

NeurIPS 2026 · Main Conference Poster #36660

Relevance Is Not Necessity:

Selecting the Necessary API Set for Tool-Using LLMs

Chen Wang^{1,2*}

¹College of Systems Engineering, National University of Defense Technology, Changsha 410073, Hunan, China

²National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410073, Hunan, China

*Corresponding author

Poster Session Info

NeurIPS 2026 Main Track

Paper ID: #36660

Benchmark: HSR-Bench
($|\mathcal{A}| = 1,129$ APIs, 300 tasks)

Method: TC-MASS Refiner

Core Claim

Whole-set exact supervision,
not relevance ranking alone

+9.0% to +10.3%

Plan Exact Match gain on
HSR-Bench ($p < 10^{-4}$)

0 Extra LLM Calls

Offline refiner over frozen
traces (19 ms CPU / task)

2,201 Evaluations

Theory-predicted transfer
& 3-family blind judges

1. The Relevance–Necessity Gap

Why ranking APIs by query similarity fails multi-tool set verification

Retrieve-then-Plan vs. Set-Exact Verification

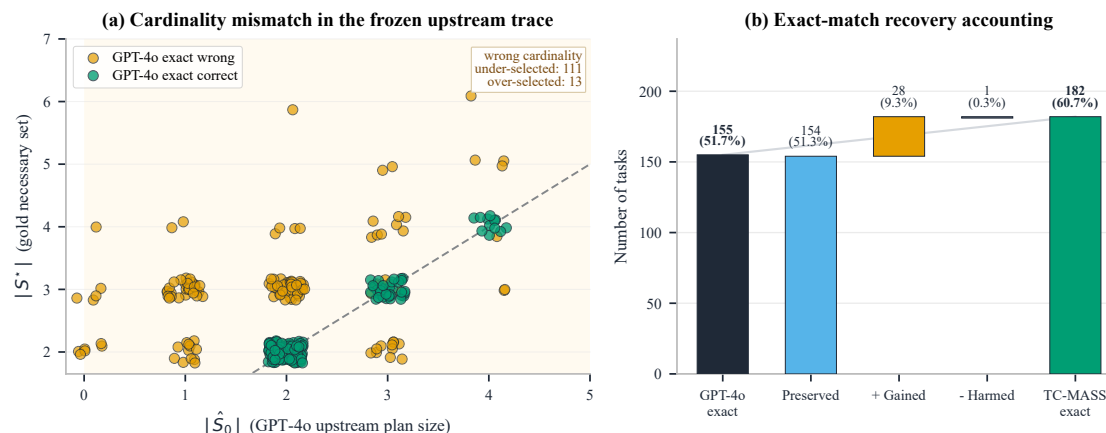
In large-pool tool use ($|\mathcal{A}| = 1,129$), a retriever narrows $\mathcal{A}$ to C by **relevance** $\text{rel}(s \mid x) = \Pr[s \in C \mid x]$, and an upstream LLM emits $\hat{S}_0 \subseteq C$. Multi-tool verifiers score the **gold necessary set** S^* under strict whole-set loss:

$$\ell_{\text{exact}}(\hat{S}, S^*) = \mathbb{I}[\hat{S} \neq S^*].$$

- **Confusable siblings:** If $S^* = \{\text{flight_search}, \text{hotel_book}\}$ and the plan calls $\{\text{flight_search}, \text{hotel_search}\}$, both hotel APIs are relevant and in C , yet the set fails.
- **Cardinality mismatch:** Fixed Top- k ignores task-dependent $|S^*|$.
- **Zero set supervision:** Per-API scores never see $\mathbb{I}[S = S^*]$.

Operational Scope: “Necessary” means membership in verifier-defined S^* in retrieval-conditioned, large-pool, multi-tool settings.

Upstream Cardinality Mismatch & TC-MASS Recovery (300 Tasks)



Left: Frozen GPT-4o plans frequently miss the gold set size $|S^*|$ (off-diagonal points).

Right: TC-MASS repairs **28 non-exact upstream plans** while regressing on only **1 task** (+9.0% net exact gain), using the *same* frozen trace and zero extra LLM calls.

2. Formal Boundary Theory & Dependency Diagnostics

NeurIPS 2026 · Slide 3 / 10

When per-API marginals fail, when Top- k suffices, and why set BCE is Bayes-consistent

Three Boundary Propositions

Proposition 1 (Decomposability Gap):

Any decomposable selector uses only marginals

$\hat{S}(x) = \{s \in C : \phi(\text{nec}(s | x), k) = 1\}$. On $|C| = 4$ with

$P(S^* = \{1, 3\}) = P(S^* = \{2, 4\}) = \frac{1}{2}$, marginals are uniform $\text{nec}(s) = \frac{1}{2}$; every symmetric decomposable rule incurs exact-match loss $\geq 2/3$ vs. $1/2$ for Bayes $\arg \max_S P(S | x)$ (**strict gap** $\geq 1/6$).

Proposition 2 (Top- k Sufficiency):

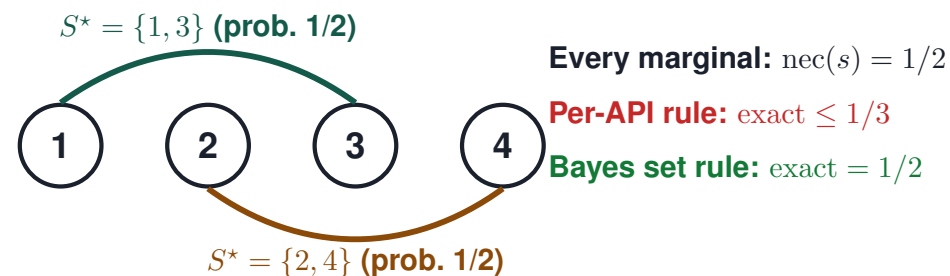
With oracle $k = |S^*|$, Top- k on $\widehat{\text{nec}}$ is exact iff indicators $\{\mathbb{I}[s \in S^*]\}$ are conditionally **independent** with margin $\gamma > 2\|\widehat{\text{nec}} - \text{nec}\|_\infty$.

Proposition 3 (Proper Set Supervision):

Over finite menu $S_x \subseteq 2^C$, binary cross-entropy on $\mathbb{I}[S = S^*]$ is Bayes-consistent for $\arg \max_{S \in S_x} P(S^* = S | x)$, bounded by $\Pr[\text{exact}] \leq \Pr[S^* \in S_x]$.

Theoretical Takeaway: Pairwise co-occurrence and sibling exclusivity require scoring candidate *sets* $S \in S_x$, not isolated APIs.

Proposition 1 in One Picture



Empirical Preconditions on HSR-Bench

41%

tasks contain a
same-family gold pair

0.58

family-confusable pool
mass η (0.89 on G3)

9×+

top-20 gold
pairs co-occur
vs. marginal product

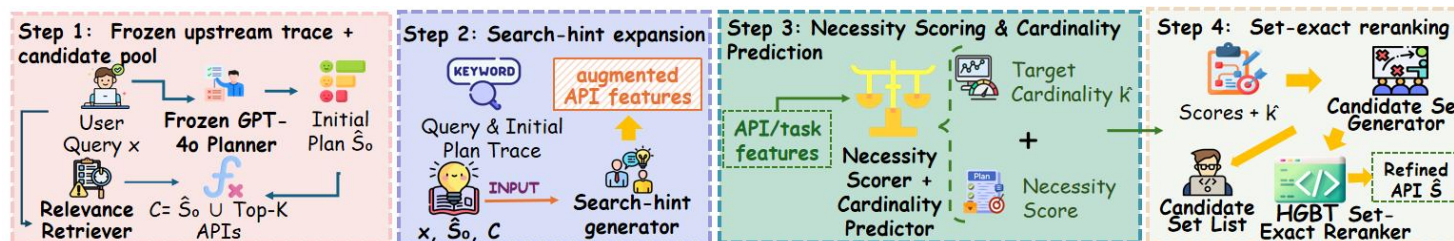
Diagnosis: Strong positive/negative API dependencies violate Proposition 2's independence condition on HSR-Bench.

3. TC-MASS: Offline Set Refiner Over Frozen Traces

NeurIPS 2026 · Slide 4 / 10

Four deterministic CPU stages with zero marginal LLM calls

Pipeline Overview



Four Stages (Algorithm 1)

- **(a) Augmented Candidate Pool**: $C \cup \hat{S}_0$ ($|C| \leq 40$) with compact search hints built from family, identifier tokens, description, and schema text.
- **(b) Per-API Necessity Scorer** f_{api} : Class-balanced logistic regression over **17 features in 6 groups** (rank, lexical/semantic overlap, name/param match, family redundancy, utility/hard-negative penalties, upstream flag).
- **(c) Cardinality Predictor** g : Task-level bucket model $\hat{k} = g(x, |\hat{S}_0|)$ commits to a target set size.
- **(d) Set-Exact Reranker** h : Enumerates menu $\mathcal{S}_x$ around $\hat{k}$ (score floors, family caps, add/drop/replace edits, family beam; mean **61 sets/task**) and selects $\arg \max_{S \in \mathcal{S}_x} (h(x, S) - \lambda |S|)$.

Deterministic Refiner: Operates strictly on the frozen upstream trace $\hat{S}_0$ and candidate pool C without re-prompting the LLM.

Efficiency & Top Set Signals

0

extra LLM calls

19 ms

CPU inference / task

61

candidate sets / task

3.6 s

training / fold

Top HGBT Permutation Importances:

1. Mean necessity rank (0.127 ± 0.026)
2. Max family frac. (0.085 ± 0.030)
3. Pairwise score prod. (0.081 ± 0.010)
4. Upstream Jaccard (0.073 ± 0.019)

4. Main Results on Leakage-Free HSR-Bench (300 Tasks)

NeurIPS 2026 · Slide 5 / 10

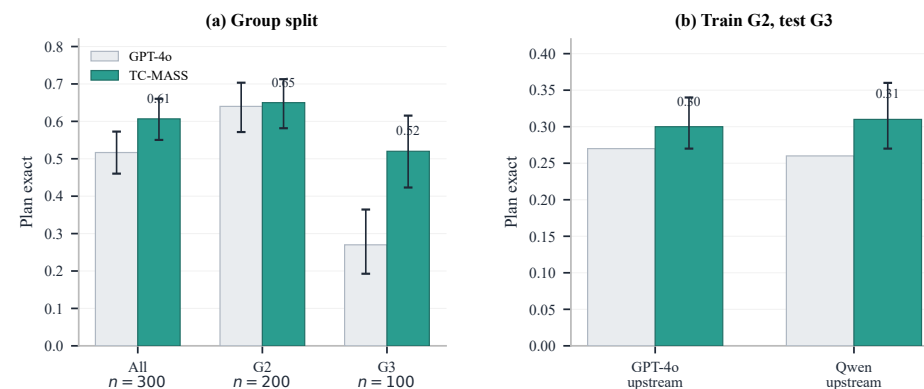
Set-exact refinement beats stronger retrievers and re-prompting on the same pool

5-Fold Task-Level CV ($B = 2,000$ Paired Bootstrap)

Method	Recall	Exact	Δ Exact [95% CI]	Calls
ToolBench retriever	0.118	0.107	—	1.41
HSR learned ranker	0.283	0.273	—	2.23
TF-IDF retriever (best 1st-hop)	0.395	0.377	—	2.21
Frozen GPT-4o Direct	0.568	0.517	Reference (0.000)	2.24
GPT-4o + Chain-of-Thought	0.548	0.523	+0.007 [−0.027, +0.040]	2.24
GPT-4o + Draft-and-Self-Verify	0.563	0.530	+0.013 [−0.017, +0.047]	2.27
TC-MASS Compact ($\lambda = 0.12$)	0.640	0.597	+0.080 [+0.047, +0.117]	2.42
TC-MASS Balanced ($\lambda = 0.06$)	0.657	0.603	+0.087 [+0.053, +0.123]	2.47
★ TC-MASS HGBT ($\lambda = 0.04$)	0.663	0.607	+0.090 [+0.057, +0.123]	2.48
TC-MASS Set-Transformer	0.677	0.620	+0.103 [+0.067, +0.143]	2.48

Key Comparison: Re-prompting GPT-4o (CoT / Self-Verify) yields $\leq +1.3\%$ non-significant gains, whereas offline set-exact refinement adds **+9.0%** to **+10.3%** with zero extra LLM calls.

Group Split & G2→G3 Transfer



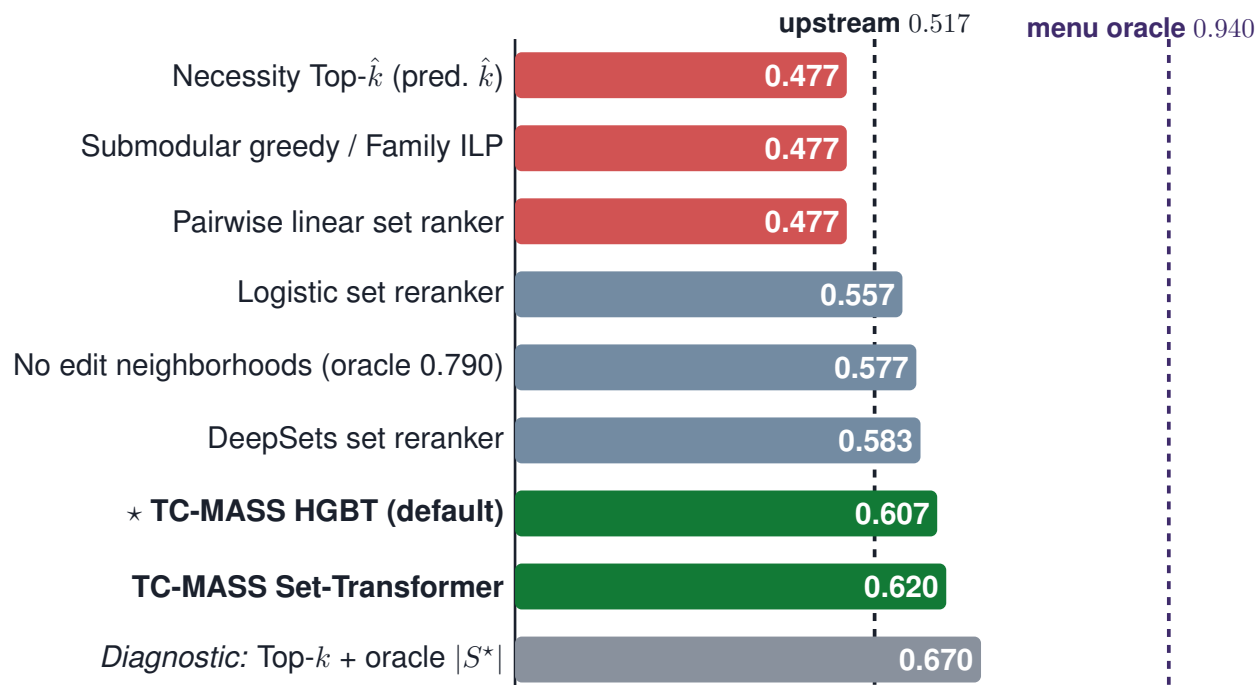
- **Where the 29 flips come from:** 28 gains (27 omission repairs + 1 over-call prune) vs. 1 regression ($p < 10^{-4}$).
- **Harder strata gain most:** +0.250 on cross-family G3 (0.270 $\rightarrow$ 0.520) and +0.202 on $|S^*| = 3$.
- **5 fold-shuffle seeds:** 0.590 ± 0.008 exact ($\Delta = +0.073 \pm 0.008$, every CI > 0).

5. Mechanism Ablations & Oracle Coverage

NeurIPS 2026 · Slide 6 / 10

Isolating set-level supervision from per-API scoring and cardinality prediction

Same Candidate Menu & Per-API Features (5-Fold CV, $n = 300$)



Controlled Comparison: Every row above (except the no-edit ablation) evaluates the *exact same* candidate menu $\mathcal{S}_x$ and per-API necessity scores f_{api} .

Three Mechanism Takeaways

- Decomposable ceiling (0.477):** Top- $\hat{k}$, submodular greedy, family-penalty ILP, and pairwise linear ranking all collide at 0.477 (-0.040 vs. upstream), matching Proposition 1.
- Set supervision drives the jump:** On the *exact same* menu (oracle = 0.940), set-level BCE lifts exact match from 0.477 to 0.607 (HGBT) and 0.620 (Set-Transformer).
- Menu edits & features:** Removing edit neighborhoods drops oracle coverage $0.940 \rightarrow 0.790$ (0.577 exact); dropping cardinality gives 0.580; keeping only the 4 structural features still reaches 0.587 ($+0.070$).

Conclusion: Neither per-API ranking nor heuristic diversity penalties can substitute for proper whole-set supervision.

6. Cross-Distribution & Cross-Upstream Regimes

NeurIPS 2026 · Slide 7 / 10

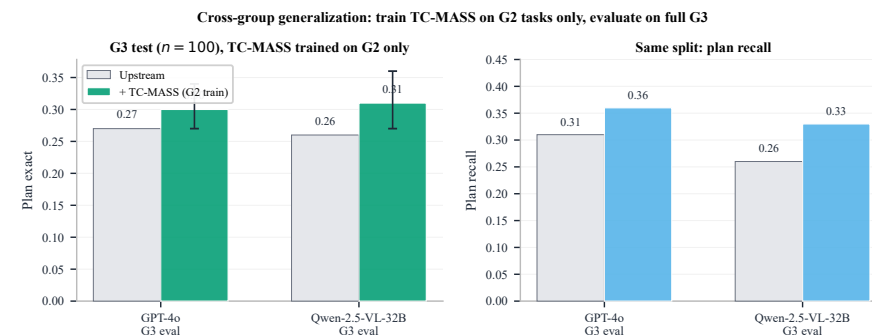
Zero-shot transfer across 2,201 evaluations follows Proposition 2

External Validity Grid (Zero-Shot Transfer)

Regime & Benchmark	Upstream	n	Transfer Result
Regime A: Multi-tool ($S^* > 1$)			
HSR-Bench Full Trace	Qwen-2.5-VL-32B	300	+0.110 [+0.073, +0.150]
ToolBench G2-Train $\rightarrow$ G3-Test	GPT-4o / Qwen	100	+0.030 / +0.050
Regime B: Near-saturated			
BFCL parallel_multiple	GPT-4o Direct	200	Small negative (near ceiling)
Regime C: Single-tool ($S^* = 1$)			
BFCL multiple	GPT-4o Direct	200	Prior mismatch $\rightarrow$ recovery
BFCL live_multiple	GPT-4o Direct	1053	Prior mismatch $\rightarrow$ recovery
APIBank-real	GPT-4o Direct	48	Prior mismatch $\rightarrow$ recovery

Deployment Rule: Estimate target $|S^*|$ on a small pilot batch; if single-tool ($|S^*| \approx 1$), replace the multi-tool cardinality prior with fixed upstream cardinality $\hat{k} = |\hat{S}_0|$ ($\lambda = 0.30$).

Out-of-Group Generalization (G2 $\rightarrow$ G3)



Zero API Overlap: Trained *only* on G2 (all G3 API and task IDs held out), TC-MASS improves full-G3 exact match on both GPT-4o (0.270 $\rightarrow$ 0.300) and Qwen-2.5-VL-32B (0.260 $\rightarrow$ 0.310, +0.050).

7. Method-Blind Path Judges & Full-Library Robustness

NeurIPS 2026 · Slide 8 / 10

Attributed functional evaluation on changed sets + retrieval/planner stress tests

Gold-Free Judges on Changed Predictions ($n = 73$)

Paths are cached by *selected API set* and scored by three judges outside the generator family:

Method-Blind Judge	Upstr.	TC-MASS	Δ Pass	G/L (p)
Qwen3.7-Max ($n = 73$)	0.794	0.904	+0.110	11/3 (0.057)
Gemini-3.6-Flash ($n = 73$)	0.753	0.931	+0.178	15/2 (0.002)
Grok-4.5 ($n = 73$)	0.616	0.808	+0.192	20/6 (0.009)
3-Family Majority ($n = 73$)	0.740	0.918	+0.178	15/2 (0.002)
Majority, non-empty ($n = 62$)	0.726	0.903	+0.177	13/2 (0.007)

Why only changed sets? On the 83% unchanged tasks, any verdict difference is path-generation variance, not API selection quality.

Retrieval & Upstream Stress Tests

- **Full 1,129-API library retrieval ($K = 20 \rightarrow 400$):** Rebuilding C with TF-IDF, BM25, or Jaccard keeps Δ exact positive across all 7 settings (+0.010 to +0.030), peaking at $K \in [50, 100]$ (417–733 ms/task) while menu size saturates near 195 sets.
- **Degraded retrieval recall:** Dropping gold APIs at $p \in \{0.25, 0.50\}$ keeps Δ exact at +0.050 and +0.033; oracle retrieval raises it to +0.137 (0.653 exact).
- **Noisy upstream planner:** Under call drop/swap/add ($q \in \{0.25, 0.50\}$), TC-MASS gain *grows* up to +0.257 because pool coverage stays near 0.94.

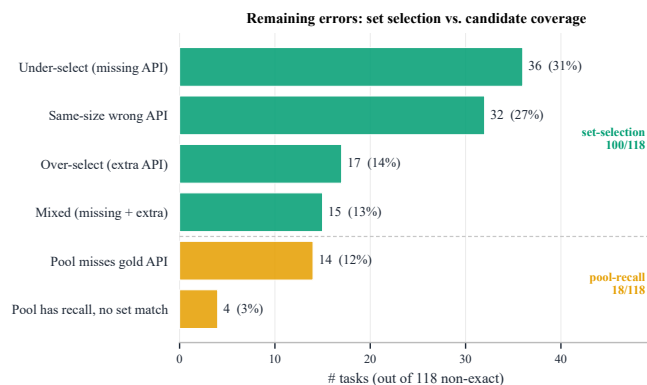
Stress-Test Takeaway: TC-MASS remains net-positive across full-library retrieval, degraded candidate recall, and corrupted upstream traces.

8. Residual Error Taxonomy, Sensitivity & Label Audit

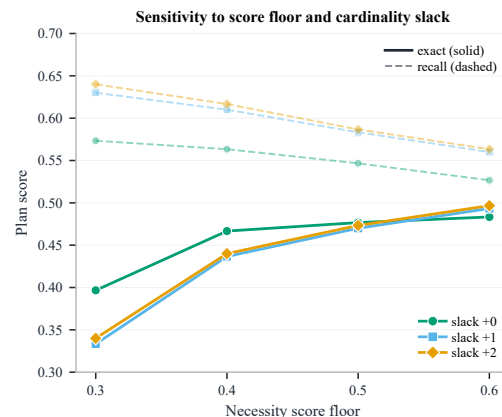
NeurIPS 2026 · Slide 9 / 10

Most remaining errors are set-selection ties or benchmark label mismatches

Residual Errors ($n = 118$) & Hyperparameter Sensitivity



(a) Residual-error taxonomy



(b) Floor & slack sensitivity

(a) 100/118 residual errors occur at set selection despite gold coverage in the pool; only 18/118 are pool-recall misses.

(b) Score floor 0.4 and slack -1 sit on a smooth recall–exact trade-off plateau.

100-Task Human + 2-LLM Gold Label Audit

- **Label-bound necessity:** In **47%** of audited tasks, at least one benchmark gold API is a redundant sibling or non-indispensable helper.
- **Residual “errors” ($n = 50$):** **64%** (32/50) of TC-MASS non-exact outputs are judged functionally acceptable label mismatches (human–LLM agreement 0.82–0.88).

0.874 [0.819, 0.926]

Audit-adjusted functional rate
(vs. 0.607 strict verifier exact match)

Implication: Over half of remaining “errors” against static S^* are valid alternative API combinations.

9. Scope, Limitations & Conclusion

NeurIPS 2026 · Slide 10 / 10

Fixed-pool API-set recovery should be trained and evaluated as set selection

What This Study Establishes

- **Formal gap:** When gold API membership is correlated, per-API marginal selectors hit a strict exact-match ceiling (Prop. 1–2); whole-set BCE over a finite menu is Bayes-consistent (Prop. 3).
- **Practical refiner:** TC-MASS edits frozen upstream traces in **19 ms CPU/task** with **0 extra LLM calls**, lifting HSR-Bench exact match by **+0.090** (HGBT) to **+0.103** (Set-Transformer).
- **Attributed validation:** On the 73 changed predictions, 3-family method-blind judges prefer TC-MASS 0.918 vs. 0.740 ($p = 0.002$).

Core Principle: Relevance narrows the candidate pool, but necessity is a joint property of the *complete API set*.

Explicit Scope Boundaries

1. **Verifier-consistent necessity:** Targets benchmark S^* , not execution-proven minimal sets.
2. **Set equality endpoint:** Does not model call ordering, argument binding, or live API runtime errors.
3. **Pool coverage upper bound:** Bounded by $\Pr[S^* \in \mathcal{S}_x]$; cannot recover APIs missed by both retrieval and upstream.
4. **Regime scope:** Designed for retrieval-conditioned, large-pool, multi-tool settings ($|S^*| > 1$).

Reproducibility: Full code, frozen traces, and 2,201 evaluation records are provided in the supplementary material.

Take-Home Message: Relevance narrows the pool, but necessity is a property of the *complete API set*—supervise the set directly.