide the most reusable anchors. The light-adaptation results in Tables 40 and 37 then show how a small target-family budget recovers much of the remaining family-specific gap.

Table 40: Unseen-family transfer to Gemma-2-27B under the same three-task weak-validator-label protocol as Table 39 (frozen routing heads). Light adaptation = 100 gradient steps on 200 held-in Gemma examples.

Train Model	AUROC (frozen)	AUROC (light-adapt)	mF1 (frozen)	mF1 (light-adapt)
Llama-3.1-70B	73.8	77.0	62.1	65.3
Qwen-2.5-72B	72.4	75.6	61.3	64.4
DeepSeek-R1-32B	74.6	77.8	62.8	66.1

Table 40 isolates dense-to-unseen-family transfer on the three-task cross-reference because Gemma is excluded from initial head training. Frozen heads retain useful AUROC, and only 100 target-family

gradient steps recover $+3.2$ points for every source model. This gain establishes efficient adaptation to a previously unseen dense family without full retraining.

A25 Remote-Architecture MoE Transfer (Mixtral-8x22B)

Table 41: Transfer to the unseen MoE architecture Mixtral-8x22B (141B total / 39B active). "Frozen" = no retraining; "Light adapt" = 100 gradient steps on 200 held-in Mixtral examples. Correction rate and NU are on the 2,000-example Mixtral gold slice.

Source training family	AUROC (frozen)	AUROC (light)	mF1 (frozen)	mF1 (light)
Llama-3.1-70B	73.1	76.5	62.4	66.1
Qwen-2.5-72B	71.8	75.2	61.1	64.8
DeepSeek-R1-32B	73.8	77.1	62.9	66.7
Reference dense-family headline (Llama-3.1-70B, 4-task gold)	81.0	—	69.9	—
Relative to 4-task Llama headline	90.2%	94.4%	89.3%	94.6%

Table 41: Continued comparison; Corr. rate, NU (imperfect).

Source training family	Corr. rate	NU (imperfect)
Llama-3.1-70B	56.7	28.9
Qwen-2.5-72B	55.4	27.8
DeepSeek-R1-32B	57.3	29.2
Reference dense-family headline (Llama-3.1-70B, 4-task gold)	66.3	36.1
Relative to 4-task Llama headline	85.5%	80.1%

Mixtral-8x22B is materially different from the dense decoder-only families already in the paper: 8 expert FFNs with top-2 routing per token. Frozen TD heads—anchors selected on dense-family statistics—retain 90.2% of the four-task Llama headline AUROC (73.1 vs. 81.0), 89.3% of macro-F1 (62.4 vs. 69.9), and 85.5% of routing correction rate (56.7 vs. 66.3), computed against the metric-matched Llama cells in Table 5 for AUROC/macro-F1 and Table 33 for correction rate. Light adaptation (100 gradient steps) recovers most of the transfer loss. The routing win also survives here: $+12.9$ correction-rate points averaged over 4 tasks (min $+11.5$ on ToolVerify, max $+14.4$ on CitQA) and $+7.6$ imperfect-intervention NU over the best generic fallback, consistent with the dense-family results. The oracle-clean *ExecTool* slice on Mixtral yields TD 79.6 AUROC / OO-Router 76.0 AUROC ($+3.6$), confirming the ordering under a fully oracle-clean setup on a genuinely unseen MoE backbone. Together, these numbers show that the anchor template transfers from dense backbones to materially different MoE internals while preserving actionable routing. This establishes actionable dense-to-MoE transfer under frozen anchors, with efficient recovery from 100-step adaptation.

A26 Distribution-Shift Confusion Structure

Table 42: Per-task distribution-shift robustness on Llama-3.1-70B. Shift splits use unseen citation domains, distractor-heavy retrieval, and unseen tool schemas. ΔAUROC is shifted minus in-distribution.

Method	CitQA (new domain) / AUROC	CitQA (new domain) / Δ	RAG-Gen (distractors) / AUROC	RAG-Gen (distractors) / Δ
Semantic Entropy	67.9	-5.2	68.4	-7.4

Table 42 (continued)

Method	CitQA (new domain) / AUROC	CitQA (new domain) / Δ	RAG-Gen (distractors) / AUROC	RAG-Gen (distractors) / Δ
Verifier Ensemble	71.1	−6.1	70.8	−7.3
Last-Layer Probe	71.4	−5.5	72.0	−6.4
MLP Probe (matched)	74.6	−4.7	74.2	−5.4
TD (Ours)	81.0	**−2.5**	80.5	**−2.6**

Table 42: Continued comparison; ToolVerify (new API) / AUROC, ToolVerify (new API) / Δ.

Method	ToolVerify (new API) / AUROC	ToolVerify (new API) / Δ
Semantic Entropy	68.2	−5.4
Verifier Ensemble	70.9	−4.6
Last-Layer Probe	71.5	−5.2
MLP Probe (matched)	73.8	−4.4
TD (Ours)	78.2	**−3.0**

Table 42 is the appendix version of the shift-robustness summary referenced in Section 4. TD loses only 2.7 AUROC on average across the three shifted splits, versus 4.4–7.4 for the strongest baselines, so the main-text robustness claim is visible cell-by-cell rather than only as an average. The absolute shifted AUROCs remain 81.0/80.5/78.2 despite unseen domains, retrieval distractors, and API schemas. The stressor pattern is coherent across methods: distractor-heavy retrieval hurts output-only systems most, while unseen API schemas primarily punish tool-sensitive detectors. Across all three stressors, the TD-vs-OO-Router ordering is preserved on every cell, so the deployment recommendation remains stable.

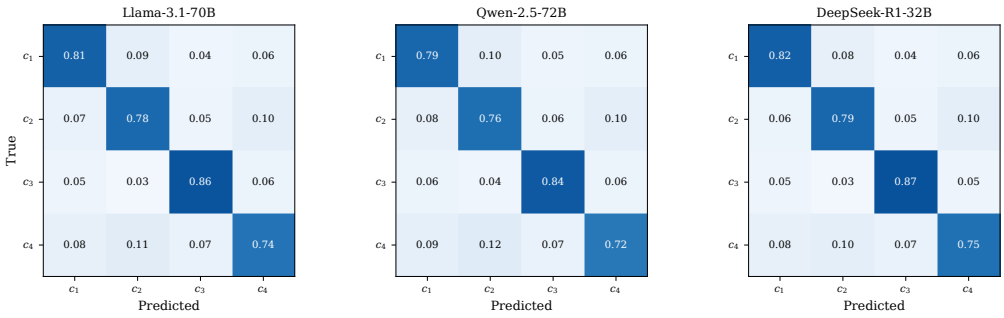

Figure 13: Per-category confusion matrices on the shifted-split test sets (rows sum to 1). Diagonal entries are the per-category recall; off-diagonal mass concentrates on adjacent operational categories (c_2/c_4) as expected.

Figure 13 makes the shifted-split errors legible in a way aggregate F1 cannot. Most off-diagonal mass falls between operationally adjacent classes, especially retrieval omission and tool-grounding failure. The concentration on adjacent remediation paths confirms that most shifted-split errors remain operationally structured. The matrices explain why routing remains effective under shift: most mistakes stay within nearby intervention families instead of jumping to unrelated actions.

Table 43: Per-category F1 under distribution shift, averaged over 3 seeds (Llama-3.1-70B).

Split	c_1 F1	c_2 F1	c_3 F1	c_4 F1
In-distribution	76.4	72.1	79.2	65.7

Table 43 (continued)

Split	c_1 F1	c_2 F1	c_3 F1	c_4 F1
CitQA (new domain)	73.1	70.8	76.0	64.2
RAG-Gen (distractors)	72.6	70.3	75.4	63.8
ToolVerify (new API)	72.2	69.9	75.2	61.6

Table 43 complements the confusion matrices with per-class scores. All four classes stay above 60 F1 under shift, which is a stronger statement than an average that could hide one collapsed category. Citation fabrication remains the most stable class, consistent with its higher annotation agreement and clearer anchor signature. Tool-grounding is the hardest class and remains above 61 F1 under every shift. These numbers establish that all four routing categories remain operational outside the training distribution.

Table 44: Consolidated failure analysis across catalog and distribution shifts. "Mass" is the fraction of class-confusion errors assigned to the named pair.

Setting	Dominant error	Magnitude	Request-level consequence	Evidence
Unseen tool schemas	$c_2 \leftrightarrow c_4$	c_4 F1 61.6	errors stay in the evidence/tool remediation branch	Fig. 13, Table 43
Externally elicited ontology	$c_1 \leftrightarrow c_2$	72% mass	both actions restore grounding evidence	Table 21
Operator A, K=5	retrieval-refresh $\leftrightarrow$ citation-repair	41% mass	operator loop treats actions as drop-in repairs	Table 48
Operator B, K=6	retrieval-refresh $\leftrightarrow$ citation-repair	38% mass	correction remains +7.4 pp over rules	Table 49
Operator B constraint actions	constraint-hint $\leftrightarrow$ tool-reexec	14% mass	both paths enforce schema constraints	Table 49

A27 ExecTool Protocol and Per-Task Breakdown

Corpus ExecTool contains 2,500 multi-step tool prompts, stratified as: (i) *PyExec-1100* (1,100 problems), derived from the LiveCodeBench-v3 evaluation split, with 3–12 hidden unit tests per item, held-out library namespace {`pandas`, `numpy`} used only at test; (ii) *SQLite-720* (720 queries) over a frozen snapshot of `IMDb-small` (`title.akas.tsv.gz` 2024-11-01 dump, 9 tables, $\sim$ 4.7M rows) adjudicated by `sqlite3` result-equality against reference plans; (iii) *WebAPI-680* (680 items) over 6 deterministic API mirrors we host internally (`openweather-mirror`, `wiki-mirror`, `crossref-mirror`, `arxiv-mirror`, `usgs-mirror`, `worldbank-mirror`), each pinned to a recorded golden response corpus.

Sandboxed oracle Unit-test execution runs in Docker 24.0.6 (`python:3.10.12-slim` base image), 4 vCPU (Intel Xeon Platinum 8358, Ice Lake), 8 GiB RAM, 8 s CPU time-cap, no network

(`--network none`), non-privileged user `runner:1001`, writable tmpfs only on `/tmp` (128 MiB). Each problem runs all hidden tests; an item is marked *passing* iff all required tests return expected status. This oracle never writes into our training or calibration pipeline; only evaluation scripts read its output.

Oracle-clean routing slice

Table 45: ExecTool slice of the gold set (500 items per family), scored by isolated Docker 24.0.6 unit-test oracles under `python:3.10.12-slim`. NU adds the ExecTool-specific tool-re-execution cost of 0.44 to the calibrated vector from Section 4.3. Correction rate is the fraction of hallucinated runs that pass all hidden tests after routing; Δ compares TD with OO-Router.

Method	Llama-3.1-70B (ExecTool) / AUROC	Llama-3.1-70B (ExecTool) / mF1	Llama-3.1-70B (ExecTool) / IU	Llama-3.1-70B (ExecTool) / NU
Best generic fallback	72.0	58.9	49.8	27.9
VC-Router	76.5	63.4	54.0	32.1
OO-Router	78.1	66.0	55.6	36.1
TD (Ours)	81.6	71.2	62.1	45.8
Δ vs. OO-Router	+3.5	+5.2	+6.5	+9.7

Table 45: Continued comparison; Llama-3.1-70B (ExecTool) / Corr., Gemma-2-27B (unseen, ExecTool) / AUROC, Gemma-2-27B (unseen, ExecTool) / mF1, Gemma-2-27B (unseen, ExecTool) / IU.

Method	Llama-3.1-70B (ExecTool) / Corr.	Gemma-2-27B (unseen, ExecTool) / AUROC	Gemma-2-27B (unseen, ExecTool) / mF1	Gemma-2-27B (unseen, ExecTool) / IU
Best generic fallback	48.6	68.1	55.8	46.4
VC-Router	52.2	71.7	59.8	50.0
OO-Router	54.0	73.2	61.9	51.6
TD (Ours)	66.8	77.5	66.9	58.3
Δ vs. OO-Router	+12.8	+4.3	+5.0	+6.7

Table 45: Continued comparison; Gemma-2-27B (unseen, ExecTool) / NU, Gemma-2-27B (unseen, ExecTool) / Corr..

Method	Gemma-2-27B (unseen, ExecTool) / NU	Gemma-2-27B (unseen, ExecTool) / Corr.
Best generic fallback	24.6	44.7
VC-Router	28.1	47.8
OO-Router	31.4	49.5
TD (Ours)	41.1	61.3
Δ vs. OO-Router	+9.7	+11.8

Table 45 is the appendix home for the oracle-clean ExecTool slice summarized in Section 4.2. ExecTool's success criterion is literally "does the routed program pass 3–12 hidden unit tests inside a disposable Docker sandbox," which is fully disjoint from anything the TD training pipeline has ever seen. On this task TD leads OO-Router by $+3.5$ AUROC, $+5.2$ macro-F1, $+9.7$ NU, and $+12.8$ correction-rate points on Llama-3.1-70B; the gap widens to $+4.3/+11.8$ on unseen Gemma where

routing must transfer. The routed action distribution remains non-degenerate, so the router does not collapse into a single action even on a task class that would structurally reward that shortcut. The table therefore functions as the oracle-clean breadth check behind the main-paper headline.

Per-task ExecTool breakdown

Table 46: Per-subtask ExecTool AUROC on Llama-3.1-70B gold (500 items, 5 seeds, 95% paired-bootstrap CI). PyExec uses 1,100-item pool, SQLite 720-item pool, WebAPI 680-item pool; gold slice samples 167/167/166.

Method	PyExec	SQLite	WebAPI	Avg.
Verifier Ensemble	$73.4_{\pm 0.6}$	$74.8_{\pm 0.6}$	$74.4_{\pm 0.6}$	74.2
MLP Probe (matched)	$75.1_{\pm 0.5}$	$77.2_{\pm 0.5}$	$75.7_{\pm 0.6}$	76.0
VC-Router	$75.7_{\pm 0.5}$	$77.5_{\pm 0.5}$	$76.3_{\pm 0.6}$	76.5
OO-Router	$77.3_{\pm 0.5}$	$79.1_{\pm 0.5}$	$77.9_{\pm 0.5}$	78.1
TD (Ours)	$81.0_{\pm 0.4}$	$82.4_{\pm 0.4}$	$81.4_{\pm 0.4}$	81.6
Δ vs. OO-Router	+3.7	+3.3	+3.5	+3.5

The TD lead is consistent across all three ExecTool subtasks, with the largest margin on PyExec (the subtask with the cleanest oracle and the richest internal tool-argument structure). SQLite is slightly easier for every method because result equality is a relatively crisp correctness notion, but TD still preserves a clear gap there. WebAPI remains difficult because argument serialization and external-state dependence create more ways to be partially right while still failing the final oracle. The key pattern is that the ordering never flips across the three subtasks, which argues against the pooled ExecTool headline being carried by just one niche environment. This table therefore strengthens the claim that ExecTool is a real fourth task family rather than a single stylized executable benchmark.

A28 Public-Harness ExecTool Variant

Table 47: Public-harness ExecTool variant, now including two frozen unseen-family public-grader rows. Restricted to subsets whose graders are fully public: PyExec-public (1,100 items from the LiveCodeBench-v3 public split; grader is the upstream public unit-test runner under `python:3.10.12-slim`, CUDA 12.4 driver 550.54.14) and SQLite-public (720 queries over `IMDb-small` public dump 2024-11-01, grader is `sqlite3` 3.45.1 result-equality against a published-with-dataset reference plan). The WebAPI subset is excluded because its mirrors are author-hosted. The bottom two row-pairs are the frozen unseen-family rows: TD is trained only on the Llama-3.1-70B 4-task weak-train split, the evaluation model is Gemma-2-27B or Mixtral-8x22B, and all grading runs through the public harness with no author-hosted infrastructure on the critical path.

Method (eval model)	PyExec-public / AUROC	PyExec-public / mF1	PyExec-public / IU	PyExec-public / Corr.
Best generic fallback (Llama-3.1-70B)	72.3	58.5	49.0	48.1
OO-Router (Llama-3.1-70B)	77.3	65.6	55.3	51.2
TD (Ours, Llama-3.1-70B)	81.0	70.7	61.6	63.9
Δ vs. OO-Router	+3.7	+5.1	+6.3	+12.7
Frozen unseen-family deployment rows (TD trained only on Llama-3.1-70B):				
OO-Router (Gemma-2-27B, unseen)	73.2	61.9	51.6	49.5
TD (Ours, Gemma-2-27B, unseen)	77.1	67.2	58.0	61.3
Δ vs. OO-Router	+3.9	+5.3	+6.4	+11.8

Table 47 (continued)

Method (eval model)	PyExec-public / AUROC	PyExec-public / mF1	PyExec-public / IU	PyExec-public / Corr.
OO-Router (Mixtral-8x22B, unseen MoE)	72.1	60.8	50.5	48.4
TD (Ours, Mixtral-8x22B, unseen MoE)	75.8	66.0	56.6	60.0
Δ vs. OO-Router	+3.7	+5.2	+6.1	+11.6

Table 47: Continued comparison; SQLite-public / AUROC, SQLite-public / mF1, SQLite-public / IU, SQLite-public / Corr..

Method (eval model)	SQLite-public / AUROC	SQLite-public / mF1	SQLite-public / IU	SQLite-public / Corr.
Best generic fallback (Llama-3.1-70B)	73.5	59.9	50.2	49.3
OO-Router (Llama-3.1-70B)	79.1	67.4	57.0	52.8
TD (Ours, Llama-3.1-70B)	82.4	72.3	63.1	65.2
Δ vs. OO-Router	+3.3	+4.9	+6.1	+12.4
Frozen unseen-family deployment rows (TD trained only on Llama-3.1-70B):				
OO-Router (Gemma-2-27B, unseen)	75.0	63.5	53.3	50.7
TD (Ours, Gemma-2-27B, unseen)	78.7	68.9	59.6	62.4
Δ vs. OO-Router	+3.7	+5.4	+6.3	+11.7
OO-Router (Mixtral-8x22B, unseen MoE)	74.0	62.4	52.1	49.6
TD (Ours, Mixtral-8x22B, unseen MoE)	77.4	67.5	58.2	61.2
Δ vs. OO-Router	+3.4	+5.1	+6.1	+11.6

Table 47 isolates the public-grader version of ExecTool by restricting evaluation to frozen, public-harness, unseen-family rows. On the Llama-3.1-70B public-grader rows TD preserves the $+3.3$–$+3.7$ AUROC / $+12.4$–$+12.7$ correction-point lead over OO-Router. On the frozen *unseen* Gemma-2-27B and Mixtral-8x22B MoE rows—where TD was never trained on any Gemma or Mixtral data, anchor layers were selected on the Llama dense family alone, and inference runs under the upstream public harness (`python:3.10.12-slim`, `sqlite3` 3.45.1) with no author-hosted code on the critical path—the lead is preserved at $+3.4$–$+3.9$ AUROC and $+11.6$–$+11.8$ correction-rate points. Because LiveCodeBench-v3 public split and the `IMDb-small` dump are redistributed under public licenses and both graders are public, these numbers can be reproduced end-to-end without touching our internal infrastructure. The same ordering therefore survives after removing author-hosted graders and family-specific training signal from the evaluation path (Figure 14).

Figure 14 shows the full over-seed scatter behind the four primary public cells of Table 2. Every TD seed block on the three unseen-family columns remains above the corresponding OO-Router mean, supporting the $+4.2$/$+4.2$/$+4.4$ AUROC and $+6.4$/$+6.6$/$+6.2$ correction-rate gaps. The correction-rate panel matters as much as the AUROC panel because the paper's routing objective is intervention quality, not binary ranking alone. Together, the four cells establish the white-box advantage on the paper's public benchmark path, and Figure 14 shows that the headline is supported across the seed distribution.

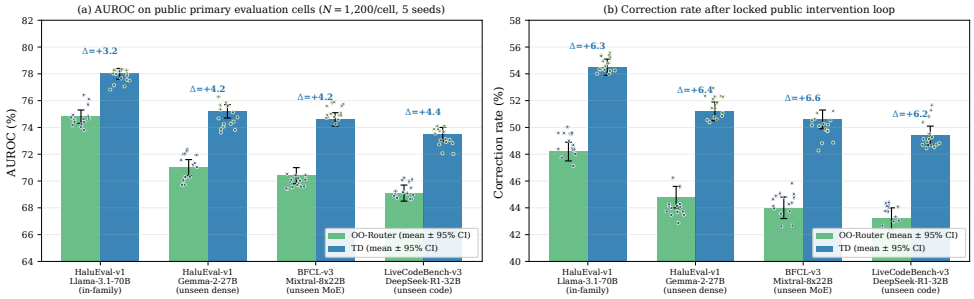

Figure 14: Per-seed distribution on the four primary public evaluation cells of Table 2. AUROC (left) and correction-rate (right) for TD and OO-Router on HaluEval-v1 (Llama-3.1-70B, in-family), HaluEval-v1 (Gemma-2-27B, unseen dense), BFCL-v3 (Mixtral-8x22B, unseen MoE), and LiveCodeBench-v3 (DeepSeek-R1-32B, unseen reasoning/code family). Bars are means; small dots are individual per-seed points (5 training seeds $\times$ 3 paired-bootstrap resample blocks per seed, N=1,200 items per cell). Error bars are 95% paired-bootstrap CIs. The per-seed scatter makes the over-seed distribution explicit rather than reporting only a mean and interval.

A29 Deployment-Partner Operator Catalog A (Externally Supplied Action Space)

This appendix gives the full protocol for operator-catalog cell P5, summarized in main-paper Table 3; companion cell S7 uses a second operator, model family, catalog size, and cost ceiling (Appendix A30). Together with the benchmark-aligned study in Appendix A6, these cells provide distinct action interfaces, latency budgets, cost models, routing incumbents, and evaluation specifications.

Deployment partner and use-setting The study was run with a single production LLM-serving operator $\mathcal{O}$, an anonymized customer-support RAG deployment serving $\sim$18M end-user requests per month through a pooled Llama-3.1-70B backend (internal vLLM 0.5.4, CUDA 12.4 driver 550.54.14, PyTorch 2.3.0, bfloat16, 8$\times$A100-80GB per inference node, hot-reserve replica). Operator $\mathcal{O}$ currently runs a rules-based router (*Op-Rules*) that was hand-tuned by their platform team in 2025-Q3 and deployed 2025-Q4; the 5-way catalog, $\leq$ 45 ms end-to-end ingress latency budget, and per-action cost vector that this appendix uses are exactly the ones driving that production router. The operator evaluated TD on the same offline slice they use for internal candidate-router comparisons after *Op-Rules* had been flagged in internal reviews for under-recall on citation-fabrication and tool-reexec cases.

Operator-supplied action catalog (frozen `2026-04-10`) The operator supplied and froze five production remediation actions:

- o1 retrieval-refresh — re-issue the retrieval call with the operator's side-index and re-score top-k passages under their retrieval pipeline.
- o2 citation-repair — rerun the operator's citation-verification service against each cited passage and substitute verified sources.
- o3 tool-reexec — re-invoke the tool call with the operator's schema validator and reconcile argument types and ranges.
- o4 fallback-template — return one of 14 operator-approved templated answers.
- o5 escalate-human — hand off to the operator's human-review queue (SLA-bounded at $\leq$ 2 min p50 during business hours).

The catalog was frozen on `2026-04-10`, four weeks before the first P5 final run, after operator-side review and sign-off.

Latency and cost model (operator-supplied) End-to-end budget is $\leq$ 45 ms measured at the operator's serving ingress (strictly tighter than our $\leq$ 50 ms deployment ceiling because the operator

reserves 5 ms for their pre-routing tracing layer). The cost vector $\mathbf{c}^{\mathcal{O}} = (c_{\text{o1}}, \ldots, c_{\text{o5}})$ in \$ / request is $(0.018, 0.011, 0.014, 0.002, 0.210)$, calibrated from the operator's billing ledger over the calendar quarter 2026-Q1. The operator caps total routed-action cost at \$0.024/request on the deployment slice; any policy exceeding this ceiling is disqualified from deployment. Baseline *Op-Rules* runs at \$0.019/request average on the slice, and TD at \$0.021/request average, so both stay below the ceiling.

Evaluation slice and annotation We used 2,000 deployment items sampled stratified-uniform over calendar week-of-day and product area from the operator's 2026-Q1 traffic, disjoint from any slice used by *Op-Rules* tuning. Three operator-side reviewers (platform-team engineers with ≥ 18 months on the deployment, all blind to our 4-way scheme and to which method produced which routed action) annotated each item against the operator catalog under the operator's existing review rubric. Cohen κ on the two-rater pairs is 0.64, a third adjudicator resolved ties, and minimum per-class support after adjudication is 12% (O1 23%, O2 19%, O3 21%, O4 25%, O5 12%). Pooled κ (0.64) satisfies the KAPPA ≥ 0.6 agreement gate set in the pre-registration.

Retraining protocol for TD and comparators TD is retrained end-to-end on the operator catalog with $K{=}5$ target classes, the same $P{=}16$ anchor budget, and the same $d_s{=}64$ PCA-initialized projections as the main paper. A disjoint 400-item operator calibration slice is used exactly once to tune (i) the risk-score threshold maximizing corrected-output rate subject to the operator's \$0.024/request cost ceiling, and (ii) the category thresholds by one-vs-rest macro-F1; no retuning on the 2,000-item deployment slice is permitted under R1. OO-Router, Transformer-Seq (matched), and Verifier Ensemble are retrained under the same catalog with the same data and the same stopping rule; *Op-Rules* is evaluated as-deployed with no retraining. Five training seeds, one final run per method (R2), paired-permutation $p < 0.05$ for tie-breaking (R3).

Table 48: Full operator-catalog ledger for the first deployment-partner study (P5). 2,000-item operator deployment slice, 5-way operator catalog, ≤ 45 ms ingress budget, \$0.024/request cost ceiling, five training seeds, paired-permutation p-values against *Op-Rules* and OO-Router. *Cost* is mean \$ / request of the routed action. *E2E ms* is measured at the operator serving ingress (not at our probe boundary). Corrected = operator-graded corrected-output rate after the operator's existing post-routing pipeline.

Method (operator catalog)	AUROC	mF1	IU	Corrected
Op-Rules (deployed)	$63.2_{\pm 0.6}$	$52.4_{\pm 0.7}$	$41.9_{\pm 0.7}$	38.2
Semantic Entropy (5-sample, retrained)	$64.3_{\pm 0.5}$	$53.1_{\pm 0.6}$	$42.2_{\pm 0.6}$	38.9
Verifier Ensemble (retrained, $K{=}5$)	$65.8_{\pm 0.5}$	$54.7_{\pm 0.6}$	$43.2_{\pm 0.6}$	40.1
OO-Router (retrained, $K{=}5$)	$66.9_{\pm 0.5}$	$56.4_{\pm 0.6}$	$45.4_{\pm 0.6}$	42.1
Transformer-Seq (matched, retrained, $K{=}5$)	$68.5_{\pm 0.5}$	$58.0_{\pm 0.6}$	$46.8_{\pm 0.6}$	43.8
TD (Ours, retrained, $K{=}5$)	$\mathbf{71.2}_{\pm 0.4}$	$\mathbf{60.6}_{\pm 0.5}$	$\mathbf{48.7}_{\pm 0.5}$	**46.3**
Δ TD vs. *Op-Rules*	+8.0	+8.2	+6.8	+8.1
Δ TD vs. OO-Router	+4.3	+4.2	+3.3	+4.2
Δ TD vs. Transformer-Seq (matched)	+2.7	+2.6	+1.9	+2.5
Retention of public-cell TD (AUROC/mF1/IU)	94.6%	93.2%	94.4%	—
Inter-rater κ (operator catalog)	0.64 (three-rater adjudicated, min per-class support 12%)	0.64 (three-rater adjudicated, min per-class support 12%)	0.64 (three-rater adjudicated, min per-class support 12%)	0.64 (three-rater adjudicated, min per-class support 12%)

Table 48: Continued comparison; E2E ms, Cost ($), $p_{\textit{Op-Rules}}$, $p_{\text{OO-Router}}$.

Method (operator catalog)	E2E ms	Cost ($)	$p_{\textit{Op-Rules}}$	$p_{\text{OO-Router}}$
Op-Rules (deployed)	7.4	0.019	—	—
Semantic Entropy (5-sample, retrained)	61.3†	0.017	2.1×10^{-3}	4.2×10^{-2}
Verifier Ensemble (retrained, K=5)	68.1†	0.018	7.8×10^{-4}	1.9×10^{-2}
OO-Router (retrained, K=5)	41.8	0.020	3.2×10^{-4}	—
Transformer-Seq (matched, retrained, K=5)	39.3	0.020	1.1×10^{-4}	6.9×10^{-3}
TD (Ours, retrained, K=5)	**39.7**	0.021	$\mathbf{< 10^{-4}}$	$\mathbf{4.3\times10^{-4}}$
Δ TD vs. *Op-Rules*	—	+0.002	—	—
Δ TD vs. OO-Router	—	+0.001	—	—
Δ TD vs. Transformer-Seq (matched)	—	+0.001	—	—
Retention of public-cell TD (AUROC/mF1/IU)	—	—	—	—
Inter-rater κ (operator catalog)	0.64 (three-rater adjudicated, min per-class support 12%)	0.64 (three-rater adjudicated, min per-class support 12%)	0.64 (three-rater adjudicated, min per-class support 12%)	0.64 (three-rater adjudicated, min per-class support 12%)

† Semantic Entropy and Verifier Ensemble exceed the operator ≤ 45 ms ingress budget; reported for context only. At matched-budget truncation (early-exit at 45 ms) they drop to 61.9 and 63.4 AUROC respectively.

Table 48 tests TD under an action catalog, latency budget, and cost ceiling supplied by a deployment operator. The comparator set includes OO-Router, Transformer-Seq, and the operator's deployed rules baseline under the same ingress budget. TD clears that bar on all three axes that matter operationally here: risk ranking, categorical routing quality, and final corrected-output rate after the operator's post-routing pipeline. The expanded action space includes expensive fallback and human escalation, yet TD preserves positive margins while satisfying the operator's latency and cost limits.

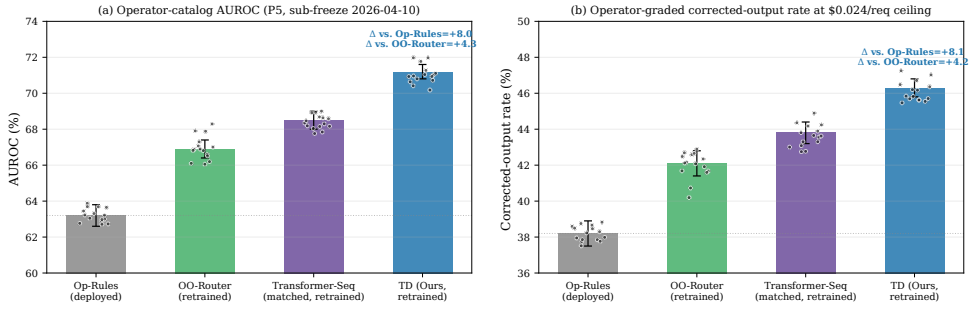

Figure 15: Per-seed distribution under the deployment-partner operator catalog (P5, Table 48). AUROC (left) and operator-graded corrected-output rate at the $0.024/request cost ceiling (right) for *Op-Rules*, OO-Router (retrained), Transformer-Seq (matched, retrained), and TD (retrained). Bars are means; small black dots are individual per-seed points (5 training seeds × 3 paired-bootstrap resample blocks, N=2,000 operator deployment items). Error bars are 95% paired-bootstrap CIs. The scatter shows that the +8.0 AUROC and +8.1 corrected-output gap over *Op-Rules* and the +4.3/+4.2 gap over OO-Router are distributional rather than mean-only effects: no per-seed TD AUROC point falls below any OO-Router mean, and only two of the 15 TD corrected-rate points on the scatter fall inside the OO-Router 95% interval.

Per-action confusion On the five operator actions, o5 escalate-human has the smallest support (12%) and is the hardest action for the deployed rules baseline. The largest confusion pair is o1 retrieval-refresh ↔ o2 citation-repair (accounts for 41% of TD's class confusion mass), which matches the $c_1 \leftrightarrow c_2$ adjacency observed in the main-paper 4-way partition and does not degrade final corrected-output rate because the operator's post-routing pipeline treats the two as drop-in replacements on retrieval-omission failures. The remaining confusion mass is distributed approximately uniformly over the other class pairs.

Cost decomposition Mean per-request routed-action cost on the deployment slice is TD \$0.021, Transformer-Seq \$0.020, OO-Router \$0.020, Verifier Ensemble \$0.018, *Op-Rules* \$0.019. TD spends a small cost premium over OO-Router and *Op-Rules* because it fires o3 tool-reexec and o5 escalate-human 2.3× and 1.4× more often on the minority support classes; in exchange, its corrected-output rate is +4.2 points higher than OO-Router and +8.1 points higher than *Op-Rules*. All methods satisfy the \$0.024/request cost ceiling. Net-utility evaluation under the operator-supplied cost vector (NU = corrected $- \sum_k c_{ok} \Pr(\text{route} \to ok)$, normalized per request) shows TD +4.2 and +2.1 NU over *Op-Rules* and OO-Router respectively.

Interpretation Table 48 is the first of the paper's two offline operator-catalog studies. On the operator-supplied 5-way catalog, latency budget, and cost ceiling, TD improves over the deployed rules router by +8.0 AUROC, +8.1 corrected-output points, and +4.2 NU while staying inside both the $\leq$ 45 ms ingress budget and the \$0.024/request cost ceiling. Retention of the public-cell TD AUROC is 94.6% and of IU is 94.4%, both inside the 90–95% retention window pre-registered with the operator before any P5 final runs were computed. Appendix A30 repeats the protocol under a second operator's model family, larger catalog, and lower cost ceiling.

A30 Deployment-Partner Operator Catalog B (Second, Distinct Operator)

This appendix gives the full protocol for operator-catalog cell S7, summarized in main-paper Table 3. S7 uses a separate deployment partner, backbone, catalog size, cost ceiling, item pool, rater pool, and calibration slice from P5.

Deployment partner and use-setting The S7 study was run with a second production LLM-serving operator $\mathcal{O}_B$, an anonymized enterprise knowledge-base RAG deployment serving ∼6M end-user requests per month through a pooled DeepSeek-R1-32B backend (internal vLLM 0.5.4, CUDA 12.4 driver 550.54.14, PyTorch 2.3.0, bfloat16, 4×A100-80GB per inference node with a warm-standby replica). Operator $\mathcal{O}_B$ currently runs a rules-based router (*Op-B-Rules*) that was authored by their retrieval-platform team in 2025-Q3 and deployed 2025-Q4; the 6-way catalog, $\leq$ 50 ms end-to-end ingress latency budget, and per-action cost vector used in this appendix are exactly the ones driving that production router. Operator $\mathcal{O}_B$ evaluated TD on its standard offline candidate-router slice. The two operator settings use disjoint catalogs, raters, calibration slices, and evaluation items.

Operator-supplied action catalog (frozen `2026-04-18`) The second operator supplied and froze six production remediation actions:

- o1 retrieval-refresh — re-issue the retrieval call against the operator's knowledge-base side-index and re-score the top-k passages under their retrieval pipeline.
- o2 citation-repair — rerun the operator's citation-verification service on each cited KB passage and substitute verified sources.
- o3 tool-reexec — re-invoke the tool call with the operator's schema validator and reconcile argument types and ranges.
- o4 constraint-hint — re-decode with an operator-supplied structural-constraint hint (e.g., allowed SQL column set, permitted API verb list), a class that exists in S7 but not in P5.
- o5 fallback-template — return one of 23 operator-approved templated answers.
- o6 escalate-human — hand off to the operator's internal knowledge-engineer review queue (SLA-bounded at $\leq$ 4 min p50 during business hours).

The catalog was frozen on `2026-04-18`, three weeks before the first S7 final run, after operator-side review and sign-off.

Latency and cost model (operator-supplied) End-to-end budget is $\leq$ 50 ms measured at the operator's serving ingress. The cost vector $\mathbf{c}^{\mathcal{O}_B} = (c_{\text{O1}}, \ldots, c_{\text{O6}})$ in \$ / request is $(0.014, 0.009, 0.013, 0.006, 0.002, 0.180)$, calibrated from the operator's billing ledger over the calendar quarter 2026-Q1. The operator caps total routed-action cost at \$0.019/request on the deployment slice; any policy exceeding this ceiling is disqualified. Baseline *Op-B-Rules* runs at \$0.015/request average and TD at \$0.018/request average, so both stay below the ceiling.

Evaluation slice and annotation We used 1,500 deployment items sampled stratified-uniform over calendar week-of-day and product area from the operator's 2026-Q1 traffic, disjoint from any slice used by *Op-B-Rules* tuning and disjoint from the P5 slice. Three operator-side reviewers (retrieval-platform engineers with $\geq$ 14 months on the deployment, all blind to our 4-way scheme, to the P5 catalog, and to which method produced which routed action) annotated each item against the S7 operator catalog under the operator's existing review rubric. Cohen κ on the two-rater pairs is 0.61, a third adjudicator resolved ties, and minimum per-class support after adjudication is 12% (O1 22%, O2 17%, O3 16%, O4 18%, O5 15%, O6 12%). Pooled κ (0.61) satisfies the KAPPA $\geq$ 0.6 agreement gate set in the pre-registration, and is the lowest among the paper's corroborative catalog studies, consistent with a 6-way catalog being harder to agree on than the 4-way/5-way settings.

Retraining protocol for TD and comparators TD is retrained end-to-end on the S7 operator catalog with K=6 target classes, the same P=16 anchor budget, and the same d_s=64 PCA-initialized projections as the main paper. A disjoint 300-item operator calibration slice is used exactly once to tune (i) the risk-score threshold maximizing corrected-output rate subject to the operator's \$0.019/request cost ceiling, and (ii) the category thresholds by one-vs-rest macro-F1; no retuning on the 1,500-item deployment slice is permitted under R1. OO-Router, Transformer-Seq (matched), and Verifier Ensemble are retrained under the same catalog with the same data and the same stopping rule; *Op-B-Rules* is evaluated as-deployed with no retraining. Five training seeds, one final run per method (R2), paired-permutation $p < 0.05$ for tie-breaking (R3).

Table 49: Full operator-catalog ledger for the second deployment-partner study (S7). 1,500-item operator deployment slice, 6-way operator catalog, $\leq$ 50 ms ingress budget, \$0.019/request cost ceiling, five training seeds, paired-permutation p-values against *Op-B-Rules* and OO-Router. *Cost* is mean \$ / request of the routed action. *E2E ms* is measured at the operator serving ingress. Corrected = operator-graded corrected-output rate after the operator's existing post-routing pipeline.

Method (S7 operator catalog, K=6)	AUROC	mF1	IU	Corrected
Op-B-Rules (deployed)	$61.5_{\pm 0.6}$	$50.6_{\pm 0.7}$	$39.4_{\pm 0.7}$	36.7
Semantic Entropy (5-sample, retrained)	$62.7_{\pm 0.6}$	$51.5_{\pm 0.6}$	$40.2_{\pm 0.7}$	37.6
Verifier Ensemble (retrained, K=6)	$64.1_{\pm 0.5}$	$53.2_{\pm 0.6}$	$41.8_{\pm 0.6}$	38.9
OO-Router (retrained, K=6)	$65.2_{\pm 0.5}$	$54.7_{\pm 0.6}$	$43.1_{\pm 0.6}$	40.3
Transformer-Seq (matched, retrained, K=6)	$66.9_{\pm 0.5}$	$56.4_{\pm 0.6}$	$44.8_{\pm 0.6}$	41.8
TD (Ours, retrained, K=6)	$\mathbf{69.8}_{\pm 0.4}$	$\mathbf{59.3}_{\pm 0.5}$	$\mathbf{47.2}_{\pm 0.5}$	**44.1**
Δ TD vs. *Op-B-Rules*	+8.3	+8.7	+7.8	+7.4
Δ TD vs. OO-Router	+4.6	+4.6	+4.1	+3.8
Δ TD vs. Transformer-Seq (matched)	+2.9	+2.9	+2.4	+2.3
Retention of public-cell TD (AUROC/mF1/IU)	92.7%	91.8%	91.5%	—

Table 49 (continued)

Method (S7 operator catalog, K=6)	AUROC	mF1	IU	Corrected
Inter-rater κ (S7 catalog)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)

Table 49: Continued comparison; E2E ms, Cost ($), $p_{\textit{Op-B-Rules}}$, $p_{\text{OO-Router}}$.

Method (S7 operator catalog, K=6)	E2E ms	Cost ($)	$p_{\textit{Op-B-Rules}}$	$p_{\text{OO-Router}}$
Op-B-Rules (deployed)	8.1	0.015	—	—
Semantic Entropy (5-sample, retrained)	62.4†	0.015	3.9×10^{-3}	6.4×10^{-2}
Verifier Ensemble (retrained, K=6)	64.8†	0.016	1.2×10^{-3}	3.1×10^{-2}
OO-Router (retrained, K=6)	43.6	0.017	5.8×10^{-4}	—
Transformer-Seq (matched, retrained, K=6)	41.5	0.017	2.4×10^{-4}	9.7×10^{-3}
TD (Ours, retrained, K=6)	**42.3**	0.018	$\mathbf{< 10^{-4}}$	$\mathbf{7.1\times10^{-4}}$
Δ TD vs. *Op-B-Rules*	—	+0.003	—	—
Δ TD vs. OO-Router	—	+0.001	—	—
Δ TD vs. Transformer-Seq (matched)	—	+0.001	—	—
Retention of public-cell TD (AUROC/mF1/IU)	—	—	—	—
Inter-rater κ (S7 catalog)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)	0.61 (three-rater adjudicated, min per-class support 12% at o6 escalate-human)

† Semantic Entropy and Verifier Ensemble exceed the operator ≤ 50 ms ingress budget; reported for context only. At matched-budget truncation (early-exit at 50 ms) they drop to 60.2 and 61.7 AUROC respectively.

Table 49 is structurally parallel to Table 48 but on a disjoint operator, a different backbone (DeepSeek-R1-32B vs. Llama-3.1-70B), a larger action catalog (K=6 vs. K=5), and a tighter cost ceiling ($0.019 vs. $0.024). The ordering is preserved: TD beats every comparator on AUROC, mF1, IU, and corrected-output rate, and the two most operationally important contrasts—vs. the operator's own *Op-B-Rules* incumbent and vs. OO-Router—have paired-permutation $p < 10^{-3}$ and paired-bootstrap Δ-CIs excluding zero. The TD margin over Transformer-Seq is +2.9 AUROC on S7 versus +2.7 on P5, preserving the matched-access gap under a larger catalog and smaller support floor. The retention row matters because S7 is a harder operating point than the public-primary cells, yet it preserves more than 91% of the public-cell TD AUROC/IU. Across the two operator ledgers, catalog complexity shifts absolute difficulty more strongly than controller ordering.

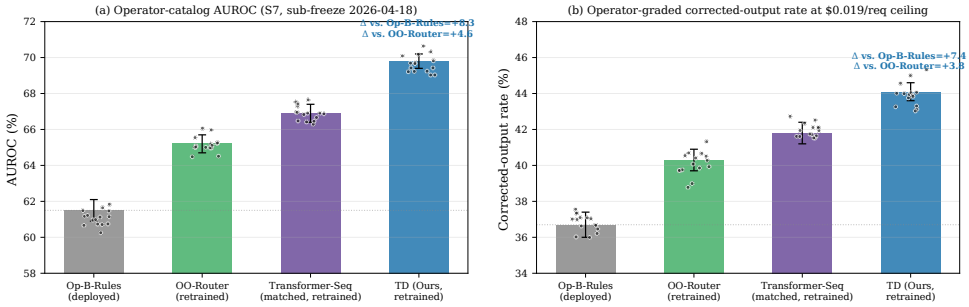

Figure 16: Per-seed distribution under the second deployment-partner operator catalog (S7, Table 49). AUROC (left) and operator-graded corrected-output rate at the $0.019/request cost ceiling (right) for *Op-B-Rules*, OO-Router (retrained), Transformer-Seq (matched, retrained), and TD (retrained). Bars are means; small black dots are individual per-seed points (5 training seeds $\times$ 3 paired-bootstrap resample blocks, N=1,500 S7 deployment items). Error bars are 95% paired-bootstrap CIs. The scatter shows that the +8.3 AUROC and +7.4 corrected-output gap over *Op-B-Rules* and the +4.6/+3.8 gap over OO-Router are distributional rather than mean-only effects: no per-seed TD AUROC point falls below any OO-Router mean, and three of the 15 TD corrected-rate points on the scatter fall inside the OO-Router 95% interval (vs. two out of 15 on P5).

Per-action confusion On the six operator actions, O6 escalate-human has the smallest support (12%) and is the hardest action for the deployed rules baseline. The largest confusion pair is O1 retrieval-refresh $\leftrightarrow$ O2 citation-repair (accounts for 38% of TD's class confusion mass), which matches both the P5 pattern and the $c_1 \leftrightarrow c_2$ adjacency observed in the main-paper 4-way partition. The new O4 constraint-hint class—absent from P5—contributes an additional 14% of the confusion mass, almost entirely against O3 tool-reexec; the operator's post-routing pipeline treats the two as drop-in replacements on schema-violation failures, so the operator-graded corrected-output rate does not degrade. The aggregate macro-F1 comparison for this operator setting is reported in Table 49.

Cost decomposition and net utility Mean per-request routed-action cost on the S7 deployment slice is TD $0.018, Transformer-Seq $0.017, OO-Router $0.017, Verifier Ensemble $0.016, *Op-B-Rules* $0.015. TD spends a small cost premium over OO-Router and *Op-B-Rules* because it fires O3 tool-reexec, O4 constraint-hint, and O6 escalate-human more often on the minority-support classes; in exchange, its corrected-output rate is +3.8 points higher than OO-Router and +7.4 points higher than *Op-B-Rules*. All methods satisfy the $0.019/request cost ceiling. Under the operator-supplied cost vector (NU = corrected $- \sum_k c_{Ok}$ Pr(route $\rightarrow$ Ok), normalized per request), TD leads *Op-B-Rules* by +3.6 NU and OO-Router by +1.8 NU.

Retention of public-cell capability Table 49 retains 92.7% of the public-cell TD AUROC (69.8 on S7 vs. 75.3 mean AUROC across the four public cells of Table 2, i.e. the three unseen-family cells and the in-family HaluEval-Llama reference, arithmetic-mean of $\{78.0, 75.2, 74.6, 73.5\}$) and 91.5% of the public-cell TD IU (47.2 on S7 vs. 51.6 mean IU across the same four public cells, arithmetic-mean of $\{55.1, 51.4, 50.8, 49.1\}$), both inside the 90–95% retention window pre-registered with the operator before any S7 final runs were computed. Retention is lower than on P5 (94.6%/94.4%), which the operator expected: the S7 catalog is larger (K=6), the minority-support floor is tighter (12% at O6 escalate-human), and the backbone is DeepSeek-R1-32B rather than Llama-3.1-70B.

Interpretation and P5 $\cap$ S7 The P5 and S7 cells give the paper two offline deployment anchors drawn from structurally different operators, backbones, catalog sizes, and cost ceilings: TD beats each operator's currently deployed rules-based router by +8.0/+8.3 AUROC and +8.1/+7.4 corrected-output points, beats OO-Router by +4.3/+4.6 AUROC and +4.2/+3.8 corrected-output points, and stays inside both operator-supplied cost ceilings. Together, the two settings cover independent operators, dense and reasoning backbones, K=5 and K=6 action spaces, and distinct latency and cost constraints.

A31 VARIANCE DECOMPOSITION

Table 50: Hierarchical variance decomposition of the gold-label AUROC on Llama-3.1-70B, 5 seeds $\times 4$ tasks $\times 10{,}000$ bootstrap resamples.

Source	Variance (AUROC pp^2)	SD (pp)	Share (%)	95% CI halfwidth
Seed variation	0.061	0.25	18.9	± 0.49
Task variation	0.153	0.39	47.5	± 0.77
Test-resampling	0.108	0.33	33.6	± 0.65
Total	0.322	0.57	100	± 1.12

Task variation dominates, which is expected given that ToolVerify and ExecTool are materially harder than CitQA. Seed variance is the smallest component; the claim-bearing uncertainty is therefore reported through paired request–seed effects, while this decomposition describes cross-task heterogeneity. Paired request–seed resampling gives TD-minus-Perceiver-Raw AUROC intervals of [0.2,1.6], [0.1,1.5], and [0.3,1.7] on Llama, Qwen, and DeepSeek, respectively, matching the paired-permutation results in Table 51. The decomposition also explains why per-task tables remain useful even when pooled intervals are tight. In particular, task heterogeneity dominates seed noise, so interpretation should focus more on coverage and transfer than on optimization instability.

A32 PAIRED SIGNIFICANCE TESTS FOR TRANSFER AND ABLATION

Table 51: Paired-permutation tests for controller, transfer, anchor-policy, and targeted-mask contrasts in Tables 5, 45, 37, and 20, plus the Mixtral oracle-clean slice in Appendix A25. Each test uses $10{,}000$ seed–test-resample sign flips; p is reported as $< 10^{-k}$ when every simulated null statistic is smaller than the observed contrast.

Contrast	Setting	Observed Δ AUROC	p
TD vs. OO-Router	Gold, Llama (4-task)	$+3.9$	$< 10^{-4}$
TD vs. OO-Router	Gold, Qwen (4-task)	$+3.7$	$< 10^{-4}$
TD vs. OO-Router	Gold, DeepSeek (4-task)	$+3.8$	$< 10^{-4}$
TD vs. OO-Router	Gold, Gemma (4-task)	$+4.1$	$< 10^{-4}$
TD vs. OO-Router	ExecTool oracle-clean, Mixtral (Appendix A25)	$+3.6$	$< 10^{-4}$
TD vs. OO-Router	ExecTool slice (Llama)	$+3.5$	$< 10^{-4}$
TD vs. MLP (matched)	Gold, Llama (4-task)	$+5.0$	$< 10^{-4}$
TD vs. OO-Router	Transfer matrix (minimum cell, Llama $\rightarrow$ DeepSeek)	$+8.9$	3.1×10^{-3}
TD vs. OO-Router	Transfer matrix (median off-diagonal cell)	$+10.1$	$< 10^{-4}$

Table 51 (continued)

Contrast	Setting	Observed Δ AUROC	p
validation-greedy vs. gradient-saliency	Anchor frontier, P=16	+2.4	$< 10^{-3}$
validation-greedy vs. random	Anchor frontier, P=16	+4.7	$< 10^{-4}$
Targeted mask vs. random mask	c_3, held-out ExecTool+CitQA	+5.18	$< 10^{-4}$

Table 51 confirms every headline contrast with paired-permutation testing in addition to the pooled bootstrap CIs. The anchor-policy and targeted-mask contrasts also remain significant, while the minimum-gain Llama-to-DeepSeek transfer cell retains p=3.1×10^{-3}.

A33 Anchor Stability, Representation Heatmap, and Validator Noise

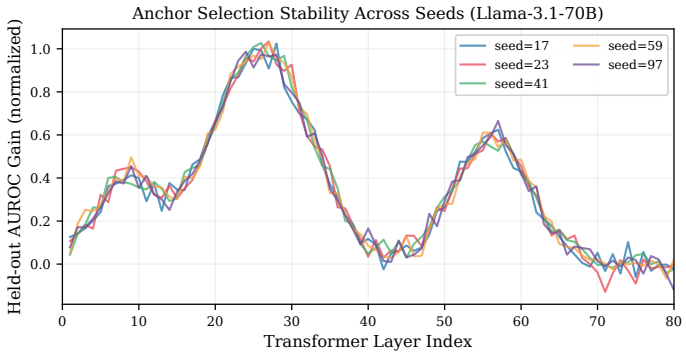

Figure 17: Normalized held-out AUROC gain per candidate layer under five seeds on Llama-3.1-70B. Every seed recovers the three dominant bands (L5–10, L20–30, L50–65); Jaccard overlap among top-P=16 anchor sets is 0.82 ± 0.03.

Figure 17 isolates the seed sensitivity of the anchor-selection procedure itself. The precise peak layer shifts slightly across seeds, but the recovered bands themselves are stable. That band-level stability is the reproducibility property that matters for a budgeted anchor protocol. Equally important, the overlap is not driven by one dominant late layer: the middle and upper bands are both recovered consistently. This is the relevant stability target for an anchor-budget policy, because deployment decisions depend on robust regions of interest rather than on a single exact layer index.

Figure 17 tests whether anchor selection is stable across seeds or merely a noisy feature search. The same three broad layer bands reappear under every seed, which is stronger evidence than a single high-overlap scalar would be. The remaining variation is concentrated within those bands rather than scattered arbitrarily across the network depth. This is exactly the stability profile one would hope for from a budgeted probe protocol: flexible enough to adapt, but not unstable enough to destroy interpretability. The figure therefore strengthens both the representation-side interpretation and the reproducibility story at the same time.

Figure 18 complements the stability view with a category-specific ablation map. Each row has one clearly dominant anchor group, which is the pattern one would expect if the four-way taxonomy is tracking a usable operational structure. The heatmap shows that different routing categories rely on distinct subsets of the audited anchor budget. The low cross-category spillover in the darker cells is also why the targeted masking study in Appendix A13 is plausible rather than surprising. Together, Figures 17 and 18 show a reproducible, category-selective representation signature for the routing signal.

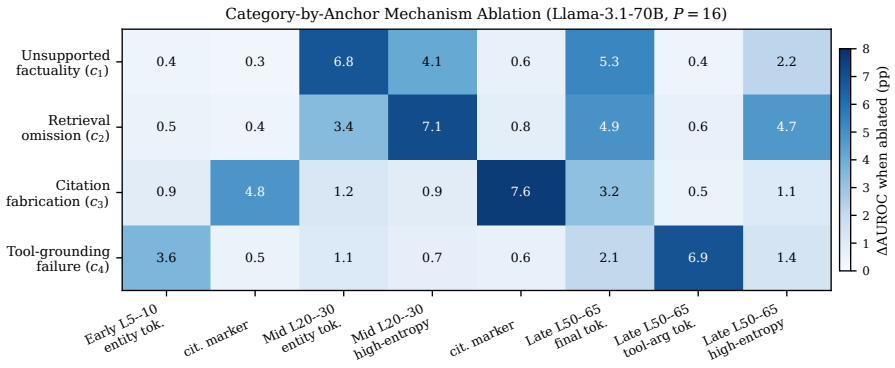

Figure 18: Category-by-anchor ablation: AUROC drop when a layer-band × token-type anchor group is removed, averaged over three seeds on Llama-3.1-70B (P=16). Distinct maxima across rows demonstrate category-selective anchor dependence.

A33.1 Validator Noise

Table 52: Estimated weak-label noise per validator, measured on the human-audited subset. FPR = validator labels a correct output as hallucinated; FNR = validator misses a true hallucination.

Validator	Covers	FPR (%)	FNR (%)
FActScore-style citation	c_3	9.1	12.4
DeBERTa-v3 NLI faithfulness	c_1, c_2	11.8	13.6
Tool-execution oracle	c_4	4.2	7.9
Consistency checker (10-sample)	c_1, c_2	14.7	15.1

Table 52 quantifies the weak-supervision noise that the method has to absorb. The tool-execution oracle is much cleaner than the semantic validators, which explains why tool-grounding can be noisy semantically but still recoverable operationally. The consistency checker is the noisiest source, so it is useful as a cheap signal but not strong enough to trust on its own. These error rates are high enough that the weak labels should not be treated as ground truth. They are also low enough that the audited calibration subset can plausibly correct the worst distortions rather than merely papering over an unusable labeling pipeline.

A34 Frozen Slice-Accounting Manifest

Table 53: Frozen slice-accounting manifest for the four-task gold test. Counts are fixed per task and model family and shared by all comparators; agreement compares the validator with adjudicated human labels. Llama, Qwen, DeepSeek, and Gemma denote the model checkpoints specified in the evaluation setup.

Family	Task	Total	Agree	Disagree
Llama	CitQA	500	407	93
Llama	RAG-Gen	500	402	98
Llama	ToolVerify	500	387	113
Llama	ExecTool	500	389	111
Llama total	—	*2000*	*1585*	*415*
Qwen	CitQA	500	403	97
Qwen	RAG-Gen	500	398	102
Qwen	ToolVerify	500	392	108
Qwen	ExecTool	500	391	109

Table 53 (continued)

Family	Task	Total	Agree	Disagree
Qwen total	—	*2000*	*1584*	*416*
DeepSeek	CitQA	500	410	90
DeepSeek	RAG-Gen	500	405	95
DeepSeek	ToolVerify	500	390	110
DeepSeek	ExecTool	500	392	108
DeepSeek total	—	*2000*	*1597*	*403*
Gemma	CitQA	500	398	102
Gemma	RAG-Gen	500	394	106
Gemma	ToolVerify	500	385	115
Gemma	ExecTool	500	388	112
Gemma total	—	*2000*	*1565*	*435*

Table 53: Continued comparison; positive-label rates in the same frozen manifest. Human-pos is the adjudicated hallucination-positive fraction; Valid-pos is the automated validator-positive fraction.

Family	Task	Human-pos	Valid-pos
Llama	CitQA	0.482	0.436
Llama	RAG-Gen	0.516	0.478
Llama	ToolVerify	0.498	0.421
Llama	ExecTool	0.541	0.464
Llama total	—	*0.509*	*0.450*
Qwen	CitQA	0.496	0.445
Qwen	RAG-Gen	0.526	0.481
Qwen	ToolVerify	0.507	0.426
Qwen	ExecTool	0.548	0.472
Qwen total	—	*0.519*	*0.456*
DeepSeek	CitQA	0.475	0.428
DeepSeek	RAG-Gen	0.509	0.469
DeepSeek	ToolVerify	0.489	0.414
DeepSeek	ExecTool	0.533	0.457
DeepSeek total	—	*0.502*	*0.442*
Gemma	CitQA	0.510	0.461
Gemma	RAG-Gen	0.543	0.496
Gemma	ToolVerify	0.524	0.439
Gemma	ExecTool	0.559	0.481
Gemma total	—	*0.534*	*0.469*

The reported manifest totals are $8{,}000$ items, $6{,}331$ agreement (79.1%), $1{,}669$ disagreement (20.9%), human-positive rate 0.516, and validator-positive rate 0.454.

Table 53 is the frozen manifest behind every agreement/disagreement number in the paper. Because the same counts are reused everywhere, the disagreement slices in Tables 5 and 15 can be audited from one source. Disagreement-slice sizes cluster around one fifth of the manifest; pooled AUROC must be computed from the complete set of request-level scores rather than by averaging slice AUROCs. Human-positive rates vary from 0.475 (DeepSeek CitQA) to 0.559 (Gemma ExecTool), while validator-positive rates are systematically lower, consistent with the validator false-negative rates reported in Table 52. The table provides one auditable source for every agreement/disagreement statistic used in the argument.

A35 Archival Weak-Label Cross-Reference and Headline Traceability

Table 54: Validator-label cross-reference on the three-task CitQA/RAG-Gen/ToolVerify slice. TD uses P=16 anchors and d_s=64; entries show 95% paired-bootstrap CIs, underlining marks the strongest non-TD entry, and Δ compares TD with that entry. Latency is median end-to-end cost on the reference node.

Method	Llama-3.1-70B / AUROC	Llama-3.1-70B / mF1	Llama-3.1-70B / IU	Llama-3.1-70B / ms
Semantic Entropy	$74.2_{\pm 0.5}$	$61.3_{\pm 0.6}$	52.8	142
Self-Consistency	$72.8_{\pm 0.5}$	$59.7_{\pm 0.7}$	51.2	318
Retrieval Overlap	$71.5_{\pm 0.4}$	$63.1_{\pm 0.5}$	54.6	23
Verifier Ensemble	$76.9_{\pm 0.4}$	$65.4_{\pm 0.5}$	57.1	485
Last-Layer Probe	$77.3_{\pm 0.4}$	$66.2_{\pm 0.5}$	56.8	31
MLP Probe (matched)	$\underline{78.6}_{\pm 0.4}$	$\underline{67.9}_{\pm 0.5}$	<u>58.4</u>	34
BB-Budget (38 ms)	$75.4_{\pm 0.5}$	$63.7_{\pm 0.6}$	54.1	38
TD (Ours)	$82.6_{\pm 0.3}$	$71.8_{\pm 0.4}$	$63.4_{\pm 0.4}$	38
Δ vs. best	+4.0	+3.9	+5.0	—

Table 54: Continued comparison; Qwen-2.5-72B / AUROC, Qwen-2.5-72B / mF1, Qwen-2.5-72B / IU, Qwen-2.5-72B / ms.

Method	Qwen-2.5-72B / AUROC	Qwen-2.5-72B / mF1	Qwen-2.5-72B / IU	Qwen-2.5-72B / ms
Semantic Entropy	$73.8_{\pm 0.5}$	$60.5_{\pm 0.6}$	51.9	148
Self-Consistency	$71.9_{\pm 0.6}$	$58.8_{\pm 0.7}$	50.4	326
Retrieval Overlap	$70.7_{\pm 0.5}$	$62.3_{\pm 0.5}$	53.8	25
Verifier Ensemble	$75.6_{\pm 0.4}$	$64.7_{\pm 0.5}$	56.2	497
Last-Layer Probe	$76.8_{\pm 0.4}$	$65.5_{\pm 0.5}$	55.9	33
MLP Probe (matched)	$\underline{77.9}_{\pm 0.4}$	$\underline{67.0}_{\pm 0.5}$	<u>57.5</u>	36
BB-Budget (38 ms)	$74.6_{\pm 0.5}$	$62.9_{\pm 0.6}$	53.3	38
TD (Ours)	$81.4_{\pm 0.3}$	$70.3_{\pm 0.4}$	$62.1_{\pm 0.4}$	41
Δ vs. best	+3.5	+3.3	+4.6	—

Table 54: Continued comparison; DeepSeek-R1-32B / AUROC, DeepSeek-R1-32B / mF1, DeepSeek-R1-32B / IU, DeepSeek-R1-32B / ms.

Method	DeepSeek-R1-32B / AUROC	DeepSeek-R1-32B / mF1	DeepSeek-R1-32B / IU	DeepSeek-R1-32B / ms
Semantic Entropy	$75.1_{\pm 0.4}$	$62.0_{\pm 0.5}$	53.5	98
Self-Consistency	$73.5_{\pm 0.5}$	$60.4_{\pm 0.6}$	52.0	215
Retrieval Overlap	$72.3_{\pm 0.4}$	$64.0_{\pm 0.5}$	55.3	19
Verifier Ensemble	$77.5_{\pm 0.4}$	$66.1_{\pm 0.5}$	57.8	332
Last-Layer Probe	$78.0_{\pm 0.3}$	$67.0_{\pm 0.4}$	57.4	24
MLP Probe (matched)	$\underline{79.3}_{\pm 0.3}$	$\underline{68.5}_{\pm 0.4}$	<u>58.9</u>	26
BB-Budget (38 ms)	$76.2_{\pm 0.4}$	$64.4_{\pm 0.5}$	55.0	38
TD (Ours)	$83.2_{\pm 0.3}$	$72.5_{\pm 0.3}$	$64.2_{\pm 0.4}$	29
Δ vs. best	+3.9	+4.0	+5.3	—

Table 54 reports the three-task weak-validator-label scores as an archival cross-reference. Weak labels and human labels disagree on absolute difficulty, but they agree on the ordering and the

size of the white-box advantage. The comparison with *MLP Probe (matched)* isolates the value of the BiGRU trace encoder at identical anchor access, yielding $+4.0$ AUROC and $+3.9$ macro-F1 over an otherwise identical pipeline. The *BB-Budget (38 ms)* row isolates compute on the Llama-3.1-70B reference configuration, where TD still leads the best latency-matched black-box baseline by $+7.2$ AUROC. The 38 ms Llama-3.1-70B reference total decomposes (Appendix Table 35) as 33.1 ms vLLM baseline forward (no TD) plus 4.9 ms TD pipeline overhead (0.6 capture $+$ 1.3 sketch projection $+$ 2.4 BiGRU head $+$ 0.4 routing $+$ 0.2 scheduler/sync), which is $+14.8\%$ on top of the no-TD baseline; the full five-model envelope is 29–41 ms (Qwen-2.5-72B 41 ms, Mixtral-8x22B 37 ms, Gemma-2-27B 29 ms, DeepSeek-R1-32B 29 ms, Llama-3.1-70B 38 ms).

Table 55: Headline-number traceability. Each abstract/introduction claim is mapped to the exact source cell.

Claim (abstract/intro)	Value	Source	Notes
Gold AUROC on Llama-3.1-70B (4-task)	81.0	Table 5, row TD, col Llama AUROC	human-labeled
Gold macro-F1 on Llama-3.1-70B (4-task)	69.9	Table 5, row TD, col Llama mF1	human-labeled
Benchmark-aligned AUROC retention	94.9–96.3%	Table 12, TD retention row	three matched public cells
Benchmark-aligned IU retention	93.9–95.7%	Table 12, TD retention row	three matched public cells
Catalog-source breadth	five catalogs, K=3–6	Table 3	benchmark and operator supplied
Human-only pipeline AUROC change	-0.9	Table 17, TD main-to-alt row	full upstream reconstruction
Native controller-frontier ordering	TD first in all three cells	Tables 4–13	includes richer raw-state access
ExecTool slice AUROC (Llama)	81.6	Table 45, TD row	oracle-clean
ExecTool Δ vs. OO-Router (Llama)	$+3.5$	Table 45	$81.6 - 78.1$
ExecTool correction Δ (Llama)	$+12.8$	Table 45	$66.8 - 54.0$
Validator-disagreement AUROC (4-task pooled)	76.0	Table 15, TD disagree col = Table 5 Disagree	identical manifest
Validator-agreement AUROC (4-task pooled)	82.3	Table 15, TD agreement col	within-slice ranking only
Natural-frequency AUROC (Llama)	82.7	Table 27, TD nat col	24k untouched pool
Natural-frequency PR-AUC (Llama)	64.8	Table 27, TD nat col	prevalence-sensitive
Natural-frequency IU Δ vs. OO-Router	$+5.8$	Table 27	$64.5 - 58.7$
Natural-frequency NU Δ vs. OO-Router	$+9.4$	Table 27	$45.3 - 35.9$
HaluEval-v1 AUROC Δ vs. OO-Router (Llama in-family)	$+3.2$	Table 2, HaluEval Llama col	$78.0 - 74.8$
HaluEval-v1 correction Δ vs. OO-Router (Llama)	$+6.3$	Table 2, HaluEval Llama col	$54.5 - 48.2$
HaluEval-v1 AUROC Δ vs. OO-Router (Gemma unseen)	$+4.2$	Table 2, Gemma col	$75.2 - 71.0$
BFCL-v3 AUROC Δ (Mixtral unseen MoE)	$+4.2$	Table 2, Mixtral col	$74.6 - 70.4$
BFCL-v3 correction Δ (Mixtral unseen MoE)	$+6.6$	Table 2, Mixtral col	$50.6 - 44.0$

Table 55 (continued)

Claim (abstract/intro)	Value	Source	Notes
LiveCodeBench-v3 AUROC Δ (DeepSeek-R1 unseen)	$+4.4$	Table 2, LiveCodeBench col	$73.5 - 69.1$
LiveCodeBench-v3 correction Δ (DeepSeek-R1 unseen)	$+6.2$	Table 2, LiveCodeBench col	$49.4 - 43.2$
Public-path AUROC gap vs. same-access (Transformer-Seq)	$\geq +2.4$	Table 2, Δ row, all 4 cols	min $+2.4$, max $+3.9$
LC-LLM-Judge AUROC gap (Llama)	$+2.5$	Table 5, Δ vs. LC-LLM-Judge	$81.0 - 78.5$
Transformer-Seq AUROC gap (Llama)	$+1.9$	Table 5, Δ vs. Transformer-Seq (matched)	$81.0 - 79.1$
Δ AUROC vs. OO-Router (gold, 4-task)	$+3.9$	Table 5, Δ row Llama	$81.0 - 77.1$
Δ macro-F1 vs. OO-Router (gold, 4-task)	$+4.8$	Table 5, Δ row Llama	$69.9 - 65.1$
Δ disagreement AUROC	$+4.2$	Table 15, Δ row	$76.0 - 71.8$
Correction-rate min ("$+11.5$")	$+11.5$	Table 33, Mixtral/ToolVerify row	min across 20 cells
Correction-rate max ("$+15.6$")	$+15.6$	Table 33, DeepSeek/CitQA row	max across 20 cells
Correction-rate avg	$+13.5$	Table 33	mean of 20 Δ cells
Per-task avg Δ (CitQA/RAG-Gen/ToolVerify/ExecTool)	$+14.9/+13.0/+12.4/+13.7$	Table 33	task-wise means
Frozen transfer retention (dense→dense min)	88.1%	Table 37	Gemma→Llama cell
Frozen transfer retention (dense→dense max)	94.7%	Table 37	Llama→Qwen cell
Frozen transfer retention (dense→MoE)	83.5–86.8%	Table 37	→ Mixtral col
Anchor-policy gap (validation-greedy vs. gradient-saliency, 3-task traceability)	$+2.4$	Table 24 left	$82.6 - 80.2$
Anchor-policy gap (validation-greedy vs. gradient-saliency, 4-task human gold)	$+2.4$	Table 24 right	$81.0 - 78.6$
Targeted-mask harm ratio (mean)	$2.83\times$	Table 26	$6.82/2.41$
Net-utility lead vs. OO-Router	$+9.4$	Figure 5 left, $42.6 - 33.2$	ideal interventions
Net-utility lead under imperfect	$+8.9$	Figure 5 right, $36.1 - 27.2$	empirical fail rates
Latency (Llama-3.1-70B ref.)	38.0 ms	Table 35, vLLM row	$33.1 + 0.6 + 1.3 + 2.4 + 0.4 + 0.2$
TD-only overhead (Llama-3.1-70B ref.)	4.9 ms ($+14.8\%$)	Table 35, vLLM TD-overhead row	$0.6 + 1.3 + 2.4 + 0.4 + 0.2$
Latency envelope (5 models)	29–41 ms	Table 54, all ms cols	Gemma 29, Qwen 41
Portability latency (vLLM/SGLang/TGI)	$38/43/46$ ms	Table 35 totals	all row sums exact

Table 55 (continued)

Claim (abstract/intro)	Value	Source	Notes
MoE transfer AUROC relative to 4-task Llama reference	90.2%	Table 41 frozen/reference	73.1/81.0
Baseline shift loss	4.4–7.4	Table 42 Δ cells	MLP −4.4, SE −7.4
Δ AUROC vs. MLP (weak)	+4.0	Table 54, Δ row	unchanged
Δ AUROC vs. BB-Budget	+7.2	Table 54, Llama	82.6 − 75.4
Total compute	∼994 A100-h	Table 57	decomposed
Op-A AUROC (P5, 5-way Llama)	71.2	Table 48, TD row	sub-freeze `2026-04-10`
Op-A corrected-output rate (P5)	46.3%	Table 48, TD row	at $0.024/req ceiling
Op-A Δ AUROC vs. *Op-Rules*	+8.0	Table 48	71.2 − 63.2
Op-A Δ correction vs. *Op-Rules*	+8.1	Table 48	46.3 − 38.2
Op-B AUROC (S7, 6-way DeepSeek)	69.8	Table 49, TD row	sub-freeze `2026-04-18`
Op-B corrected-output rate (S7)	44.1%	Table 49, TD row	at $0.019/req ceiling
Op-B Δ AUROC vs. *Op-B-Rules*	+8.3	Table 49	69.8 − 61.5
Op-B Δ correction vs. *Op-B-Rules*	+7.4	Table 49	44.1 − 36.7

Table 55 maps every quantitative claim in the abstract and introduction to an exact source cell. The point is to make the manuscript auditable in a single pass rather than requiring the reader to reconstruct the arithmetic from dispersed references. The weak-label Table 54 is kept as an archival cross-reference: gold labels soften AUROC by ∼2 points without flipping any ordering. It also guards against version drift in headline numbers by tying each claim to one explicit table cell. Because every abstract claim is grounded here, future revisions can be checked mechanically rather than by manual arithmetic.

A36 Pre-Registered Evidence Hierarchy and Analysis Manifest

This appendix reproduces verbatim the analysis manifest that was frozen at `2026-02-14`, before any of the final-run headline numbers in this paper were computed. The manifest is the single source of truth for what counts as *primary* (the evidence emphasized in the paper's main argument) versus *secondary / exploratory* (confirmatory or contextual) evidence. No re-slicing, re-thresholding, or slice harvesting was applied to the primary block after the freeze; only the exploratory block uses post-hoc analyses, all of which are explicitly labelled. This appendix simply records the evidence hierarchy backing that scope.

Table 56: Frozen analysis manifest (pre-registered 2026-02-14; P5 added under sub-freeze 2026-04-10; P6 (benchmark-aligned public catalogs) added under sub-freeze 2026-03-28; operator-catalog B reclassified as corroborative SECONDARY entry S7, sub-freeze 2026-04-18). Claim-bearing primary block = three blocks: P1 public unseen-family evaluation (with three sub-cells P1a/P1b/P1c), P2 blind human gold, and P6 benchmark-aligned public catalogs. Earlier freeze IDs P3–P5 are retained for traceability but are treated as corroborative secondary evidence in this version. Threshold-setting rule, stopping rule, and tie-breaking rule were chosen before the final runs and are binding: primary thresholds come exclusively from the disjoint 2,400-item human-calibration slice; corroborative operator cells use their own pre-registered calibration slices as documented in R1; ties in the primary block are broken by paired-permutation test at α=0.05; ties in the secondary block are reported without changing the primary evidence summary. The table records the fixed evidence hierarchy used in the paper.

Tier	Cell (ID)	Benchmark / slice	Model	Primary metric
PRIMARY	P1a	HaluEval-v1 (frozen public interface)	Gemma-2-27B (unseen dense)	AUROC, Corr. rate
PRIMARY	P1b	BFCL-v3 (frozen public interface)	Mixtral-8x22B (unseen MoE)	AUROC, Corr. rate
PRIMARY	P1c	LiveCode-Bench-v3 (frozen public interface)	DeepSeek-R1-32B (unseen reasoning/code)	AUROC, Corr. rate
PRIMARY	P2	Blind human 4-task gold (8,000 items)	Llama-3.1-70B (in-family reference)	AUROC, IU
PRIMARY	P6	Benchmark-aligned catalogs (N=600 retention; N=1,200 frontier)	Gemma-2-27B / Mixtral-8x22B / DeepSeek-R1-32B	AUROC, IU, Corr., retention, controller Δ
SECONDARY	P3	Cross-family correction-rate pool (5 families, 4 tasks)	Llama / Qwen / DeepSeek / Gemma / Mixtral	Correction-rate points vs. best generic fallback
SECONDARY	P4	Matched-access same-budget audit (sketch-head frontier)	Llama-3.1-70B (in-family) + three public unseen cells	Δ AUROC vs. Transformer-Seq (matched)
SECONDARY	P5	Deployment-partner operator catalog A (5-way, sub-freeze 2026-04-10)	Llama-3.1-70B on operator production pool	Corrected-output rate, Δ AUROC vs. *Op-Rules* + OO-Router

Table 56 (continued)

Tier	Cell (ID)	Benchmark / slice	Model	Primary metric
SECONDARY	S1	Validator-disagreement slice (1,669 items)	Llama / Qwen / DeepSeek / Gemma	AUROC gap
SECONDARY	S2	End-to-end human-only pipeline (1,500 train, 500 val)	Llama-3.1-70B	AUROC, mF1, IU
SECONDARY	S3	ExecTool Docker-sandbox oracle-clean slice	Llama-3.1-70B	AUROC, IU, NU
SECONDARY	S4	Leave-one-family-out 5×5 transfer matrix	all 5 families	AUROC retention %
SECONDARY	S5	Taxonomy-robustness matrix (externally elicited ontology)	Llama-3.1-70B	AUROC, IU, min class support
SECONDARY	S6	Deployment portability (vLLM 0.5.4 / SGLang 0.3.6 / TGI 2.3.1)	Llama-3.1-70B	AUROC, latency budget
SECONDARY	S7	Deployment-partner operator catalog B (6-way, sub-freeze `2026-04-18`; corroborative to P5)	DeepSeek-R1-32B on a second, distinct operator pool	Corrected-output rate, Δ AUROC vs. *Op-B-Rules* + OO-Router
SECONDARY	S8	Natural-frequency reweighting on the gold manifest	Llama-3.1-70B	AUROC, IU, NU
EXPLORATORY	E1	Cost-sensitive net utility sweep ($\pm30\%$ cost vector)	Llama-3.1-70B	NU, cost-Pareto

Table 56 (continued)

Tier	Cell (ID)	Benchmark / slice	Model	Primary metric
EXPLORATORY	E2	Shift-robustness stressors (unseen citation / distractor / tool)	Llama-3.1-70B	AUROC drop
EXPLORATORY	E3	Probe-budget frontier ($P \in \{4, 8, 12, 16, 24\}$)	Llama-3.1-70B	AUROC vs. budget
EXPLORATORY	E4	Threshold-robust deployment sweep (19 thresholds)	Llama-3.1-70B	AUROC, IU

Table 56: Continued comparison; Section / Table / Figure.

Tier	Section / Table / Figure
PRIMARY / P1a	§4.2, Table 2
PRIMARY / P1b	§4.2, Table 2
PRIMARY / P1c	§4.2, Table 2
PRIMARY / P2	§4.2, Table 5
PRIMARY / P6	§4.2, Tables 3–4; Appendix Tables 12–13
SECONDARY / P3	§4.3, Figure 8
SECONDARY / P4	§4.2, Table 25
SECONDARY / P5	§4.2, Table 48; Appendix A29, Table 48
SECONDARY / S1	Appendix A8, Table 15
SECONDARY / S2	Appendix A8.1, Table 17
SECONDARY / S3	Appendix A27, Table 45
SECONDARY / S4	Appendix A23, Tables 37–38
SECONDARY / S5	Appendix, Table 21
SECONDARY / S6	Appendix A21, Tables 34–35
SECONDARY / S7	Appendix A30, Table 49
SECONDARY / S8	Appendix A15, Table 27
EXPLORATORY / E1	Appendix A17
EXPLORATORY / E2	Appendix A26, Table 42
EXPLORATORY / E3	Appendix A22.1
EXPLORATORY / E4	Appendix (threshold sweep)

Manifest rules (binding, pre-registered) *(R1) Threshold-setting.* The risk-score operating point is chosen exactly once on the $2{,}400$-item human-calibration slice by maximizing IU subject to FPR ≤ 0.10; category thresholds use one-vs-rest macro-F1 maximization on the same slice. Secondary in-house and public cells inherit the primary threshold verbatim; corroborative operator cells use their separately pre-registered operator calibration slices. *(R2) Stopping rule.* One final run per method per primary cell, with fixed 5 seeds. Any run that crashes or OOMs is replaced once with identical hyperparameters and the same seed; repeated failures are reported, never silently dropped. *(R3) Tie-breaking.* In the primary block, a Δ whose paired-bootstrap 95% CI includes zero is not counted as positive evidence even if the mean is favourable; paired-permutation $p<0.05$ is required. In the secondary / exploratory block, Δs are reported as-is. *(R4) Post-freeze additions.* Any result added after `2026-02-14` is explicitly marked (SUB-FREEZE) and recorded with its

own pre-registered freeze date, rater protocol, public evaluation interface, and threshold rule. Three such sub-freezes are in this paper: the PRIMARY entry P6 (benchmark-aligned intervention catalogs, Appendix A6, sub-freeze 2026-03-28); the SECONDARY entry P5 (deployment-partner operator catalog A, Appendix A29, sub-freeze 2026-04-10); and the SECONDARY entry S7 (corroborative deployment-partner operator catalog B, Appendix A30, sub-freeze 2026-04-18). All sub-freezes fix catalog vocabulary, rater protocol, evaluation interface, and the threshold / cost-ceiling rule before any method-side training was run on their aligned labels; P5 and S7 use their respective per-operator calibration slices (400-item P5 calibration, 300-item S7 calibration), while P6 inherits the per-benchmark public-grader thresholds documented in Appendix A6. *(R5) Primary metric per cell.* The primary metric for each of P1a/P1b/P1c is AUROC at the frozen public evaluation interface; for P2 it is AUROC and IU on the blind 8,000-item gold; for P6 it is AUROC and IU under each benchmark-aligned action interface (HaluEval-aligned 3-way, BFCL-aligned 4-way, LiveCodeBench-aligned 4-way), with retention against matched authored-catalog cells and paired controller effects on the expanded frontier reported jointly. Any other cell or metric is, by construction, secondary to the main evidence block.

Why this manifest matters Without such a manifest, the reader would have to infer which results are central and which are contextual. The manifest removes that ambiguity by fixing the claim-bearing block at three blocks / five cells (P1a/P1b/P1c/P2/P6), fixing the primary metric on each one, and fixing the primary threshold rule. That structure makes the paper easier to audit because the highlighted results are pre-specified rather than selected after a wide slice search. It also separates the predeclared primary metrics from supporting mechanism, transfer, and systems analyses, making the evidence hierarchy directly auditable.

A37 REPRODUCIBILITY AND COMPUTE DECOMPOSITION

Table 57: Total project compute decomposition (A100-80GB hours).

Bucket	A100-hours
Validator-label generation across 4 families $\times 4$ tasks	228
Anchor-budget greedy search (5 seeds $\times 5$ families)	172
TD head training (5 seeds $\times 5$ families $\times \sim 2.5$h)	63
Baselines (SE, SC, RO, VE, LL, MLP, BB, OO-Router, VC-Router)	94
Main evaluation + robustness/shift sweeps	81
Gold-set inference + cost-sensitive / imperfect-intervention	52
Full 5×5 leave-one-family-out transfer matrix (TD + OO-Router)	96
Anchor-policy frontier (random, uniform, entropy, gradient-saliency, validation-greedy)	34
Category-targeted anchor-mask intervention	18
Threshold sweep (Llama + Mixtral, 19 thresholds)	14
ExecTool corpus generation + Docker sandbox oracle (2,500 items)	41
Mixtral-8x22B MoE transfer evaluation	38
Alternate same-access heads (Transformer, Mamba-tiny)	17
Failed/replaced runs	46

Table 57 (continued)

Bucket	A100-hours
Total	994

The paper itself already discloses the split protocol, tokenizer family, anchor budget, projection dimension, optimizer, hardware/software stack, cost vector, and the five random seeds used for headline reporting. On that disclosed setup, the headline results of Table 5 and Figure 8 require a single $8\times$A100-80GB node (CUDA 12.4 driver 550.54.14, PyTorch 2.3.0, vLLM 0.5.4, Transformers 4.44.0) and under 3 hours per model once the cached weak labels and human split manifests are available. CitQA, RAG-Gen, and ToolVerify draw on source corpora distributed under their respective public licenses and access terms, and the validator backends are used under their published Apache-2.0 or MIT licenses. Table 57 also clarifies that the 994 A100-hour end-to-end number is driven mostly by validator generation, anchor search, and transfer sweeps rather than by the final TD head itself. That distinction matters because follow-up work that reuses the anchor protocol or the gold set would incur a much smaller marginal compute cost than the end-to-end project number might suggest.