
When Are Recursive Language Models Useful? A Cost-Aware, Task-Conditional Evaluation

Odysseus Kaloidas
UC Berkeley

Abstract

1 Recursive Language Models (RLMs) handle long inputs by writing programs over
2 externalized context and making additional model calls. For a question asking
3 which label is most frequent for one user, an RLM can infer a program while
4 a route that knows the operation can label and count directly. We ask when
5 the RLM’s added adaptivity justifies its failures, latency and cost. We compare
6 released standard RLM with deterministic operators, retrieval, semantic workers
7 and direct inference under common information, scoring and budget contracts. On
8 controlled operations and Oolong tasks with a known operation family, matched
9 routes generally match or outperform standard RLM at lower cost. A replication on
10 unused Oolong source corpora preserves this direction, although its clustered exact-
11 rate difference includes zero. On open multi-document search, RLM beats BM25
12 on a first untouched 45-query shard, but neither a prospectively locked equal-
13 cap comparison nor a same-query repeat yields an RLM contrast that survives
14 multiplicity correction. Route ordering changes across these rollouts while RLM
15 remains less reliable and 5.8–15.9× costlier than stronger alternatives. Controls
16 separate operation choice, answer formatting and execution failure. Restricting
17 comparison to completed queries can reverse the ordering. We therefore require
18 any deployment advantage to hold against a strong task-matched alternative on all
19 attempted queries.

20 1 Introduction

21 A known aggregation can use a local worker and fixed code; open search instead requires choosing
22 what to retrieve and when to stop. Recursive programs can serve both settings, but capability alone
23 does not determine whether their added adaptivity is worth its failures, latency and cost.

24 RLMs [Zhang et al., 2026] make this distinction consequential: a controller writes Python over
25 externalized context and may recursively query models. Existing studies compare task-agnostic long-
26 context and coding scaffolds, vary depth or training and add self-reflective search, typed combinators
27 and recursive harnesses [Zhang et al., 2026, Bertsch et al., 2026, Alizadeh et al., 2026, Roy et al.,
28 2026, Lumer et al., 2026]. Deployment alternatives include RAG–long-context routing [Li et al.,
29 2024, Chen et al., 2026a], retrieval [Lewis et al., 2020], prompt compression [Jiang et al., 2023],
30 deterministic operators and decomposed workers [Khot et al., 2023]. The relevant comparison is
31 RLM against the strongest affordable route permitted by the task contract.

32 We define a task family by its operation and answer schema. An information contract states the
33 context, tools and prior knowledge available to each route; a budget covers calls or decisions, time
34 and measured cost. We then ask whether adaptive program generation yields a stable accuracy–cost
35 advantage over operation-matched alternatives. This extends cost-aware agent auditing [Kapoor et al.,
36 2025, 2026, Bragg et al., 2026] and differs from RLM trajectory diagnosis: the unit is the deployed

37 route, not recursion depth or one component. Declared specialization, paired uncertainty and rollout
38 stability determine the credibility of the task-conditional route frontier [Yu and Kumbier, 2020].

39 We compare released standard RLM with deterministic, retrieval, semantic-worker and direct routes.
40 Oolong and an RLM trajectory already decompose labeling from aggregation [Zhang et al., 2026,
41 App. E.2]; our study adds a same-row comparison of complete deployed routes. Table 3 summarizes
42 the boundary. We contribute:

- 43 1. An operation-matched protocol that fixes rows and scorers, declares route-specific prior
44 knowledge and reports accuracy, failures, latency, calls and cost. An answer-contract repair
45 separates finalization failure from route quality.
- 46 2. Wrong-family, same-operator, finalization and trace controls that distinguish operation
47 knowledge from tool access and localize differences in semantic judgment, output schema
48 and scaffold reliability; prior RLM trajectory diagnoses do not combine these controls
49 [Zhang et al., 2026, Wang, 2026].
- 50 3. A bounded deployment comparison: known-operation routes lead on the evaluated Oolong
51 slices, while adaptive-search differences remain unresolved. Paired rollout and completion
52 analyses show how the same outputs can support different apparent rankings when the
53 evaluation population changes.

54 2 Operation-Matched Evaluation

55 **Scope.** The audited system is the official `r1ms` v0.1.3 inference-time scaffold: a controller generates
56 Python over externalized context and may call additional models before answering. We compare this
57 implementation with task-matched routes on identical instances. The exact-aggregation, semantic-
58 counting and Oolong tasks have direct standard-RLM comparisons; the classification task is diagnostic.
59 Tested controller changes and official trained-policy/code checks probe robustness, but do not
60 reproduce the original SRLM, λ -RLM or RAH benchmark protocols. Conclusions remain specific to
61 the evaluated implementation and settings.

62 **Tasks and independence.** Table 1 defines three controlled families: exact structured aggregation (T-
63 STRUCT), directional semantic counting (T-MIXED) and standard classification (T-NLP, diagnostic
64 only). Oolong separates local semantic labeling from aggregation. Label-provided slices (LP-80 and
65 LP-150) isolate aggregation and answer formatting rather than end-to-end performance. The standard-
66 unlabeled slice (SU-150) removes those labels but reuses development contexts, so it supplies scale
67 evidence. Two 45-case slices test different limits: the context-disjoint slice (CD-45) excludes prior
68 context windows and spans 1k, 4k and 262k tokens; the validation-source slice (VC-45) uses two
69 previously unused corpora from the official validation split. Oolong supplies no source-document
70 IDs, so neither slice establishes document-level independence.

71 BrowseComp-Plus tests search in fixed pools of 1,000 documents per query [Chen et al., 2026b].
72 The original 80-query study supplies exploratory evidence; a first untouched 45-query shard tests
73 replication. A second untouched 45-query shard adds iterative search and common decision/time
74 caps before scoring. A later seed repeat reuses those same queries and is a rollout sensitivity, not 45
75 new independent cases. LongBench-v2 Academic QA ($N=94$) remains a separate format-validity
76 diagnostic [Bai et al., 2025].

77 **What the typed route does.** Consider an Oolong question asking whether one label is more
78 frequent than another for a named user. The worker labels individual records; fixed code filters that
79 user’s records, counts the two labels and formats their relation. The route knows the operator family
80 but receives no gold local labels. An RLM must instead choose and execute its program. Comparing
81 complete routes therefore tests the value of that additional adaptivity under declared specialization.

82 **Evaluation design.** Figure 1 specifies the information available to each route. A complete route
83 includes input handling, model or tool calls, answer finalization and failures. Task-specific operators
84 and answer schemas are intentional deployment specialization, so comparisons are not pure scaffold
85 effects. Wrong-family, same-tools and finalizer controls test whether an observed advantage comes
86 from operation knowledge, access to tools or answer formatting. Frozen rows, scorer versions,
87 decision/time caps and failure records link each result to its comparison contract.

Table 1: Controlled task-family construction. Each family targets a different long-context operation with a prespecified alternative route.

Family	Operation	N	Matched alternative
T-STRUCT	Exact pair enumeration	30	SLM extraction + fixed operator
T-MIXED	Directional semantic counting	30	Direct LLM / local judgments
T-NLP (diagnostic)	Standard classification	24	SLM direct

Table 2: Model roles and their corresponding evaluation settings.

Role	Model	Use
RLM controller	Gemini 2.5 Flash; GPT-5; Qwen3.6-35B-A3B	Standard scaffold; direct matched Gemini; BrowseComp matched routes
Trained/code RLM	Qwen3-30B-A3B; λ /Qwen3-8B	Frozen CD-15 diagnostics
Direct LLM	Gemini 2.5 Flash; Qwen3.6-35B-A3B	Direct-context and matched-controller baselines
SLM worker	Gemma 4 26B-A4B	Direct, CoT, labeling and fallback routes
Cheap controller	GPT-4o-mini	Deliberately weak cheap controller; T-NLP orchestration only

Baselines and model roles. Table 2 identifies the models. The primary Oolong RLM controller is Gemini 2.5 Flash; T-STRUCT uses GPT-5. All BrowseComp routes use Qwen3.6-35B-A3B. There, the text route issues one batch of textual subqueries, while iterative search updates its lexical query after each retrieved batch; neither generates and executes Python. Within each endpoint, decoding is matched: the first 45-query and equal-cap endpoints use temperature 0.7, seed 20 and thinking disabled; the repeat changes only the seed to 21. The original 80-query endpoint and both judges use temperature 0. Self-hosted models are pinned by revision and container digest. Appendix H gives the full settings, hosted-model identifiers and protocol differences.

Evidence roles. LP isolates aggregation; SU tests end-to-end scale on reused contexts; VC tests unused-source transfer; and CD tests specialization and shared tools. For search, the prospectively locked equal-cap endpoint is primary, the two earlier studies provide context and the same-query seed repeat is a rollout sensitivity. These roles are not pooled.

Estimands, scoring and uncertainty. For paired routes r, s on n rows, let Y_{ir} be exact correctness and C_{ir} measured cost for row i under route r . We estimate $\hat{\Delta}_A = n^{-1} \sum_i (Y_{ir} - Y_{is})$ and $\hat{\Delta}_C = n^{-1} \sum_i (C_{ir} - C_{is})$; failed or invalid executions have $Y = 0$. A route is empirically dominated if another has no lower observed accuracy and lower cost. This point-estimate relation does not resolve uncertainty or supply a deployment-specific value for accuracy.

Paired accuracy comparisons use exact McNemar tests. For routes reported in order, b counts cases correct only for the first and c cases correct only for the second. BrowseComp difference intervals use 50,000 paired-row bootstrap draws. Oolong intervals resample context-window clusters (5,000 draws): 72 for development $N=120$, 73 for SU-150, 32 for CD-45 and 20 for VC-45. Intervals are conditional on observed executions. Within each equal-cap rollout, Holm correction covers the three RLM comparisons against BM25, text and iterative search. Subgroup, controller and mechanism probes are descriptive; there is no global familywise claim.

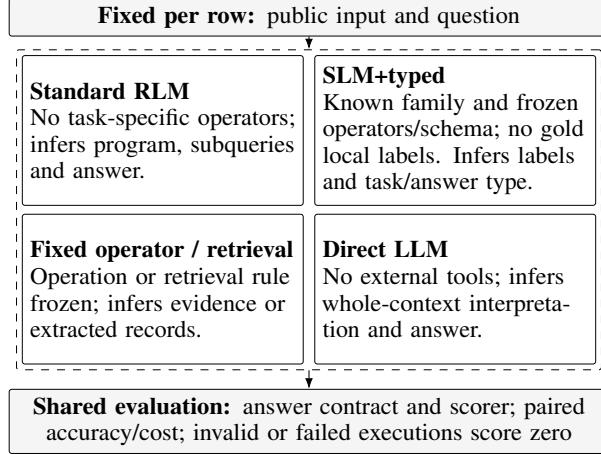


Figure 1: Route and information contracts. Allowed deployment prior knowledge is declared before comparison. The estimates compare complete deployed routes. They do not isolate a scaffold-only causal effect or evaluate an automatic router.

Oolong reports binary exact accuracy and the benchmark’s partial-credit mean score: numeric near-misses receive $0.75^{|\Delta|}$ and relative-frequency answers accept the prescribed relation substrings. BrowseComp uses a method-blinded Gemma 4 semantic judge and a separate GPT-5.6 audit. Both compare predictions with benchmark gold answers; cross-judge agreement measures stability, not human validity. MCQ diagnostics map predictions back after answer shuffling.

Cost accounting. RLM cost is decomposed into controller generation, child-query execution and final extraction. API totals capture successful completions and are lower bounds where failed calls return no usage. Direct and SLM baselines are single-call unless retrieval-augmented; fixed operators have zero marginal model cost but nonzero engineering cost. Engineering amortization is an explicitly hypothetical sensitivity (Appendix A), not measured labor or total cost of ownership. Self-hosted BrowseComp costs use active runtime by route with generation-plus-judge total reported separately. Reported statistics and cost totals are computed from stored logs.

3 Results

3.1 Known-operation tasks

Matched routes lead on controlled tasks. On T-STRUCT, Gemma extracts records and a fixed Python operator performs the aggregation, which is exact by construction, so this row measures extraction success and cost. It reaches 100% accuracy versus 96.7% for RLM at about $99\times$ lower captured deployed-configuration API cost; because both are near ceiling, the result establishes cost dominance only, and supports neither a reliable accuracy difference nor a pure matched-model scaffold-overhead estimate. T-MIXED requires semantic judgments. A matched Gemini ablation separates finalization from semantic error: the count-only finalizer returns no valid answers while the JSON finalizer returns valid answers on all 20 examples and solves 13. A frozen $N=30$ endpoint compares direct Gemini with JSON-finalized RLM: direct Gemini solves 28/30 at \$0.0223, RLM solves 18/30 at \$0.1089 (paired exact $b=12, c=2, p=0.0129$) and remains about $4.9\times$ costlier. We treat T-NLP only as an interface diagnostic because no official RLM accuracy run is logged; moreover, the classification target requires categorical labels while the current aggregation scaffold assumes numeric-style answers.

Oolong specialization persists across labels and sources. On label-provided LP-80, RLM reaches 82.4% mean score against 70.0% for a generic fixed-parser/SLM fallback, although that difference is uncertain ($p=0.089$); its largest gain is on timeline questions and deterministic parsing remains perfect on counting. Gold-label typed operators reach 100% on the same split, so RLM is not the strongest LP-80 route. On frozen held-out LP-150, RLM reaches 84.7% exact while typed

Table 3: Main route results and boundaries. LP-80 is mean score; LP-150 and all other accuracy entries are exact. “Best non-RLM” prioritizes accuracy on T-MIXED. T-STRUCT compares deployed configurations; CD-45 is diagnostic. Equal-cap BrowseComp active-runtime cost is \$13.279 (\$15.015 including lower-bound interruption billing and judging).

Role / regime	Best non-RLM	Standard RLM	Boundary supported by this row
T-STRUCT controlled aggregation	Gemma extract + fixed agg.: 100%, \$0.0002/ex	96.7%, \$0.0198/ex	Near-ceiling accuracy; about 99× lower captured cost in this cross-model comparison. This does not isolate scaffold overhead.
T-MIXED controlled count	Direct Gemini: 28/30, \$0.0223 total	RLM: 18/30, \$0.1089 total	Same-model direct inference wins on the frozen $N=30$ endpoint; the $N=20$ finalizer ablation isolates answer-contract failure.
Oolong CD-45 mechanism controls	Correct-family SLM+typed: 36/45; wrong-family: 2/45	Same-tools RLM: 31/45 or 29/45; GPT-5/GPT-5.6 Luna: 23/45, 26/45; at least 2.3× costlier	Choosing the operator family matters here; shared tools and tested controllers do not reverse the ordering.
Oolong LP-80/LP-150 aggregation-only	Typed ops: 100% / 97.3%	RLM: 82.4% / 84.7%	Provided labels isolate aggregation and answer formatting; these are not end-to-end results.
Oolong SU-150/VC-45 end-to-end	SLM-labeler + typed ops: 98/150, \$0.839; unused-source VC-45 35/45, \$0.312	RLM 68/150, \$4.145; VC-45 28/45, \$1.139	SU-150 reuses contexts. VC-45 tests new sources: its exact-rate difference includes zero, while its mean-score difference does not.
BrowseComp equal-cap primary ($N=45$)	Text: 37/45, \$1.400; iterative: 33/45, \$1.710; BM25: 25/45, \$0.296	RLM: 31/45, \$9.873	All three RLM comparisons remain unresolved after Holm correction; RLM costs 5.8–7.1× more than iterative/text routes.

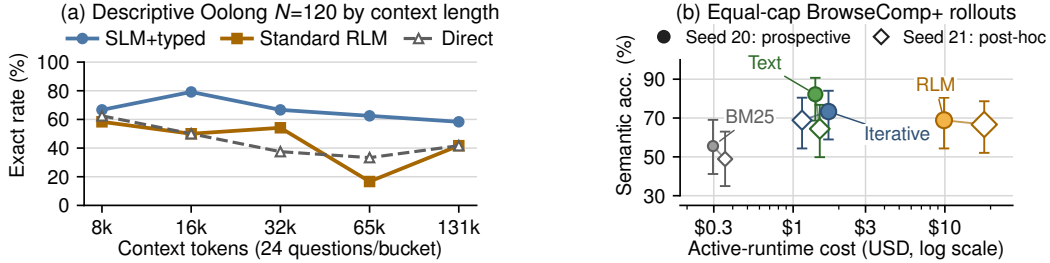


Figure 2: Accuracy and cost across task regimes. (a) Descriptive exact rates on standard-input Oolong development data ($N=120$; 24 per bucket). The inferential contrast is the paired cluster-resampled difference reported in the text, not a per-cell interval. (b) Accuracy versus active-runtime cost on the matched-Qwen equal-cap endpoint (filled circles: prospective seed 20; hollow diamonds: same-query seed 21; $N=45$ each). Whiskers are Wilson intervals, and larger markers mean more all-row model calls (RLM/iterative/text/BM25: 476/426/311/42 and 667/276/312/42). No RLM contrast survives Holm correction in either rollout.

144 rules fixed before evaluation reach 97.3% ($b=23, c=4, p<0.001$). Gold local labels make these
145 aggregation/output-schema ablations, not end-to-end Oolong results. They show that RLM handles
146 flexible schemas better than a generic parser but not that open-ended Python is uniquely required
147 when the schema family is known.

148 All central standard-unlabeled checks use `context_window_text` without gold local labels. An
149 $N=120$ development replication established the SLM>RLM direction across every 8k–131k bucket
150 (Figure 2(a)), with a paired 22.5-point exact-rate advantage ($b=37, c=10, p=9.8 \times 10^{-5}$; 72-cluster
151 interval 11.9–33.0). The fixed pipeline was evaluated on SU-150, CD-45 and VC-45, which was
152 frozen before model outputs were generated. Before scoring, one parser repair was applied only to
153 rows returning no call, a condition visible without labels.

154 On SU-150, SLM+typed solves 98/150 at \$0.839, versus RLM 68/150 at \$4.145 and direct Gemini
155 65/150 at \$1.939. Discordance favors SLM over RLM ($b=44, c=14, p=1.00 \times 10^{-4}$); no direct–RLM

156 difference is detected ($p=0.788$). The 73-cluster interval gives a 10.6–29.9 point SLM advantage.
 157 Because all rows reuse previously evaluated context windows, SU-150 shows that the locked method
 158 retains its advantage at larger N but is not an independence check.

159 VC-45 removes the known source-corpus overlap: its official validation rows come from spam
 160 and `trec_coarse`, absent from the test split and all previous Oolong evaluations. A replay of
 161 the unchanged question parser agrees with the recorded task and answer-type metadata on all 150
 162 SU-150, 45 CD-45 and 45 VC-45 rows. This validates the parser on these benchmark templates, not
 163 on arbitrary paraphrases or unseen operator families. SLM+typed solves 35/45 at \$0.312, versus RLM
 164 28/45 at \$1.139 and direct Gemini 23/45 at \$0.658. SLM–RLM discordance is $b=10$, $c=3$, $p=0.092$
 165 (Holm $p=0.185$); the 20-cluster exact-difference interval is 0.0–31.1 points while the mean-score
 166 interval is 3.0–30.9. RLM reaches 13/15 on user questions but 6/15 on counting. Task and source
 167 are not factorially balanced, so subgroup differences are descriptive. The specialized route remains
 168 cheaper and directionally favored under source-corpus transfer but VC-45 alone establishes neither a
 169 significant exact-rate advantage nor document-level independence.

170 CD-45 retains the specialization controls: using the wrong operator family collapses the typed route
 171 from 36/45 to 2/45. Giving standard RLM the same operator library yields 31/45 when the library
 172 includes a high-level convenience solver and 29/45 with only lower-level atomic tools, but remains
 173 at least $2.3\times$ costlier. An independent standard-RLM repeat drops from 27/45 to 21/45. Together,
 174 SU-150 supplies scale, VC-45 supplies source-corpus transfer and CD-45 isolates specialization and
 175 same-tools behavior; none is a full Oolong leaderboard.

176 **Tested controllers preserve the Oolong ordering.** Strict-stop GPT-5 and GPT-5.6 Luna standard-
 177 RLM runs solve 23/45 and 26/45 CD-45 cases, below SLM+typed’s 36/45 (descriptive paired
 178 $p=0.00235$ and $p=0.041$). On CD-15, the official trained Qwen3-30B-A3B policy solves 10/15 and
 179 unmodified public λ -RLM code with Qwen3-8B solves 8/15, versus 15/15 for SLM+typed. The λ run
 180 counts two 1,800-second timeouts incorrect. These sensitivities do not reverse the observed ordering,
 181 but differing populations and fidelity tiers preclude pooling. They support a limited robustness claim,
 182 not full official-family benchmark parity (Appendix A).

183 3.2 Adaptive search

184 **Stronger alternatives leave the comparison unresolved.** BrowseComp-Plus supports controlled
 185 retrieval/agent comparisons. We separate three stages: an exploratory 80-query study, a first untouched
 186 45-query replication and the prospective equal-cap comparison below. The exploratory study gave
 187 RLM/text/BM25 54/48/39 correct, but completion across two episodes after partial score visibility
 188 limits its evidential role. On the first untouched shard, the same routes scored 33/29/21. RLM
 189 exceeded BM25 by 26.7 points (paired 95% interval 11.1–44.4; Holm $p=0.0151$), while its 8.9-point
 190 lead over text remained uncertain (−6.7–24.4). RLM completed 42/45, cost \$6.746 and had a
 191 425.7-second p95 latency; text completed 45/45 at \$1.610 and 44.9 seconds. A blinded second-model
 192 audit agreed on 128/129 successful outputs ($\kappa=0.983$). The positive RLM–BM25 result is retained;
 193 it does not establish superiority to adaptive search.

194 The prospective equal-cap endpoint tests four routes on a second untouched 45-query shard. We added
 195 iterative lexical search and attempted all 180 Qwen3.6 route cells under common 20-decision/1,200-
 196 second caps before semantic scoring. RLM scores 31/45 (68.9%), versus iterative 33/45 (73.3%),
 197 text 37/45 (82.2%) and BM25 25/45 (55.6%). Differences are −4.4, −13.3 and +13.3 points,
 198 with discordances 6/8, 5/11 and 10/4. Paired 95% intervals are [−20.0, 11.1], [−31.1, 4.4] and
 199 [−2.2, 28.9] points; no contrast survives Holm correction ($p_{\text{Holm}} \geq 0.539$). The second judge closely
 200 agrees ($\kappa=0.985$).

201 RLM completes 41/45 with 50.1/628.8-second median/p95 latency. Iterative/text complete all 45
 202 with 14.1/90.2 and 22.1/47.6 seconds; BM25 completes 42/45. On 39 common completions, scores
 203 are 30/28/31/22; this post-treatment sensitivity leaves the adaptive routes near parity. A question-only
 204 Qwen leakage probe scores 0/45 under both judges: neither finds a correct answer without the corpus.
 205 This endpoint does not establish the earlier RLM–BM25 advantage; because both queries and the
 206 resource contract changed, the difference cannot be attributed to equal caps alone; Figure 2(b) plots
 207 both equal-cap rollouts against active-runtime cost. Open search alone is not a sufficient routing
 208 signal.

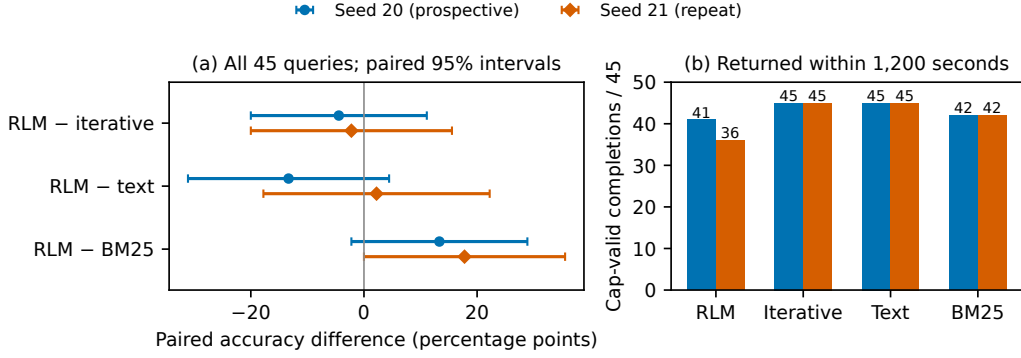


Figure 3: Paired accuracy and completion on the same 45 search queries. (a) RLM minus each alternative, with paired 95% bootstrap intervals conditional on each rollout; these marginal contrast intervals are not simultaneous confidence bands. Holm-adjusted tests resolve no RLM advantage. (b) Outputs returned within the 1,200-second cap. The repeat has 37 successful RLM executions but only 36 cap-valid completions. Dropping incomplete cases changes the comparison population.

209 **Rollouts and completion subsets answer different questions.** A post-hoc seed-21 repeat also
 210 leaves the RLM comparisons unresolved. RLM/iterative/text/BM25 complete 37/45, 45/45, 45/45
 211 and 42/45 and score 30/31/29/22; no RLM contrast survives Holm ($p_{\text{Holm}} \geq 0.346$). A cap-valid
 212 completion finishes within 1,200 seconds: RLM has 36/45, and on 34 rows cap-valid for every
 213 route the scores are 28/23/20/18. This post-treatment subset puts RLM first, but selecting cases
 214 by completion changes the population; it does not identify a causal effect of attrition. RLM costs
 215 12.1–15.9 $\times$ the adaptive routes. Sol agrees on 166/169 outputs ($\kappa=0.961$); the repeat is not pooled
 216 with the prospective endpoint.

217 On seed 21, RLM trails iterative search by 2.2 points over all 45 queries but leads by 14.7 points
 218 on the 34 common cap-valid completions: a 16.9-point change from selection alone. For text,
 219 the corresponding differences are +2.2 and +23.5 points. A paired comparison of the RLM–text
 220 difference across seeds changes by 15.6 points, with a case-bootstrap interval of $[-8.9, 40.0]$. This
 221 descriptive interval is conditional on the two observed rollouts; it does not establish a population-level
 222 seed effect. Figure 3 displays the all-case comparisons and their completion denominators.

223 **Implications for route choice.** Known operation families warrant fixed-operator and semantic-
 224 worker baselines. Open search warrants strong retrieval and adaptive nonrecursive alternatives. Our
 225 results support these comparisons, not an automatic router: flexible schemas do not uniquely require
 226 generated Python, and open search alone does not establish an RLM advantage.

227 4 Mechanism Analysis

228 Failures occur in task interpretation, evidence selection, aggregation, tool use and finalization rather
 229 than one RLM component. Appendix C.1 groups these observations by known-operator overhead,
 230 answer-contract sensitivity, semantic-local error, output-schema rigidity, search instability, context
 231 constraints and multiple-choice bias.

232 **Additional orchestration need not help.** A strong LLM can delegate local interpretation to cheaper
 233 workers [Narayan et al., 2025], but this adds little when the operation is already fixed. In the T-NLP
 234 diagnostic, LLM+SLM scores 83.3% versus 87.5% for SLM direct at about 3.4 $\times$ the cost. This
 235 supports checking a direct worker before adding orchestration; it is not an official RLM comparison.

236 **Traces expose failures at distinct stages.** CD-45 case 710070026 illustrates a typed-route failure:
 237 all seven local labels are correct, yet the route returns “contradiction” rather than the gold least-
 238 frequent label “neutral”; RLM and direct inference are correct. Thus accurate local labeling does
 239 not guarantee correct aggregation or tie handling. Conversely, case 312030020 is solved by RLM
 240 and the typed route while direct inference returns the wrong user. These outcome-selected examples

241 explain distinct failure stages, not their prevalence (Appendix C.2). Stored T-MIXED traces likewise
242 show successful dual-gate filters but also vocabulary overreach and brittle finalization. These traces
243 localize where failures surface; they do not attribute them to the scaffold as a whole.

244 **Multiple-choice position bias limits the QA diagnostic.** LongBench-v2 Academic QA is multiple-
245 choice so answer-position priors confound retrieval-vs.-RLM comparisons. On the $N=50$ discovery
246 set, canonical BM25 predicts A 68% of the time while gold-A is 18%. Rotating priming improves
247 pooled $N=94$ ($p=0.015$) but not held-out $N=44$ ($p=1.0$); single-call accuracy is unmeasured. We
248 therefore keep it as a validity diagnostic, with prior debiasing comparisons in Appendix B.

249 5 Related Work

250 RLM externalizes context through controller-generated programs and recursive calls [Zhang et al.,
251 2026]. Follow-ups test whether recursion drives gains or add typed operators, coding agents and
252 recursive harnesses [Alizadeh et al., 2026, Roy et al., 2026, Ehrlich and Blackman, 2026, Cao et al.,
253 2026, Lumer et al., 2026]; Wang reports task-dependent depth, cost and strong-model degradation
254 [Wang, 2026]. Unlike trajectory studies, our unit is the complete route, compared on matched rows
255 and scorers with paired uncertainty.

256 Cost-aware agent evaluation measures answers and the resources used to obtain them [Kapoor et al.,
257 2025, 2026, Bragg et al., 2026]; RAG inference scaling allocates a given input-token budget [Yue
258 et al., 2025]. We apply both ideas to complete long-context routes under explicit information, scoring
259 and budget contracts and test recommendation stability across rollouts.

260 Oolong separates local labeling from aggregation, which supports our mechanism ablations [Bertsch
261 et al., 2026]. Long-context suites test reasoning, effective context and retrieval/length effects [Bai
262 et al., 2025, Chen et al., 2026c, Hsieh et al., 2024, Yen et al., 2025, Zhang et al., 2024, Modarressi
263 et al., 2025, Li et al., 2025]. Decomposition, search and routing systems select evidence or execution
264 routes [Xu et al., 2026, Song et al., 2026, Ong et al., 2025, Li et al., 2024, Chen et al., 2026a];
265 search-space scale can also change lexical-versus-agentic ordering [Wang et al., 2026]. Our claims
266 remain limited to the declared fixed pools.

267 6 Limitations

268 **External validity.** We diagnose task-conditional failure modes, not average long-context perfor-
269 mance. VC-45 transfers source corpora but not documents because Oolong lacks IDs; SU-150 reuses
270 windows. Official-family evidence covers CD-15 and the controlled curve ends at 131k; longer rows
271 are stress tests.

272 **Measurement.** The 180 equal-cap cells run under fixed caps across three execution episodes. Cross-
273 family judge agreement and a question-only leakage control assess stability, not human validity; there
274 is no human adjudication amid known contextual judge fragility [Tan et al., 2025, Xu et al., 2025].

275 **Cost and scope.** Costs exclude engineering and TCO; alternatives are not exhaustive and recursion
276 depth is fixed. BrowseComp bins as small as $n=10$ preclude subgroup claims. The below-RLM
277 same-row embedding rerank is not full-corpus dense retrieval (Appendix B). Seven repeat RLM calls
278 exceed the cap and score incorrect; self-hosted and API dollars remain separate. Automated routing
279 could suppress useful methods in untested domains.

280 **Ethics and AI assistance.** Public and synthetic benchmarks involve no human subjects. AI tools
281 assisted coding, experiment orchestration and editing; the authors verified all outputs.

282 7 Conclusion

283 Specialized routes lead on tested known operations at lower cost, while equal-cap search yields
284 no RLM advantage despite $5.8\text{--}15.9\times$ higher cost. Completion filtering can change the winner,
285 so recursion should be judged against operation-matched alternatives using all-query accuracy,
286 completion and cost.

References

- Keivan Alizadeh, Parshin Shojaei, Minsik Cho, and Mehrdad Farajtabar. Recursive language models meet uncertainty: The surprising effectiveness of self-reflective program search for long context. *arXiv preprint arXiv:2603.15653*, 2026. URL <https://arxiv.org/abs/2603.15653>.
- Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pages 3639–3664, 2025. URL <https://arxiv.org/abs/2412.15204>.
- Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R. Gormley. Oolong: Evaluating long context reasoning and aggregation capabilities. In *Conference on Language Modeling*, 2026. URL <https://colm.eventhosts.cc/Conferences/2026/AcceptedPapers>.
- Jonathan Bragg, Mike D’Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefanus Candra, Jason Dunkelberger, Daniel Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Ngoc-Uyen Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=M7TNf5J26u>.
- Weili Cao, Xunjian Yin, Bhuwan Dhingra, and Shuyan Zhou. Coding agents are effective long-context processors. *arXiv preprint arXiv:2603.20432*, 2026. URL <https://arxiv.org/abs/2603.20432>.
- Yiwen Chen, Kuan Li, Fuzhen Zhuang, Deqing Wang, Zhao Zhang, Liwen Zhang, Yong Jiang, Shuai Wang, and Minhao Cheng. Route before retrieve: Activating latent routing abilities of LLMs for RAG vs. long-context selection. *arXiv preprint arXiv:2605.10235*, 2026a. URL <https://arxiv.org/abs/2605.10235>.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhui Chen, and Jimmy Lin. BrowseComp-Plus: A fair and disentangled evaluation benchmark for deep search agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026b. URL <https://aclanthology.org/2026.acl-long.1023/>.
- Ziyang Chen, Xing Wu, Junlong Jia, Chaochen Gao, Qi Fu, Debing Zhang, and Songlin Hu. LongBench pro: A more realistic and comprehensive bilingual long-context evaluation benchmark. *arXiv preprint arXiv:2601.02872*, 2026c. URL <https://arxiv.org/abs/2601.02872>.
- Clint Ehrlich and Theodore Blackman. LCM: Lossless context management. *arXiv preprint arXiv:2605.04050*, 2026. URL <https://arxiv.org/abs/2605.04050>.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What’s the real context size of your long-context language models? In *Proceedings of the First Conference on Language Modeling (COLM 2024)*, 2024. URL <https://arxiv.org/abs/2404.06654>.
- Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, 2023. URL <https://arxiv.org/abs/2310.05736>.
- Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025-May, 2025. URL <https://arxiv.org/abs/2407.01502>.

337 Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary Siegel, Boyi Wei, Tianci
338 Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin,
339 Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong
340 Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn
341 Song, Peter Henderson, Yu Su, Percy Liang, and Arvind Narayanan. Holistic agent leaderboard:
342 The missing infrastructure for AI agent evaluation. In *International Conference on Learning
343 Representations*, 2026. URL <https://openreview.net/forum?id=vUaY1t64ZZ>.

344 Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish
345 Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *Pro-
346 ceedings of the International Conference on Learning Representations (ICLR 2023)*, 2023. URL
347 <https://arxiv.org/abs/2210.02406>.

348 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
349 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe
350 Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural
351 Information Processing Systems (NeurIPS 2020)*, volume 33, pages 9459–9474, 2020. URL
352 <https://arxiv.org/abs/2005.11401>.

353 Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng.
354 LaRA: Benchmarking retrieval-augmented generation and long-context LLMs – no silver bullet for
355 LC or RAG routing. In *Proceedings of the 42nd International Conference on Machine Learning*,
356 volume 267 of *Proceedings of Machine Learning Research*, pages 36846–36867. PMLR, 2025.
357 URL <https://proceedings.mlr.press/v267/li25dv.html>.

358 Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented
359 generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings
360 of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*,
361 2024. URL <https://aclanthology.org/2024.emnlp-industry.66/>.

362 Elias Lumer, Sahil Sen, Kevin Paul, and Vamse Kumar Subbiah. Recursive agent harnesses. *arXiv
363 preprint arXiv:2606.13643*, 2026. URL <https://arxiv.org/abs/2606.13643>.

364 Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun
365 Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. In
366 *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL <https://arxiv.org/abs/2502.05167>.

368 Arvind Narayan, Dan Biderman, Sabri Eyuboglu, Avner May, Scott Linderman, James Zou, and
369 Christopher Ré. Minions: Cost-efficient collaboration between on-device and cloud language
370 models. *arXiv preprint arXiv:2502.15964*, 2025. URL <https://arxiv.org/abs/2502.15964>.

371 Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez,
372 M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data.
373 In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://arxiv.org/abs/2406.18665>.

375 Amartya Roy, Rasul Tutunov, Xiaotong Ji, Matthieu Zimmer, and Haitham Bou-Ammar. The Y-
376 combinator for LLMs: Solving long-context rot with λ -calculus. *arXiv preprint arXiv:2603.20105*,
377 2026. URL <https://arxiv.org/abs/2603.20105>.

378 Maojia Song, Renhang Liu, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, Soujanya Poria, and
379 Jingren Zhou. Demystifying deep search: A holistic evaluation with hint-free multi-hop questions
380 and factorised metrics. In *The Fourteenth International Conference on Learning Representations*,
381 2026. URL <https://openreview.net/forum?id=x4zQDewgHr>.

382 Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang
383 Wang, Raluca Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based
384 judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL
385 <https://openreview.net/forum?id=G0dksFayVq>.

386 Daren Wang. Think, but don’t overthink: Reproducing recursive language models. *arXiv preprint
387 arXiv:2603.02615*, 2026. URL <https://arxiv.org/abs/2603.02615>.

- 388 Pengyu Wang, Benfeng Xu, Shaohan Wang, Mingxuan Du, Xin Zeng, Huarui Wu, Lei Zhang, and
 389 Licheng Zhang. BM25 wins at scale: A scaling study of retrieval-augmented generation paradigms.
 390 *arXiv preprint arXiv:2607.26497*, 2026. URL <https://arxiv.org/abs/2607.26497>.
- 391 Austin Xu, Srikanth Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. Does context matter? Contextu-
 392 alJudgeBench for evaluating LLM-based judges in contextual settings. In *Proceedings of the 63rd*
 393 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages
 394 9541–9564. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.470.
 395 URL <https://aclanthology.org/2025.acl-long.470/>.
- 396 Zhen Xu, Shang Zhu, Jue Wang, Junlin Wang, Ben Athiwaratkun, Chi Wang, James Zou, and
 397 Ce Zhang. When does divide and conquer work for long context LLM? a noise decomposition
 398 framework. In *International Conference on Learning Representations*, 2026. URL [https://](https://openreview.net/forum?id=ddQFUuHDDt)
 399 openreview.net/forum?id=ddQFUuHDDt.
- 400 Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat,
 401 and Danqi Chen. HELMET: How to evaluate long-context language models effectively and
 402 thoroughly. In *Proceedings of ICLR 2025*, 2025. URL <https://arxiv.org/abs/2410.02694>.
- 403 Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*,
 404 117(8):3920–3929, 2020. doi: 10.1073/pnas.1901326117. URL [https://doi.org/10.1073/pnas.](https://doi.org/10.1073/pnas.1901326117)
 405 [1901326117](https://doi.org/10.1073/pnas.1901326117).
- 406 Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong
 407 Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval
 408 augmented generation. In *International Conference on Learning Representations*, 2025. URL
 409 <https://openreview.net/forum?id=FSjIrOm1vz>.
- 410 Alex L. Zhang, Tim Kraska, and Omar Khattab. Recursive language models. In *Advances in Neural*
 411 *Information Processing Systems*, 2026. URL <https://arxiv.org/abs/2512.24601>.
- 412 Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han,
 413 Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞ Bench: Extending long context
 414 evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for*
 415 *Computational Linguistics (ACL 2024)*, 2024. URL <https://arxiv.org/abs/2402.13718>.

416 A Protocol Details

417 **Cost-aware audit protocol.** Table 4 summarizes the protocol used throughout this study. Because
 418 the audited object is a scaffolded inference procedure, a row should not be promoted from diagnostic
 419 evidence to method evidence unless the input contract, output schema, scorer, failure accounting and
 420 cost trace are matched.

421 **Role-based model coverage.** The primary panel assigns models to deployment roles rather than
 422 constructing a broad leaderboard: Gemini 2.5 Flash controls the logged standard RLM and supplies
 423 direct frontier inference, Gemma 4 26B-A4B supplies SLM worker paths and GPT-4o-mini controls
 424 the short T-NLP LLM+SLM route. GPT-5 is used for the T-STRUCT frontier-controller exception
 425 and a CD-45 controller sensitivity. Many central alternatives are deterministic or structural so the
 426 unresolved breadth question is same-contract coverage of stronger RLM-family systems rather than
 427 additional direct-LLM columns.

428 That stronger-family evidence includes frozen-manifest adapter pressure tests, matched standard-
 429 unlabeled smokes, the CD-15 repeatability study and CD-45 same-operator-library diagnostics.
 430 The unmodified public λ -RLM repository attempts all 15 frozen CD-15 IDs with its published
 431 Qwen3-8B settings and reaches 8/15 exact with two declared timeouts. MIT OASYS’s official
 432 `r1m-qwen3-30b-a3b-v0.1` post-trained policy reaches 10/15 exact with one timeout through the
 433 canonical `r1m` harness and model-card Oolong settings. The λ row is official-code cross-task evidence
 434 rather than an original-paper reproduction; rankable official SRLM and same-contract RAH rows
 435 remain unavailable.

Step	Check	Audit rule
1	Scope	Name the scaffold, input field, answer contract, scorer and non-claims before comparing methods.
2	Match rows	Run the scaffold and alternatives on identical instances; separate development, fixed-pipeline and split-validity evidence.
3	Cheap route	Include deterministic code, retrieval/reranking, SLM workers, direct LLM and RLM only where each is natural for the operation.
4	Cost/failures	Record per-example cost, timeout/error rows, retry/stop rules and attempted-cost lower bounds.
5	Confounds	Add finalizer, schema, same-operator, wrong-family, repeatability and trace probes when results could be harness artifacts.
6	Claims	Mark rows as confirmed, diagnostic, smoke, proxy or future; report falsifiers and candidate-positive regimes separately.

Table 4: Cost-aware audit protocol used in this study. Standard RLM is compared only after cheaper operation-matched mechanisms receive the same instances, answer contract, scorer, cost logging and failure accounting.

Slice / row	RLM controller	Row cap	Interpretation
Oolong CD-45 central	Gemini 2.5 Flash	360s	Main standard-rlms row.
Oolong CD-15 repeats	Gemini 2.5 Flash	360s	Repeatability diagnostic over a frozen subset.
Oolong CD-15 official trained policy	Qwen3-30B-A3B + official LoRA	1500s	Canonical-harness, same-contract diagnostic; one timeout is scored incorrect.
Oolong CD-45 same-operator RLM	Gemini 2.5 Flash	360s	Standard-rlms custom-tool diagnostic, not official SRLM/ λ -RLM.
Oolong CD-45 controller sensitivity	GPT-5	240s	Controller+stability sensitivity, not pure controller ablation.
Oolong CD-45 GPT-5.6 complete	GPT-5.6 Luna	360s	Current-controller sensitivity; complete strict-stop row, not official SRLM/ λ -RLM/RAH.
Oolong CD-15 official λ code	Qwen3-8B	1800s	Complete unmodified-repository cross-task row; 15/15 attempted, 13 completed, 8 exact.
LB2 >500k smoke	Gemini 2.5 Flash	480s	Stop-rule feasibility smoke, not stable ranking.

Table 5: Protocol caps for complete claim-central and sensitivity rows. The rows answer different questions and should not be merged as pure model swaps.

436 Controller-freshness and compatibility checks are reported separately from family parity. Table 5
437 gives the contracts for the complete central and sensitivity rows; these checks do not constitute full
438 official SRLM/ λ -RLM/RAH benchmark parity.

439 **Cost accounting.** RLM cost aggregates controller program generation, recursive or child-query
440 execution and final answer extraction. Direct LLM/SLM baselines are single-call unless retrieval-
441 augmented. Fixed operators have zero marginal LLM cost but not zero engineering cost; hybrid
442 pipelines such as T-STRUCT include their SLM extraction calls. Current RLM cost averages are
443 successful-completion costs from row-level logged prediction/cost/error files; not every central

Method	Exact	Acc.	Cost	Invalid ex.
BM25 + Gemma reader	8/30	26.7%	\$0.0355	3
BM25 + Gemini reader	11/30	36.7%	\$0.1588	2
BM25 + Gemma rerank + Gemini	12/30	40.0%	\$0.1863	2

Table 6: Balanced LongBench-v2 train-split >500k-character retrieval/rerank diagnostic ($N = 30$, five examples from each domain, seed 20260714). All rows use local BM25 over the full context and three answer-order votes. The rerank row applies Gemma to score 24 BM25 candidates before Gemini reads the selected excerpts. This is an over-window alternative diagnostic, not a same-row RLM-family comparison.

RLM row retains a full raw generated-program/provider-payload trajectory. The release includes an attempted-cost audit script and JSON artifact that scan local `model_calls*.jsonl` files for official-RLM completion rows rather than any one headline run. Across those 85 logged attempt rows, 76 completed rows record \$1.289 total cost while 9 failed timeout/error rows record zero captured usage/cost. Thus the local logs support a recorded attempted-cost lower bound across archived official-RLM model-call traces but they cannot reconcile to a single Oolong/T-MIXED headline denominator or reconstruct provider-side billing for failed partial scaffold calls whose usage was not returned.

Because person-hour estimates are unavailable, we report a deterministic audit of implementation surface. The complete released typed-route package spans four source files and 2,017 physical lines (1,793 nonblank) including the SLM labeler, typed aggregation kernel, natural-language schema parser, standard-RLM tool wrappers, command-line interfaces, caching, tracing and analysis support. The release declares schemas for six families and five answer types. Three families appear as controlled families in the main paper; the others support supplement-only diagnostics or public anchoring. The standard-RLM diagnostic exposes five atomic/schema tools and one optional convenience solver. These counts are an auditable upper bound on the released implementation surface, not a minimal runtime kernel or a measure of human engineering time, maintenance burden or total cost of ownership.

We also parameterize the unmeasured one-time implementation cost instead of treating it as zero. On SU-150, the observed successful-completion API-cost difference is \$0.022041 per case; amortizing assumed one-time budgets of \$100, \$1,000 and \$10,000 from API savings alone requires 4,538, 45,371 and 453,704 cases when the ratio is computed from unrounded costs and rounded up to the next whole case. On VC-45, the corresponding difference is \$0.018377 and the thresholds are 5,442, 54,416 and 544,158 cases under the same rule. These are deployment-volume sensitivities, not estimates of actual labor or maintenance. They omit failed-call billing, latency value, reliability and future model-price changes; SU-150 also reuses prior windows while VC-45 is source-corpus transfer without document-level independence.

A deterministic cost-sensitivity audit reconciles the stored result files. T-STRUCT’s roughly $99\times$ gap is a deployed-configuration comparison: GPT-5-backed standard RLM costs \$0.019826/example on the stored run while the Gemma extraction plus fixed aggregation route is estimated at \$0.0002/example. It is not a pure matched-model estimate of scaffold overhead. For the Oolong $N = 30$ pilot, the corrected SLM-labeler token-trace price is \$0.007257 versus \$0.255676 for RLM ($35.2\times$); the older pre-patch SLM estimate was \$0.103732 and still leaves RLM $2.46\times$ more expensive. On the Oolong $N = 120$ public-contract run, RLM costs \$2.688340 versus \$0.658995 for SLM+typed ($4.08\times$). The larger $N = 150$ fixed-pipeline replication gives \$4.144901 versus \$0.838777 ($4.94\times$ captured successful-completion cost). Exact prices may change, but the reported cost comparisons survive the disclosed sensitivity checks within their stated deployment scope.

B Diagnostic Validity Controls

Method	Exact	Acc.	Cost	Errors
Standard r1ms	1/6	16.7%	\$0.4698	1
BM25 + Gemma reader	1/6	16.7%	\$0.0065	0
BM25 + Gemini reader	2/6	33.3%	\$0.0284	0
BM25 + Gemma rerank + Gemini	2/6	33.3%	\$0.0326	0

Table 7: Same-row standard-r1ms smoke on the balanced LongBench-v2 train-split >500k-character manifest. The six rows are the shortest eligible context in each balanced domain. This is a stop-rule feasibility check, not a full $N = 30$ ranking.

482 The balanced diagnostic covers Code Repository Understanding, Long In-context Learning, Long
483 Structured Data Understanding, Long-dialogue History Understanding, Multi-Document QA and
484 Single-Document QA. The strongest row, BM25+Gemma-rerank+Gemini, reaches 3/5 on code, 2/5
485 on in-context learning, 2/5 on structured data, 1/5 on dialogue, 3/5 on multi-document QA and
486 1/5 on single-document QA. Reranking changes three rows and improves only one net example
487 over BM25+Gemini. We therefore use the result to calibrate the over-window boundary: cheap
488 retrieval/reranking is a necessary comparison for RLM but this particular route does not close the
489 extreme-context region.

490 **Multiple-choice position control.** On the LongBench-v2 Academic QA discovery set ($N = 50$),
491 canonical BM25 predicts option A on 68% of rows while only 18% of gold answers are A. Rotating
492 answer-order priming improves the pooled $N = 94$ comparison ($p = 0.015$) but not the held-out
493 $N = 44$ slice alone ($p = 1.0$). Single-call accuracy was not separately measured, so this remains a
494 format-validity diagnostic rather than evidence for a retrieval or RLM ranking.

495 **Post-hoc dense-rerank sensitivity.** On the attempted- $N = 49$ BrowseComp-Plus replication rows,
496 a same-row post-hoc route reranks 64 BM25 candidates with Qwen3-Embedding-8B before the same
497 Qwen answer model and Gemma judge. The hybrid scores 20/49 at \$0.516 versus 22/49 at \$0.509
498 for matched BM25 replay and 35/49 for RLM; RLM-hybrid is +30.6 points (interval 14.3–46.9;
499 $p = 0.0026$). Required-document recall rises only from 0.256 to 0.261. This is not full-corpus dense
500 retrieval and all recall diagnostics were computed after output.

501 C Mechanism and Trace Analysis

502 C.1 Mechanism Inventory

Table 8: Mechanism-level error analysis grouped by layer, setting and deployment implication.

Mechanism	Where observed	What fails	Paper implication
Known exact operator	T-STRUCT, Oolong counting	Programs add calls when the symbolic operation is known.	Typed extraction plus deterministic code dominates standard RLM at the observed operating points.
Final-answer contract	T-MIXED, T-NLP, official λ row	Valid content is not finalized under the answer contract.	Evidence localizes a finalization layer but not the whole scaffold.
Semantic local judgment	T-MIXED	Generated lexical programs miss directional meaning.	Direct local judgment beats repaired RLM in this family.
Output-schema rigidity	Oolong timeline/user	A generic parser assumes numeric outputs where labels or user IDs are required.	Typed operators remove this failure only when the schema family is inferable.
Open search	BrowseComp-Plus	Descriptive route ordering changes across equal-cap rollouts.	Open search alone does not select RLM; matched alternatives, completion and cost tails matter.
Context-window constraint	LB2 >500k	Direct input is infeasible and retrieval can miss distributed evidence.	Over-window inputs motivate decomposition, not an untested RLM benefit claim.
Evaluation-format bias	LB2 multiple choice	Retrieval outputs are sensitive to answer-position priors.	Debiasing is required before interpreting retrieval–RLM differences.

503 C.2 Representative Trace Evidence

Table 9: Representative released CD-45 traces selected by outcome pattern. These are illustrations; aggregate conclusions remain tied to the full frozen rows.

Case	Observed behavior	Localized failure	What it establishes
912090010	Gold is “less common.” SLM+typed, direct Gemini and GPT-5 RLM are correct; standard RLM/Gemini returns “same frequency.”	Relation/count error	Direct and structured routes recover an error made by one scaffold roll-out.
312030020	Gold user is 16226. Both RLM controllers and SLM+typed recover it; direct Gemini returns 72294.	Direct whole-context aggregation error	RLM can be useful on an individual case even when it is not the strongest aggregate route.
710070026	Gold least label is neutral. Standard RLM and direct Gemini are correct; SLM+typed returns contradiction despite 7/7 correct local labels.	Typed tie/normalization edge	Typed pipelines also fail above the local-label layer.
410040043	Convenience same-tools RLM succeeds after 9 iterations, 7 code blocks and 3 tool calls; atomic RLM uses 16 iterations, 15 blocks and 14 calls but emits a non-matching final form.	Tool-interface confusion and finalization	Shared operators can narrow accuracy differences while retaining orchestration overhead.

504 D Label-Provided Oolong Aggregation Ablations

505 The subgroup pattern isolates a mechanism under the label-provided aggregation contract. Counting
506 favors deterministic parsing, whereas timeline/user identification favors RLM over the generic
507 fp+SLM fallback because the fallback does not fix the required output type correctly in advance. The
508 typed proxy localizes this gap to output-schema and harness design: once the Oolong task family
509 and output metadata are expressed as bounded typed operators, the same slice is solved at 100%.
510 This does not show that open-ended generated Python is necessary when labels are already supplied.
511 The benchmark scorer gives three RLM examples fractional credit. Thresholding those scores at 1.0
512 leaves the ordering unchanged: RLM is 64/80 (80.0%) strict exact, versus 56/80 (70.0%) for fp+SLM,
513 43/80 (53.8%) for SLM direct and 37/80 (46.2%) for SLM+CoT. Both the RLM timeline advantage
514 and the typed-proxy repair remain hypothesis-generating diagnostics because the subgroup tests are
515 small, the input is label-provided and the proxy was added after failure inspection.

516 On exact outcomes over the frozen label-provided $N = 150$, typed v1 beats RLM with discordance
517 $b = 23, c = 4$ (two-sided exact $p < 0.001$); the repaired typed row beats RLM with $b = 23, c = 0$
518 ($p < 10^{-6}$). RLM remains substantially stronger than the pure fixed parser ($b = 72, c = 11$,
519 $p < 10^{-10}$). The public λ -RLM implementation pressure test scores 43.1% mean score / 42.0%
520 exact, below standard RLM ($b = 73, c = 9, p < 10^{-12}$) and typed v1 ($b = 86, c = 3, p < 10^{-20}$)
521 and statistically similar to the pure fixed parser on exact outcomes ($b = 27, c = 30$, n.s.). The prehend
522 SRLM-style pressure test is much stronger than the λ adapter on the matched $N = 60$ subset (83.0%
523 mean score; $b = 30, c = 3, p = 1.4 \times 10^{-6}$) and is not significantly different from standard RLM on
524 paired exact outcomes ($b = 4, c = 10, p = 0.180$) but it remains below typed v1 ($b = 0, c = 10$,
525 $p = 0.002$). Its failures are wrong numeric counts (4), wrong relations (3), wrong user IDs (3) and
526 wrong labels (2). Thus the result is not that RLM-family search is weak but that under a gold-label
527 aggregation contract a task-family typed harness is a better comparator than a generic parser fallback
528 or generic recursive/reflective QA adapter.

Group	N	Oracle typed	RLM	fp+SLM	SLM direct
Overall	80	100.0%	82.4%	70.0%	53.8%
Counting	29	100.0%	85.3%	100.0%	65.5%
Timeline	29	100.0%	79.9%	44.8%	37.9%
User	22	100.0%	81.8%	63.6%	59.1%

Table 10: Same-slice label-provided Oolong-synth aggregation ablation ($N = 80$, seed 42, max context 8192 tokens). All methods receive `context_window_text_with_labels` so gold local labels are visible; this is not standard unlabeled Oolong classification-plus-aggregation. RLM/fp+SLM/SLM rows are logged methods under this input contract; typed proxy is a post-hoc deterministic typed-combinator diagnostic with no model/API calls that also uses benchmark task/output metadata. Entries are mean Oolong scores: exact match for categorical/date-style answers, accepted substrings for relative-frequency comparison strings and $0.75^{|\Delta|}$ partial credit for numeric near-misses.

Method	N	Overall	Counting	Timeline/User
Fixed parser	150	44.0%	100.0%	14.0/18.0%
λ -RLM public impl.	150	43.1%	63.2%	32.0/34.0%
prehend SRLM-style	60	83.0%	90.3%	83.8/75.0%
Oracle typed v1	150	97.3%	100.0%	96.0/96.0%
Oracle typed repair	150	100.0%	100.0%	100.0/100.0%
RLM	150	86.9%	89.5%	79.1/92.0%

Table 11: LP-150: frozen held-out label-provided Oolong aggregation ablation (seed 20260712, no overlap with the same-slice LP-80). All rows use `context_window_text_with_labels`; standard unlabeled Oolong is not evaluated here. The v1 typed rules were fixed before new model calls; the repaired row adds two bounded operators after inspecting four misses. The RLM row uses the same OpenRouter Gemini 2.5 Flash `r1ms` configuration as the same-slice run. The λ -RLM row uses the public implementation through a project-side frozen-manifest adapter, default 100k-character context window. The prehend SRLM-style row uses the public fork’s $K = 3$ self-consistency/confidence/trace-length selection on the first balanced $N = 60$ frozen examples. Entries are mean Oolong scores under this ablation.

529 E Standard-Unlabeled Oolong Evaluations

530 **$N = 120$: uncertainty and direct control.** This public-contract replication covers five context
531 lengths and 72 context clusters. Paired exact outcomes favor stored SLM-labeler+typed over standard
532 RLM ($b = 37, c = 10, p = 9.85 \times 10^{-5}$). The N120 three-way controller JSON (full path in the
533 manifest) gives context-cluster intervals of 66.7% [58.2, 75.0] for SLM+typed and 44.2% [35.6, 53.0]
534 for RLM; paired differences are 11.9–33.0 exact-rate points and 14.1–33.7 mean-score points. A
535 post-run no-call schema audit infers task/output type from each question and preserves 80/120 but
536 reuses stored SLM labels rather than rerunning a prospectively locked parser. Direct Gemini, using
537 the RLM controller model without recursion or code execution, reaches 54/120 exact and 50.0%
538 mean score at \$1.546563. It is not detectably different from RLM ($b = 26, c = 25, p = 1.0$; cluster
539 exact-difference interval -10.9–13.0) while SLM+typed still wins ($b = 30, c = 4, p = 6.16 \times 10^{-6}$;
540 12.4–30.3). This is therefore a deployed-stack comparison, not a scaffold-only causal estimate.
541 Counting/timeline/user exact counts are 20/31/29 for SLM+typed, 15/16/23 for direct Gemini and
542 9/19/25 for RLM.

543 **SU-150: scale and overlap.** SU-150 is the largest standard-unlabeled Oolong tier by nominal
544 N ; VC-45 remains the cleanest source-corpus transfer. The parser, typed operators, SLM prompt,
545 scorer and cost rules were fixed before calls. A no-call question-parser audit agrees with metadata on
546 150/150 rows and reproduces the 98/150 SLM+typed result from stored labels. Paired exact outcomes
547 favor SLM+typed over RLM ($b = 44, c = 14, p = 1.00 \times 10^{-4}$) and direct Gemini ($b = 42, c =$
548 $9, p = 3.39 \times 10^{-6}$); direct and RLM do not differ detectably ($b = 26, c = 29, p = 0.788$). Context-
549 cluster intervals give SLM+typed 65.3% [57.3, 73.3] and RLM 45.3% [36.9, 54.0], with a paired
550 advantage of 10.6–29.9 points. These intervals describe evaluated windows, not independence from
551 development contexts: the split audit finds 20 prior ID overlaps and prior-window overlap for all 150
552 rows. Counting/timeline/user exact counts are 23/31/44 for SLM+typed, 18/19/28 for direct Gemini

Method	Exact	Mean	Cost	Err.	\$/exact
Q-parsed SLM+typed	80/120	72.7%	\$0.659	0	\$0.0082
Direct Gemini	54/120	50.0%	\$1.547	0	\$0.0286
RLM	53/120	48.9%	\$2.688	9	\$0.0507

Table 12: Standard-unlabeled Oolong-synth multi-length result on a frozen balanced manifest ($N = 120$, three task groups $\times$ five context lengths $\times$ eight examples; 72 context clusters, max two questions per context). All rows use `context_window_text`; gold local labels are not present. The stored SLM-labeler + typed row is paired with a post-run no-call question-inferred schema audit within the known Oolong task family: it infers task/output schema from the natural-language question before aggregation and gives 100% agreement with benchmark metadata on these stored labels. This is deployability evidence, not a fully prospective open-taxonomy parser validation. Costs are corrected/logged API costs from traced calls.

Method	Exact	Mean	Cost	Err.	\$/exact
Q-parsed SLM+typed	98/150	71.4%	\$0.839	0	\$0.0086
Direct Gemini	65/150	47.8%	\$1.939	0	\$0.0298
Standard RLM	68/150	48.4%	\$4.145	6	\$0.0610

Table 13: SU-150: larger standard-unlabeled Oolong fixed-pipeline replication after freezing the question parser, typed operators, SLM prompt, scorer and cost rules. The manifest balances three task groups across five context lengths (10 examples per group/length cell; 73 context clusters; max context reuse 3). A split-overlap audit finds that it is not context-window-disjoint from earlier Oolong work. All rows use `context_window_text`; gold local labels are not present.

Method	Exact	Mean	Cost	Err.	\$/exact
Question-parsed SLM+typed	35/45	82.1%	\$0.312	0	\$0.0089
Direct Gemini	23/45	55.2%	\$0.658	0	\$0.0286
Standard RLM	28/45	65.6%	\$1.139	0	\$0.0407

Table 14: VC-45: pre-output-locked standard-unlabeled Oolong replication after one outcome-blind no-call parser repair. Its `spam` and `trec_coarse` source corpora do not occur in the official test split or prior project artifacts. Task and length margins are balanced but not every task-source cell; the official schema exposes no source-document identifier.

553 and 14/23/31 for RLM. SLM+typed also has the highest mean score in every length bucket (79.8,
554 67.8, 78.0, 71.8 and 59.5% from 8k to 131k); RLM is closest at 16k but has six timeout/error rows
555 and the highest cost per exact answer. SU-150 is a larger known-family fixed-pipeline replication,
556 not a context-disjoint holdout or open-domain leaderboard.

557 VC-45 was frozen before paid outputs after an exact-schema audit of the pinned Oolong snapshot. The
558 freezer verifies zero overlap with the test split and prior project artifacts on row ID, context-window
559 ID and upstream corpus. A no-call preflight exposed one narrow parser grammar omission for “spam
560 or ham (i.e., not spam)”; admitting the parenthesis before any paid output gave 45/45 parser-valid
561 rows. The post-run question-only parser agrees with benchmark task and answer metadata on all 45
562 rows and exactly reproduces the 35/45 typed result.

563 SLM+typed has the highest point estimate and the lowest captured successful-completion cost.
564 Against standard RLM, paired exact discordance is $b = 10, c = 3$ (two-sided exact $p = 0.0923$;
565 three-pair Holm $p = 0.1846$); the 20-cluster exact-rate difference interval is 0.0–31.1 points while the
566 mean-score difference interval is 3.0–30.9 points. SLM+typed beats direct Gemini with $b = 14, c = 2$
567 ($p = 0.00418$; Holm $p = 0.0125$). No direct–RLM difference is detected ($b = 8, c = 13, p = 0.383$).
568 By group, SLM+typed / RLM / direct exact counts are 10/6/9 for counting, 14/9/8 for timeline and
569 11/13/6 for user. Timeline rows all come from `spam` while counting and user rows span both corpora;
570 subgroup differences are therefore descriptive, not source-independent. Captured retained-row cost is
571 \$2.109802; the OpenRouter account delta is \$2.304960 because an aborted duplicate-worker launch
572 and partial in-flight calls add \$0.195158. All 45 SLM prompts/responses, direct prompts/responses

Method	Exact	Mean	Cost	Err.	\$/exact
Q-parsed SLM+typed	36/45	83.7%	\$0.416	0	\$0.0116
Direct Gemini	21/45	54.9%	\$1.025	0	\$0.0488
Standard RLM	27/45	62.3%	\$1.001	0	\$0.0371
Standard RLM + same typed ops	31/45	69.3%	\$1.314	2	\$0.0424

Table 15: CD-45: project-history context-window-ID-disjoint standard-unlabeled Oolong validation after excluding all previously used example IDs and context-window IDs found in archived artifacts. The manifest balances counting, timeline and user tasks across 1k, 4k and 262k windows, has 32 context clusters and permits at most two questions per window. Gold local labels are absent. This is not proven source-document-disjoint and its length mix differs from SU-150. Bootstrap intervals resample context-window clusters conditional on the observed method executions; a separate RLM repeat probes rollout sensitivity. Costs are captured successful-completion totals and are lower bounds for rows with failures.

and standard-RLM input contexts, raw answers, 205 generated code blocks and trajectories are retained.

CD-45: split validity. CD-45 is the cleanest project-history split check: after scanning prior Oolong result artifacts, its manifest has zero overlap on example ID and context-window ID. The question parser agrees with task/answer metadata on 45/45 rows and the same typed operators aggregate stored SLM labels. Exact outcomes favor SLM+typed over RLM ($b = 11, c = 2, p = 0.0225$) and direct Gemini ($b = 16, c = 1, p = 2.75 \times 10^{-4}$); direct is directionally below RLM ($b = 2, c = 8, p = 0.109$). Context-cluster intervals are 80.0% [68.6, 91.1] for SLM+typed, 60.0% [45.5, 74.4] for RLM and 46.7% [30.4, 63.4] for direct; SLM-RLM intervals are 6.0–35.6 exact-rate points and 7.4–36.1 mean-score points. Counting/timeline/user exact counts are 12/11/13 for SLM+typed, 6/8/13 for RLM and 8/5/8 for direct. At 1k/4k/262k, SLM+typed scores 13/11/12; direct falls to 4/15 at 262k while RLM remains intermediate. Oolong exposes no stronger source-document key and CD-45’s 1k/4k/262k mix differs from SU-150’s 8k–131k range. CD-45 therefore supports split validity under a shifted length mix, not document disjointness or a pure overlap ablation of SU-150.

A no-call selection audit makes that last caveat explicit. SU-150 has 150 rows over 8k/16k/32k/65k/131k windows and 73 context clusters but its archival split audit finds prior context-window overlap. CD-45 has zero prior project example-ID and context-window-ID overlap but uses 1k/4k/262k windows and all three methods score higher: SLM+typed 36/45 versus 98/150 on SU-150, standard RLM 27/45 versus 68/150 and direct Gemini 21/45 versus 65/150. The released JSON/MD selection audit supports the following interpretation: SU-150 supplies larger fixed-pipeline scale and method-lock evidence while CD-45 supplies cleaner split-validity evidence under a different length mix.

A separate no-call specialization stress test asks whether SLM+typed benefits primarily from selecting the correct operation family. It reuses the same stored CD-45 Gemma row-label predictions and changes only that family. The metadata route and natural-language question-parsed route both solve 36/45. An intentionally wrong-family route solves 2/45 and a naive generic most-frequent-label route solves 3/45. These stress routes are not competing methods; they show that the SLM+typed advantage depends on choosing the correct operation family. This supports the paper’s deployment-specialization framing and would have weakened the thesis if the wrong-family route had remained strong.

We also ran a paid shared-operator controller diagnostic on the same CD-45 stored SLM-label predictions. Gemini 2.5 Flash received only the natural-language question, the allowed typed Oolong operator schema and the answer-type schema. It selected the task and answer type and the same typed operators then executed deterministically over the stored row labels. This diagnostic has 45/45 task agreement, 45/45 answer-type agreement, 36/45 exact, 83.7% mean score, zero parse/execution errors and \$0.00886 captured controller cost. It reduces the concern that the typed route depends on hidden metadata routing but it is not standard r1ms, not an official SRLM/ λ -RLM comparator and not evidence that a generic open-ended RLM scaffold would choose the same bounded operator program.

The direct same-operator-library standard-rlms diagnostic addresses that last point. We injected the same typed Oolong operator library through `custom_tools`, first including a high-level convenience solver and then withholding it in an atomic-only ablation while preserving standard rlms free-form program generation, standard `context_window_text` input and the natural-language question as the root prompt. Neither row is the shared-operator controller or official SRLM/ λ -RLM. On CD-45 the convenience row solves 31/45 exact (68.9%, mean score 69.3%) at \$1.314, with two 360-second timeouts, 173 tool calls and 348 generated code blocks. By group it is 11/15 on counting, 9/15 on timeline and 11/15 on user; by length it is 12/15 at 1k, 10/15 at 4k and 9/15 at 262k. Paired exact outcomes still directionally favor SLM+typed/shared-operator over same-operator RLM (SLM-only 7, same-operator-only 2, $p = 0.180$) while same-operator RLM is directionally above direct Gemini (direct-only 6, same-operator-only 16, $p = 0.052$) and not significantly different from the original standard-RLM rollout (original-only 8, same-operator-only 12, $p = 0.503$).

The atomic-only variant retains the schema-list tools and the three functions below but removes the high-level `solve_oolong_with_tools` convenience solver:

```

626         parse_oolong_rows
627         label_oolong_rows_with_slm
628         solve_labeled_oolong

```

It solves 29/45 exact (64.4%, mean score 69.3%) at \$0.959, with four timeout/error rows, 155 tool calls and 337 generated code blocks. By group it is 9/15 on counting, 8/15 on timeline and 12/15 on user; by length it is 11/15 at 1k, 11/15 at 4k and 7/15 at 262k. Paired exact outcomes favor SLM+typed/shared-operator over atomic same-operator RLM (SLM-only 9, atomic-only 2) while the convenience and atomic rows are close but not identical (convenience-only 8, atomic-only 6). The failure trace contains wrong task/output choices, malformed final answers, self-reported tool-use confusion, unsupported typed-solver calls and long-row timeouts. Supplying the operator library therefore helps standard RLM, and the effect does not rest entirely on one convenience solver. Under this bounded protocol, however, free-form orchestration still does not match the cheaper specialized SLM+typed route or deterministic shared-operator execution.

A no-call same-operator program-trace exhibit extracts ten representative cases from the completed convenience and atomic trajectory logs. The artifact covers all 45 rows for both variants, records the aggregate failure-class mix and prints selected generated code blocks without quoting long input contexts. It makes the same-operator mechanism inspectable beyond the separate three-row original-standard-RLM trajectory subset: the convenience row contains 348 generated code blocks and the atomic row contains 337, with examples of deterministic parsing successes, wrong-route choices, malformed finalization, tool-use confusion and timeout/error behavior. This remains qualitative mechanism evidence; the aggregate claims are still the completed 31/45 and 29/45 CD-45 rows above.

The trace examples in Table 16 are not a substitute for aggregate statistics; they explain why the aggregate statistics are plausible. They show three distinct mechanisms: standard RLM can still make ordinary aggregation mistakes, direct whole-context prompting can miss structured counts and SLM+typed can fail when local labels are right but the typed reduction has an output-edge case. The same artifact also freezes the full CD-45 outcome-pattern counts across standard RLM/Gemini, standard RLM/GPT-5, SLM+typed and direct Gemini. It avoids quoting long contexts and records the trace boundary directly: the aggregate CD-45 standard-RLM runs keep row predictions, costs and errors while SLM/direct paths keep prompts, raw responses and token traces.

To inspect generated programs directly, we ran a separate three-row standard-RLM/Gemini trajectory subset on the frozen CD-45 manifest with the rlms logger enabled. This qualitative artifact solves 2/3 rows, costs \$0.015245, records no harness errors and captures 14 generated code blocks. The two successes illustrate different program shapes: a least-label row uses batched local LLM classification followed by a Counter while a most-frequent-user row uses regex extraction and a Counter. The failed label-comparison row also uses batched local classification but its comparison/final-answer path returns “more common than” where the gold relation is “less common than.” The subset supplies program text for mechanism inspection without changing the CD-45 aggregate accuracy, cost or paired-significance estimates.

To estimate rollout variability, we also ran a full independent standard-RLM/Gemini repeat on the same frozen CD-45 manifest, with the same standard/unlabeled context and 360-second row cap. The repeat solves 21/45 exact, has 55.8% mean score, costs \$0.970 and records two 360-second

Case	Route outcomes	What it shows
RLM tie error 912090010, C/4k	Gemini RLM: same frequency (0); GPT-5 RLM, SLM+typed and direct: less common (1).	Fixed comparison repairs an RLM tie after accurate SLM row labels.
RLM label error + timeout 412040009, C/4k	Gemini RLM: False (0); GPT-5 RLM: timeout (0); SLM+typed and direct: true (1).	SLM+typed and direct answer correctly; GPT-5 RLM hits the row cap.
Direct aggregation error 312030020, U/4k	Both RLMs and SLM+typed: 16226 (1); direct: 72294 (0).	Direct selects the wrong most-frequent user; structured routes recover it.
Typed edge miss 710070026, C/1k	Gemini RLM and direct: neutral (1); GPT-5 RLM: timeout (0); SLM+typed: contradiction (0).	A typed tie/normalization edge fails despite exact local labels.
Long-context timeout 218020026, C/262k	Gemini RLM, SLM+typed and direct: more common (1); GPT-5 RLM: timeout (0).	Three routes agree; GPT-5 RLM times out at 262k.

Table 16: Representative CD-45 paired examples from the no-call trace artifact. Parentheses mark exact correctness. The SLM+typed and direct-controller rows have stored prompts/raw responses and token/cost traces. Aggregate CD-45 standard-RLM rows retain predictions, costs and errors; a separate three-row trajectory subset captures generated code blocks for qualitative inspection so this table remains behavior-first rather than a new generated-program benchmark.

668 timeout/error rows. The original and repeated RLM executions are exact on the same 15 examples,
669 wrong on the same 12 examples and disagree on 18 examples: 12 original-only exact and 6 repeat-
670 only exact. Compared with this repeat, SLM+typed has exact discordance 17/2 and remains 36/45
671 exact. CD-45 therefore supports the SLM+typed advantage under a second RLM execution while
672 the original bootstrap interval remains conditional on observed executions rather than covering both
673 sampled contexts and stochastic scaffold rollouts. The repeat also preserves the task-conditional
674 pattern: RLM is best on user rows (10/15 repeat exact) and weaker on counting/timeline (5/15 and
675 6/15), with the two errors on 262k-token timeline/user rows.

676 **Controller and policy fidelity ladder.** The following rows have different evidential roles. Standard
677 rlm with Gemini 2.5 Flash is the primary audited scaffold; GPT-5 and GPT-5.6 Luna are controller-
678 and-stability sensitivities within that scaffold; RLM-Qwen3-30B-A3B is a bounded official trained-
679 policy diagnostic; public SRLM-style and earlier λ -RLM adapters are pressure checks; and the
680 unmodified official λ -RLM repository contributes a complete, rankable CD-15 cross-task row. We
681 therefore do not pool these rows, treat model diversity as independent replication or promote the
682 official-code row to original-paper protocol parity.

683 A separate CD-15 repeatability stop-rule is now complete. The subset was frozen from CD-45 before
684 the repeat calls, balances five examples per task group and five per length bucket, uses 14 context
685 windows with max reuse two and was selected deterministically. On the parent CD-45 outputs,
686 that fixed subset has SLM+typed 15/15 exact at \$0.107, standard RLM/Gemini 10/15 at \$0.308,
687 GPT-5 standard RLM 10/15 at \$0.562 with four errors and direct Gemini 9/15 at \$0.336. Three
688 independent standard-RLM/Gemini repeats with the same standard/unlabeled context and 360-second
689 row cap then produced 13/15 exact at \$0.513, 11/15 exact at \$0.324 and 6/15 exact at \$0.415; their
690 combined captured OpenRouter cost is \$1.252. There are no harness timeout rows but one internal
691 RLM error-string prediction and two free-form non-answer failures are counted incorrect. Only 3/15
692 examples are correct in all three repeats while one relative-frequency row is wrong in every repeat.
693 We intentionally do not present a pooled repeat accuracy: the three outcomes per item are dependent
694 rollout observations, not 45 independent benchmark examples. This is not a benchmark-scale variance
695 estimate but it directly documents that the standard RLM scaffold is stochastic enough that a single
696 small frozen subset can range from RLM-favorable to clearly weak under the same controller and
697 row cap.

698 We replayed direct Gemini on the same CD-15 rows. The chronology matters: the subset was frozen
699 for repeatability after the parent CD-45 outcomes were known and the positive-side interpretation
700 was formulated after the RLM repeats. This is therefore a retrospective descriptive diagnostic, not
701 a prospectively selected positive regime. The direct replay solves 8/15 exact at \$0.341 with zero
702 errors; the three RLM rollouts solve 13/15, 11/15 and 6/15. Averaging the three RLM outcomes

CD-15 route	Exact	Mean	Cost	Err.
Stored SLM+typed	15/15	100.0%	\$0.107	0
Direct Gemini replay	8/15	53.3%	\$0.341	0
Standard RLM repeat 1	13/15	86.7%	\$0.513	0
Standard RLM repeat 2	11/15	78.3%	\$0.324	0
Standard RLM repeat 3	6/15	48.8%	\$0.415	0

Table 17: Retrospective CD-15 rollout diagnostic on the frozen repeatability subset. The direct replay was observed once and the three dependent RLM rollouts are reported separately. Item-clustered uncertainty spans -13.3 to 40.0 points so this is not a powered RLM-over-direct claim. Costs are observed captured costs, not projected repeated-direct spend.

Frozen CD-15 route	Exact	Mean	Complete	Cost
Stored SLM+typed	15/15	100.0%	15/15	\$0.107 API
Stored direct Gemini	9/15	65.0%	15/15	\$0.336 API
Stored standard RLM/Gemini	10/15	70.4%	15/15	\$0.308 API
Official RLM-Qwen3-30B-A3B	10/15	70.4%	14/15	\$1.719 GPU [†]
Official-code λ -RLM/Qwen3-8B	8/15	53.3%	13/15	\$5.450 GPU [‡]

Table 18: Official-policy and official-code diagnostics on the identical frozen CD-15 manifest. [†]The trained-policy cost includes cold start and its declared 1,500-second timeout cap; finalized infrastructure billing was \$12.510 after a teardown overrun. [‡]The λ cost is the creation-to-cleanup listed-rate estimate including two 1,800-second caps; finalized infrastructure billing was \$5.471. Neither row is a full SRLM/ λ -RLM/RAH benchmark reproduction.

703 within each of the 15 unique items gives a 13.3-point advantage over the one observed direct replay
704 but an item-cluster bootstrap spans -13.3 to 40.0 points. This interval is conditional on the three
705 observed RLM rollouts and one direct rollout, not total fresh-rollout uncertainty. We do not report an
706 aggregate McNemar test: duplicating the single direct replay three times would create 45 dependent
707 pseudo-pairs. The stored SLM+typed route remains 15/15 exact at \$0.107.

708 To test whether the result depends on a generic prompted controller, we served the official MIT
709 OASYS RLM-Qwen3-30B-A3B policy (base Qwen/Qwen3-30B-A3B-Instruct-2507, LoRA
710 mit-oasys/rlm-qwen3-30b-a3b-v0.1) on one 80GB-class RunPod GPU and ran it through
711 the authors' current canonical rlm harness. We matched the model-card Oolong inference set-
712 tings: depth one, at most 20 iterations, 4096 controller and child completion tokens, disabled Qwen
713 thinking, the official Oolong instruction plus the natural-language question, standard/unlabeled
714 context_window_text and the same local scorer used by the stored CD-15 baselines. The local
715 official-harness checkout was fixed at commit 72d6940142ddf84ee6be573dc999a37e633e671.
716 No task metadata, local gold labels or typed Oolong operators were supplied.

717 The policy returns 10/15 exact and 70.4% mean Oolong score. It completes 14/15 rows; the final
718 262k-window relative-frequency row returns no response within the predeclared 1,500-second outer
719 cap and is scored incorrect. By task group it reaches 3/5 exact on counting, 4/5 on timeline and
720 3/5 on user. The 14 completed trajectories contain 70 iterations and 70 generated code blocks;
721 completed-row median latency is 15.3 seconds and maximum latency is 72.9 seconds. Against
722 SLM+typed, paired exact discordance is 0 official-only versus 5 SLM-only ($p = 0.0625$); against
723 stored direct Gemini it is 5 versus 4 ($p = 1.0$) and against stored standard RLM/Gemini it is 4 versus
724 4 ($p = 1.0$). The official policy, stored standard RLM/Gemini and stored standard RLM/GPT-5 each
725 reach 10/15 exact on this subset. The diagnostic adds official trained-policy coverage and shows
726 a numerical RLM-over-direct difference, but it neither establishes a powered win nor reverses the
727 specialized route on this slice.

728 The deployment log separates scientific protocol cost from an operational incident. At the listed
729 \$2.99/hour rate, elapsed GPU time from pod rental through cold start, 14 completions and the final
730 timeout cap is \$1.719. Finalized RunPod billing was \$12.510 because the final request outlived the
731 local alarm and the detached pod remained active after the local process died; the \$10.791 difference
732 is teardown overrun, not model compute required by the declared protocol. OpenRouter spend for this
733 run was \$0.00. The anonymous release contains the adjudicated row table, analysis JSON/Markdown

CD-45 route	Exact	Mean	Cost	Err.
Q-parsed SLM+typed	36/45	83.7%	\$0.416	0
Standard RLM / Gemini 2.5 Flash	27/45	62.3%	\$1.001	0
Standard RLM repeat / Gemini 2.5 Flash	21/45	55.8%	\$0.970	2
Standard RLM + same typed ops / Gemini 2.5 Flash	31/45	69.3%	\$1.314	2
Standard RLM / GPT-5	23/45	58.1%	\$2.442	15
Standard RLM / GPT-5.6 Luna	26/45	62.9%	\$3.966	2

Table 19: Completed strict-stop CD-45 controller-freshness sensitivity. The GPT-5 row uses openai/gpt-5 via OpenRouter inside the same local standard-rlms scaffold with a 240-second per-row timeout; the GPT-5.6 Luna row uses the central 360-second cap. These are controller/stability sensitivities, not replacements for official SRLM or λ -RLM reproductions.

Variant	Robust	Strict	Exact	Mean
Original	120/120	120/120	80/120	72.7%
Answer format removed	120/120	29/120	80/120	72.7%
Deterministic paraphrase	120/120	0/120	80/120	72.7%

Table 20: No-call Oolong question-paraphrase stress test over the same frozen $N = 120$ rows. The robust parser receives only the question variant and stored SLM row-label predictions; it does not use benchmark task/output metadata. “Strict parser ok” reports whether the original format-template parser can infer task/output schema from the variant.

734 and compact generated-program trajectories; the 105MB raw trajectory file is retained locally rather
735 than included in the compact release bundle.

736 The completed GPT-5 controller run does not eliminate the CD-45 difference. SLM+typed beats
737 GPT-5 standard RLM on paired exact outcomes ($b = 15$, $c = 2$, two-sided exact $p = 0.00235$), with
738 a context-cluster exact-difference interval of 11.1–47.7 points. GPT-5 is not significantly different
739 from the logged Gemini 2.5 Flash standard-RLM row on paired exact outcomes (GPT-5-only 6,
740 Gemini-only 10, $p = 0.454$) but it is less stable under the strict stop rule: 15/45 rows record timeout
741 or scaffold-error failures. Because the run changes the top-level controller while retaining the local
742 scaffold and not recreating the original GPT-5/GPT-5-mini nested-call configuration, it should be read
743 as evidence that a stronger top-level controller alone does not remove the standard scaffold’s bounded-
744 run reliability problem on this slice, not as evidence against all frontier-controller RLM-family
745 variants.

746 **Current-controller sensitivity.** A complete GPT-5.6 Luna standard-RLM row uses the same 360-
747 second cap as the central Gemini row. Luna completes all 45 rows: 26/45 exact, 62.9% mean score,
748 two timeout/error rows and \$3.965806 captured successful-completion cost. SLM+typed remains
749 higher on the same rows (36/45 exact, \$0.416415), with paired discordance $b = 15$, $c = 5$, two-sided
750 exact $p = 0.041$ and a 32-cluster exact-difference interval of 2.3–40.9 points. Luna is statistically
751 indistinguishable from the logged Gemini 2.5 standard-RLM row (Luna-only 10, Gemini-only 11,
752 $p = 1.0$), directionally above GPT-5 (12 versus 9 discordant exact wins, $p = 0.664$) and directionally
753 above direct Gemini (12 versus 7, $p = 0.359$). It does not reverse the operation-matched SLM+typed
754 decision boundary and is about $9.5\times$ costlier than SLM+typed on captured successful-completion
755 cost.

756 Table 20 is not a new model run; it is a parser/harness stress test. It shows that the original exact-
757 format parser is template-sensitive while a slightly more semantic rule parser still recovers the task
758 and answer-type fields from format-removed and paraphrased questions. This reduces but does not
759 eliminate the Oolong-template concern: the paraphrases are deterministic perturbations of Oolong
760 questions, not independently authored questions from a new benchmark.

761 Each length bucket has only 24 examples, so we do not interpret the slice as a leaderboard. SLM-
762 labeler+typed remains ahead in every bucket while both accuracy and cost change with context
763 length; the captured RLM/SLM cost multiplier therefore persists into the longer evaluated buckets.

Len.	SLM	RLM	SLM \$	RLM \$	Ratio
8k	16/24	14/24	\$0.0199	\$0.2528	12.7×
16k	19/24	12/24	\$0.0414	\$0.2356	5.7×
32k	16/24	13/24	\$0.0852	\$0.4125	4.8×
65k	15/24	4/24	\$0.1768	\$0.4464	2.5×
131k	14/24	10/24	\$0.3357	\$1.3410	4.0×

Table 21: Length-sliced exact rates and captured successful-completion API costs for the $N = 120$ standard-unlabeled replication. Each length bucket has 24 examples. Recorded standard-RLM scaffold errors by bucket are 0, 0, 1, 5 and 3 respectively.

Noise view	Count	Interpretation
SLM local-label/model noise	20	SLM+typed wrong with row-label accuracy < 0.90 ; cheap workers are strong but not solved readers.
SLM aggregation sensitivity	20	SLM+typed wrong despite row-label accuracy ≥ 0.90 ; small local errors or schema edge cases can flip exact aggregates.
RLM scaffold/reliability noise	9	Captured RLM timeout/scaffold errors, concentrated at 65k/131k windows.
Direct whole-context model noise	66	Direct Gemini wrong without infrastructure errors; removing recursion alone does not solve aggregation.
RLM non-error answer noise	58	RLM wrong without captured scaffold error; most misses are ordinary wrong answers plus a reliability tail.

Table 22: No-call noise-decomposition diagnostic for the $N = 120$ standard-unlabeled Oolong replication, following the divide-and-conquer view that failures can arise from task/local interpretation, length/model degradation, aggregation or scaffold reliability. Categories are operational diagnostics over frozen outputs, not new gold labels.

At 131k, SLM-labeler+typed still beats standard RLM but falls to 14/24 exact, showing that cheap local labeling does not solve the benchmark. RLM is weakest at 65k (4/24 exact, five errors) and most expensive at 131k (\$1.341 for 24 examples). Attempted-cost auditing remains conservative: the stored RLM logs contain 85 official-RLM attempt rows, 76 completions and 9 error rows, all with zero recorded failed-call usage so reported attempted cost is a lower bound rather than an imputed provider bill for failed partial executions. The two-method and three-way score/cost PDFs are included in the released figures/ directory; the three-way files use the n120_threeway_score_by_length and n120_threeway_cost_by_length names.

The cleaner CD-45 split-validity slice gives the complementary long-context cost check at 1k/4k/262k windows. On those rows, SLM+typed costs \$0.00095, \$0.00610 and \$0.40937 by length and solves 13/15, 11/15 and 12/15. Standard RLM costs \$0.09659, \$0.12127 and \$0.78316 and solves 12/15, 8/15 and 7/15. Direct Gemini is cheap at 1k/4k (\$0.00352/\$0.01509) but becomes the most expensive path at 262k (\$1.00599) while solving 4/15. Giving standard r1ms the same typed operator library helps but does not remove the cost-scale pattern: the convenience row costs \$0.13746/\$0.17250/\$1.00449 at 1k/4k/262k and the atomic-only row costs \$0.17118/\$0.20065/\$0.58686, with long-row timeout/error counts of 2 and 4. Together, the two public-contract Oolong ladders cover 8k–131k and a context-window-ID-disjoint 262k stress point under explicitly different overlap properties.

The failure modes separate in a way that matches recent long-context decomposition work. RLM still beats direct/CoT Gemma overall in the $N = 30$ four-method pilot and wins several user-ID cases where whole-context Gemma copies a source row or chooses the wrong entity, but the $N = 120$ run adds nine timeout/scaffold errors and many wrong numeric or wrong-output-type predictions.

Route	Complete	Correct G/S	Median / max s	Active \$
Standard RLM	42/45	33/45 / 33/45	44.6 / 1,289.4	6.746
Text decomposition + same Qwen	45/45	29/45 / 28/45	29.6 / 62.5	1.610
BM25 + same Qwen	42/45	21/45 / 21/45	3.2 / 21.2	0.294

Table 23: Disjoint fixed- $N=45$ BrowseComp-Plus replication. G/S reports Gemma/Sol correct counts; Gemma Wilson intervals are 59.0–84.0, 49.8–76.8 and 32.9–60.9 for RLM, text and BM25. Failures count incorrect. RLM–BM25 is +26.7 points ($b=15, c=3$, paired bootstrap 11.1–44.4, exact $p=0.0075$, two-contrast Holm $p=0.0151$). RLM–text is +8.9 points ($b=9, c=5$, -6.7 –24.4, $p=0.424$) and remains uncertain.

785 Direct Gemini avoids those scaffold errors and is cheaper but still has 66 non-error wrong answers.
786 SLM-labeler-typed has no scaffold errors, yet its 40 misses split evenly between local-label/model
787 noise and aggregation sensitivity after high-accuracy local labeling. The evidence does not show that
788 SLM solves Oolong. Instead, it shows that the dominant failure allocation changes by mechanism
789 and that the current standard-RLM scaffold does not reduce model or aggregation noise enough to
790 justify its extra cost on this frozen public contract. The diagnostic JSON, CSV and PDF artifacts are
791 included in the release under the Oolong $N=120$ noise-decomposition filenames.

792 F BrowseComp-Plus Fixed- N Replication

793 To remove the stopping-rule ambiguity from the positive-regime endpoint, we froze $N=45$ rows
794 from official query shard 1 before output. The manifest uses SHA-256 ordering, contains exactly
795 1,000 unique documents per row and has zero query-ID or query-hash overlap with the earlier $N=80$
796 and $N=59$ manifests. Model revision, official r1ms commit, three route prompts, construction, caps,
797 failure-as-incorrect policy and primary judge are unchanged. A local-driver interruption occurred
798 after 107/135 cells. Before score inspection, a numbered infrastructure amendment preserved those
799 cells byte-for-byte and resumed exactly the 28 missing cells on the surviving pinned pod. The
800 combined compactor verifies 107 plus 28 non-overlapping cells, all 135 attempted cells and both raw
801 hashes; the pod was deleted before scoring.

802 The primary Gemma 4 judge attempted all 129 successful generations; 128 responses parsed and one
803 format failure was conservatively scored incorrect. GPT-5.6 Sol at high reasoning then re-judged
804 all 129 outputs under route and primary-label blinding. Agreement is 128/129 (99.2%, Cohen’s
805 $\kappa=0.983$). Sol preserves RLM and BM25 totals, changes text from 29 to 28 and leaves the two
806 RLM conclusions unchanged. This is independent second-model sensitivity, not human validation.
807 Generation and primary judging cost \$10.120 and \$0.338 at listed RunPod rates. Sol returned
808 \$0.590893 in per-call cost while the concurrent account delta was \$0.496861; both are retained
809 because the credential may be shared. Restricted model I/O stays local. The release contains the
810 public manifest, protocol, amendment, release-safe aggregates, source hashes and an independent
811 135-cell verifier.

812 G BrowseComp-Plus Equal-Cap Comparison and Rollout Sensitivity

813 The strongest BrowseComp-Plus endpoint prospectively freezes 45 further untouched shard-1 queries
814 before any route output and attempts all 180 cells. Standard RLM, iterative lexical search, textual
815 decomposition and BM25 use the same self-hosted Qwen3.6-35B-A3B-FP8 checkpoint and seed
816 20. Each route has at most 20 sequential decisions and 1,200 seconds per row; route-specific access
817 patterns differ by design. The execution required one qualification row, an interrupted immutable
818 ranks 2–19 episode and a hash-bound pre-score ranks 20–45 recovery. Failures remain incorrect in
819 the fixed denominator.

820 On the 39 rows completed by every route, RLM/iterative/text/BM25 score 30/28/31/22. This common-
821 completion analysis is post-treatment and does not replace the all-attempted endpoint. GPT-5.6 Sol
822 independently re-judges all 173 successful outputs under route and primary-label blinding and agrees
823 on 172/173 ($\kappa=0.985$). This is second-model scorer stability, not human validation.

Route	Comp. / corr.	Wilson 95%	Calls	p50/p95 seconds	\$
RLM	41/31	54.3–80.5	476	50.1/628.8	9.873
Iterative	45/33	59.0–84.0	426	14.1/90.2	1.710
Text decomp.	45/37	68.7–90.7	311	22.1/47.6	1.400
BM25	42/25	41.2–69.1	42	3.0/20.0	0.296

Table 24: Prospective equal-cap BrowseComp-Plus endpoint, seed 20 ($N = 45$). Entries under “Comp./corr.” are completed/correct counts. Costs are route-attributed active runtime; the four routes sum to \$13.279. Adding the interrupted-episode billing lower bound and primary judging gives \$15.015.

RLM contrast	Diff. Only	95% interval / exact p	Holm
Iterative search	−4.4 6 / 8	[−20.0, +11.1] / .7905	.7905
Text decomposition	−13.3 5 / 11	[−31.1, +4.4] / .2101	.5387
BM25 + Qwen	+13.3 10 / 4	[−2.2, +28.9] / .1796	.5387

Table 25: Declared seed-20 paired RLM contrasts in percentage points. “Only” gives RLM-only/other-only correct counts. Holm adjustment covers the three comparisons; none rejects and the minimum adjusted value is 0.5387.

824 **Post-hoc seed-21 repeat.** The same 45 rows were rerun with only the model-card seed changed
825 from 20 to 21. This same-row sensitivity was specified after the seed-20 result, is not an independent
826 endpoint and is not pooled with it.

Route	Raw/cap / corr.	Wilson 95%	p50/p95 seconds	\$
RLM	37/36/30	52.1–78.6	64.2/1,219.4	18.170
Iterative	45/45/31	54.3–80.5	7.9/68.3	1.145
Text decomp.	45/45/29	49.8–76.8	26.0/44.8	1.498
BM25	42/42/22	35.0–63.0	3.2/21.0	0.356

Table 26: Post-hoc seed-21 same-row rollout sensitivity ($N = 45$). “Raw/cap/corr.” gives raw completion, cap-valid completion and correct counts. Seven RLM rows exceed the 1,200-second cap; the sole late nonempty output is incorrect, so fail-closed scoring leaves 30/45 unchanged.

827 RLM versus iterative/text/BM25 differs by −2.2, +2.2 and +17.8 points, with Holm-adjusted p
828 values 1.0, 1.0 and 0.346. Route ordering changes but the non-significant RLM contrast set repeats.
829 On the 34 rows cap-valid for every route, RLM/iterative/text/BM25 score 28/23/20/18; this is
830 again a post-treatment sensitivity. RLM active-runtime cost is $15.9\times$ iterative search and $12.1\times$
831 textual decomposition. The four route-attributed totals sum to \$21.170 and generation plus primary
832 judging costs \$22.374 from pod creation through cleanup. Sol agrees on 166/169 successful outputs
833 ($\kappa=0.961$). The seed repeat supports rollout-instability caution, not equivalence or independent
834 replication.

835 **Dominance, transitions and completion across rollouts.** After freezing the equal-cap contract,
836 we recomputed row-level correctness from the retained compact generations and primary Gemma 4
837 judgments without new model calls. Published seed-20 counts 31/33/37/25 and seed-21 counts
838 30/31/29/22 reconstruct exactly. Under accuracy-up/cost-down dominance, seed 20 retains only
839 textual decomposition and BM25 as non-dominated; RLM and iterative search are dominated.
840 Seed 21 flips the adaptive frontier: iterative search and BM25 remain non-dominated while textual
841 decomposition becomes dominated, matching the main-text observation that route ordering is rollout-
842 sensitive. Mean per-row accuracy regret against an oracle that picks any correct route on that row
843 is lowest for text on seed 20 (0.156) but rises to 0.289 on seed 21, whereas RLM regret stays near
844 0.27–0.29 and BM25 remains highest (≈ 0.42 –0.44).

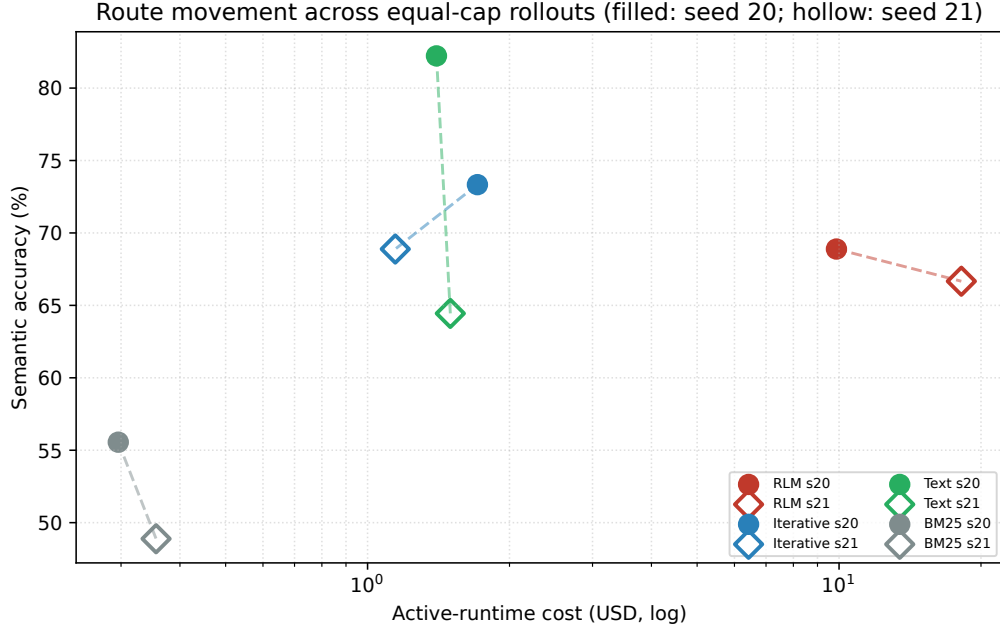


Figure 4: Equal-cap accuracy–cost movement across rollouts (filled markers: prospective seed 20; hollow markers: post-hoc seed 21). Lines connect the same route. Costs are route-attributed active runtime on a log axis. The plot does not pool seeds into one estimator.

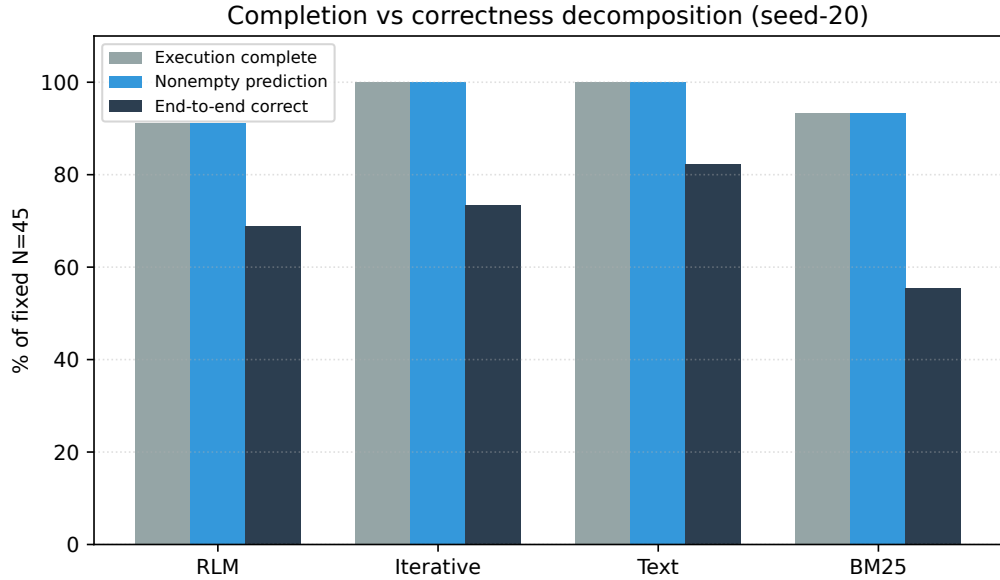


Figure 5: Completion-versus-correctness decomposition on the prospective seed-20 endpoint: execution completion, nonempty prediction and end-to-end semantic correctness under the fixed $N=45$ denominator. RLM’s gap is concentrated in completion and end-to-end correctness rather than empty predictions alone.

Route	s20	s21	TT	FT	TF	FF
RLM	31	30	23	7	8	7
Iterative	33	31	27	4	6	8
Text decomp.	37	29	28	1	9	7
BM25	25	22	22	0	3	20

Table 27: Seed-20→seed-21 correctness transitions on the same 45 rows (TT=stable correct, FT=gained, TF=lost, FF=stable incorrect). Text loses nine previously correct rows and gains one; RLM is comparatively stable in total correct count but still cost-dominated. This is rollout stability, not an independent task sample.

845 These panels are descriptive re-analyses of the frozen equal-cap evidence. They do not replace the
846 Holm-controlled paired tests, do not add a prospective dense full-pool route and do not authorize
847 pooling seed 20 with seed 21.

848 G.1 Completion, Deadlines and Paired Rollout Sensitivity

849 These analyses reuse the 45 equal-cap queries and their two stored rollouts. They add no independent
850 cases or model calls. All-case correctness keeps failed and invalid outputs in the denominator. Let
851 S be the queries for which every route returns a valid answer within 1,200 seconds. The selected
852 comparison $|S|^{-1} \sum_{i \in S} (Y_{ir} - Y_{is})$ changes both the numerator and denominator. It is descriptive
853 and cannot estimate what would have happened had an incomplete execution succeeded.

Rollout	RLM minus	All 45	Common completions	Change
Seed 20 ($ S = 39$)	Iterative	−4.4	+5.1	+9.6
	Text	−13.3	−2.6	+10.8
	BM25	+13.3	+20.5	+7.2
Seed 21 ($ S = 34$)	Iterative	−2.2	+14.7	+16.9
	Text	+2.2	+23.5	+21.3
	BM25	+17.8	+29.4	+11.6

Table 28: Accuracy differences in percentage points. The last column is the change induced by restricting to common cap-valid completions, not a causal estimate of failure removal.

854 The all-case difference decomposes without changing its denominator: $\hat{\Delta} = 45^{-1} \sum_{i \in S} (Y_{ir} - Y_{is}) +$
855 $45^{-1} \sum_{i \notin S} (Y_{ir} - Y_{is})$. For RLM versus iterative search in seed 21, the common-completion rows
856 contribute +5/45, while excluded rows contribute −6/45, leaving −1/45 overall. This explains the
857 sign reversal without imputing missing answers or attributing a causal effect to attrition.

858 A paired analysis of the change in route differences resamples query IDs jointly across both rollouts
859 and both routes (50,000 bootstrap samples). The seed-21 minus seed-20 changes for RLM versus
860 iterative, text and BM25 are respectively +2.2 points [−22.2, 26.7], +15.6 [−8.9, 40.0] and +4.4
861 [−13.3, 22.2]. These post-hoc 95% intervals describe case uncertainty conditional on two observed
862 rollouts. They neither establish a seed-population interaction nor treat two measurements of one
863 query as independent data.

864 At 60 seconds, RLM/iterative/text/BM25 have returned 22/31/37/25 correct answers in seed 20 and
865 19/30/29/22 in seed 21. At 1,200 seconds the counts are 31/33/37/25 and 30/31/29/22. Longer
866 allowed waiting reveals more correct RLM answers in these stored executions, while the relative
867 ordering remains task- and rollout-specific. No confidence claim about fresh deployments is inferred
868 from the curves alone.

869 **Question-parser audit.** The unchanged question parser was replayed using only the question
870 string; the stored benchmark task and answer-type fields were used solely for checking agreement
871 afterward. Both fields match on 150/150 SU-150 cases (73 context windows), 45/45 CD-45 cases (32
872 windows) and 45/45 VC-45 cases (20 windows), and every replay matches its original stored parser
873 output. This repairs incomplete reporting of a check already represented in the stored outputs. It does
874 not add new data or establish robustness to arbitrary paraphrases. In particular, selecting the correct

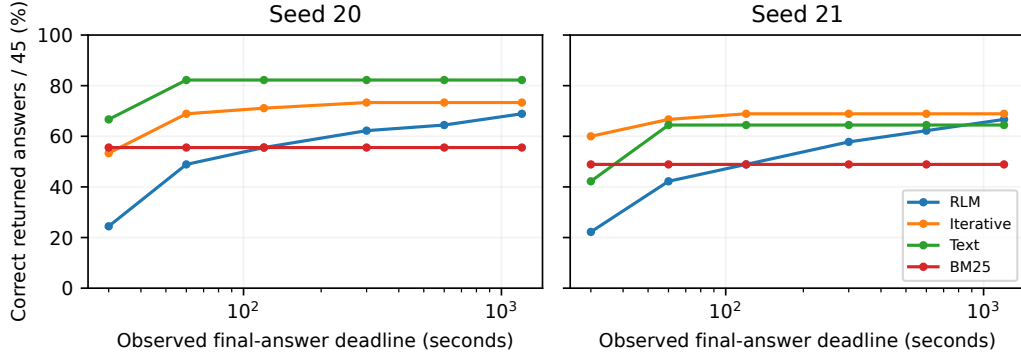


Figure 6: Correct final answers returned by a given deadline on the retained runs, with all 45 queries in every denominator. The agent policy was held at its original cap: these are observed completion curves, not reruns with shorter instructed budgets. They exclude any unreturned intermediate answer and are not dollar-budget curves.

task and answer type does not guarantee correct local labels, subset extraction, tie handling or the final answer.

The replay script writes per-case parser decisions, paired contrasts, exact McNemar tests with within-rollout Holm adjustment, selection decompositions, deadline counts and source hashes. Missing successful-output judgments, duplicate route/query cells and mismatched query IDs fail validation. All figures derive from the same analysis artifact.

H Repeatability and Cost Accounting

Controlled families. T-STRUCT and the definitive T-MIXED $N = 30$ comparison are single full-family passes. The T-MIXED old-finalizer run is a same-model $N = 20$ mechanism probe; its exact-example manifest is reconstructed from stored model-I/O traces. The current-contract extension instead uses a separately frozen stable-generator manifest and SHA-256-derived offset, avoiding the historical generator’s salted Python hash(domain) seed.

Public anchors. BrowseComp-Plus contributes four major estimates: the frozen $N = 80$ two-episode completion, the zero-overlap score-blind attempted- $N = 49$ replication stopped by its stability rule, the disjoint fixed- $N = 45$ three-route replication with all 135 cells attempted before scoring and the prospective fixed- $N = 45$ four-route equal-cap endpoint with all 180 cells attempted before scoring. The seed-21 rerun is a post-hoc same-row rollout sensitivity, not a fifth independent estimate. The earlier $N = 37$ prefix records the first study’s interruption and is not a confirmatory sample. T-NLP and T-SQ remain compatibility/proxy diagnostics; LB2 $> 500k$ and HotpotQA are bounded public diagnostics; and the public λ -RLM andprehend SRLM-style rows are pressure tests rather than official reproductions.

Cost accounting. Cost ratios compare routes within task families rather than through a global method denominator. RLM totals cover logged successful completions. The attempted-cost audit records their lower bound and the zero-usage failed rows that prevent exact reconstruction of partial failed-call billing.

Models and execution. Hosted models are identified by dated provider slug in the experiment records. The self-hosted BrowseComp endpoint uses Qwen3.6-35B-A3B-FP8 at a frozen weight revision and serving-image digest, a 65,536-token context limit, model-card decoding, disabled thinking and seed 20 (or 21 only for the labeled repeat). The official λ row pins repository commit 3874d393483d, weight revision b968826d9c46 and vllm/vllm-openai:v0.21.0 on one 80GB H100. The official trained-policy row uses an 80GB-class RunPod GPU and the canonical rlm harness at commit 72d6940142ddfbb84e6be573dc999a37e633e671.

907 **Compute accounting.** API experiments log returned tokens and list-price cost. Self-hosted
908 BrowseComp costs use route-attributed active runtime at the recorded pod rate; generation-plus-judge
909 totals are reported separately. Failed hosted calls sometimes return no usage, so attempted API cost is
910 a lower bound. The complete official λ CD-15 row required 6,562 seconds on one 80GB H100 and
911 cost \$5.471 from creation through cleanup. The prospective seed-20 equal-cap endpoint attributes
912 \$13.279 across route runtime and \$15.015 including the interrupted-episode lower bound and primary
913 judging. The seed-21 repeat attributes \$21.170 across routes and \$22.374 from pod creation through
914 primary judging and cleanup.

915 **Data and release boundary.** The study uses public or synthetic benchmark interfaces and does
916 not redistribute source-document-scale BrowseComp-Plus, Oolong, LongBench-v2 or HotpotQA
917 corpora. Access remains governed by upstream terms. The project retains frozen manifests, prompts,
918 aggregate outputs, model-I/O traces, cost ledgers and deterministic claim verifiers. A public artifact
919 is planned after anonymous review; the current single-PDF workshop submission makes no claim
920 that code or data are already openly downloadable.

921 I Reproducibility Commands

922 **Integrity and build commands.** From the extracted release root, run `python3`
923 `verify_release.py`; success is the final ANONYMOUS RELEASE VERIFIER: PASS line. This
924 deterministic command requires no network, model, API or GPU access.

925 NeurIPS Paper Checklist

926 1. Claims

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

929 Answer: [Yes].

930 Justification: The abstract and Introduction state the evaluated scaffold, matched-route
931 contribution, central uncertainty result and task-conditional scope; Sections 2–6 distinguish
932 evidence from non-claims.

933 2. Limitations

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

935 Answer: [Yes].

936 Justification: Section 6 discusses external validity, measurement, cost scope, evaluator
937 authority, split boundaries and the absence of human adjudication.

938 3. Theory assumptions and proofs

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

941 Answer: [N/A].

942 Justification: The paper is an empirical evaluation audit and makes no theorem or formal-
943 proof contribution.

944 4. Experimental result reproducibility

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

948 Answer: [Yes].

949 Justification: Section 2 and Appendix H specify inputs, routes, scoring, failures, uncertainty,
950 versions, caps, compute and cost boundaries for the claim-bearing results.

951 5. Open access to data and code

952 Question: Does the paper provide open access to the data and code, with sufficient instruc-
953 tions to faithfully reproduce the main experimental results, as described in supplemental
954 material?

955 Answer: [No].

956 Justification: The single-PDF workshop submission does not provide a public code/data

957 URL. Appendix H documents the reproduction contract and asset boundaries; an anonymized

958 artifact is planned after review.

959 **6. Experimental setting/details**

960 Question: Does the paper specify all the training and test details (e.g., data splits, hyperpa-

961 rameters, how they were chosen, type of optimizer) necessary to understand the results?

962 Answer: [Yes].

963 Justification: Section 2 specifies the frozen splits, route roles, decoding policies, seeds,

964 estimands, scorers and stopping rules. No model is trained by this work; official trained

965 checkpoints are inference-only diagnostics.

966 **7. Experiment statistical significance**

967 Question: Does the paper report error bars suitably and correctly defined or other appropriate

968 information about the statistical significance of the experiments?

969 Answer: [Yes].

970 Justification: Section 2 defines paired McNemar tests, fixed-seed row and cluster bootstraps

971 and Holm correction; Section 3 and Appendix G report intervals, discordances and adjusted

972 values for central contrasts.

973 **8. Experiments compute resources**

974 Question: For each experiment, does the paper provide sufficient information on the com-

975 puter resources (type of compute workers, memory, time of execution) needed to reproduce

976 the experiments?

977 Answer: [Yes].

978 Justification: Appendix H identifies hosted APIs, self-hosted model revisions, 80GB GPU

979 deployments, container/harness versions, elapsed time and route-level costs for the principal

980 self-hosted runs.

981 **9. Code of ethics**

982 Question: Does the research conducted in the paper conform, in every respect, with the

983 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

984 Answer: [Yes].

985 Justification: The study uses public or synthetic benchmark interfaces, does not involve

986 human subjects and does not redistribute restricted source-document-scale corpora; Section 6

987 and Appendix H state these boundaries.

988 **10. Broader impacts**

989 Question: Does the paper discuss both potential positive societal impacts and negative

990 societal impacts of the work performed?

991 Answer: [Yes].

992 Justification: Section 6 discusses lower inference cost and accessibility as potential benefits

993 and overgeneralized routing decisions as the main potential harm.

994 **11. Safeguards**

995 Question: Does the paper describe safeguards that have been put in place for responsible

996 release of data or models that have a high risk for misuse (e.g., pre-trained language models,

997 image generators, or scraped datasets)?

998 Answer: [N/A].

999 Justification: The work releases no model or high-risk scraped dataset and does not redis-

1000 tribute the source-document-scale benchmark corpora.

1001 **12. Licenses for existing assets**

1002 Question: Are the creators or original owners of assets (e.g., code, data, models), used in

1003 the paper, properly credited and are the license and terms of use explicitly mentioned and

1004 properly respected?

1005 Answer: **[No]**.

1006 Justification: The originating papers and official assets are credited, no third-party corpus is

1007 redistributed and Appendix H states the access boundary. The PDF does not reproduce a

1008 complete license inventory for every hosted model and benchmark interface.

1009 **13. New assets**

1010 Question: Are new assets introduced in the paper well documented and is the documentation

1011 provided alongside the assets?

1012 Answer: **[N/A]**.

1013 Justification: No new model or dataset asset accompanies this single-PDF workshop submis-

1014 sion; a separately documented audit artifact is planned after review.

1015 **14. Crowdsourcing and research with human subjects**

1016 Question: For crowdsourcing experiments and research with human subjects, does the paper

1017 include the full text of instructions given to participants and screenshots, if applicable, as

1018 well as details about compensation (if any)?

1019 Answer: **[N/A]**.

1020 Justification: The study does not use crowdsourcing or human-subject experiments.

1021 **15. Institutional review board (IRB) approvals or equivalent for research with human**

1022 **subjects**

1023 Question: Does the paper describe potential risks incurred by study participants, whether

1024 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)

1025 approvals (or an equivalent approval/review based on the requirements of your country or

1026 institution) were obtained?

1027 Answer: **[N/A]**.

1028 Justification: The study does not involve human subjects, so IRB review is not applicable.

1029 **16. Declaration of LLM usage**

1030 Question: Does the paper describe the usage of LLMs if it is an important, original, or

1031 non-standard component of the core methods in this research? Note that if the LLM is used

1032 only for writing, editing, or formatting purposes and does *not* impact the core methodology,

1033 scientific rigor, or originality of the research, declaration is not required.

1034 Answer: **[Yes]**.

1035 Justification: Sections 1–3 and Appendix H identify the LLM controllers, workers, direct

1036 routes and judges central to the evaluation; Section 6 separately discloses AI assistance.