

[← Go to NeurIPS 2026 Workshop LCFM homepage](#)

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

Odysseus Kaloidas¹

¹ University of California, Berkeley (berkeley.edu)

29 Aug 2026 (modified: 15 Sept 2026) NeurIPS 2026 Workshop LCFM Submission

LCFM, Reviewers, Authors Revisions CC BY 4.0

Edit ▾

Keywords: Recursive Language Models, long-context reasoning, inference-time scaling, cost-aware evaluation, task-conditional routing, agent evaluation, retrieval, robustness

TL;DR: We show that RLM value is task-conditional: operation-matched routes usually outperform it on known operations, while open-search advantages do not remain reliable under equal budgets and repeated rollouts.

Abstract:

Recursive Language Models (RLMs) handle long inputs through controller-generated programs and additional model calls. Published RLM evaluations typically compare these scaffolds with generic long-context or coding baselines. Deployment requires a different comparison: whether adaptivity improves on routes suited to the task under a common information, scoring and budget contract. We evaluate the released standard RLM against deterministic operators, retrieval, semantic workers and direct inference in an operation-matched audit. On controlled operations and known-family Oolong aggregation, matched routes generally match or outperform standard RLM at lower cost. A source-transfer replication preserves this direction although its clustered exact-rate difference includes zero. Open multi-document search is less stable. An initial untouched-shard replication favors RLM over BM25 but neither a prospectively locked equal-cap comparison nor a post-hoc same-row repeat yields an RLM contrast that survives multiplicity correction. Route ordering changes across these rollouts while RLM remains less reliable and 5.8–15.9× costlier than the stronger alternatives. Mechanism probes localize errors to operation specificity, output-schema fit, evidence selection and scaffold reliability. These results support reporting recursive inference as a task-conditional accuracy–cost frontier rather than as a single scaffold ranking.

Email Sharing: We authorize the sharing of all author emails with Program Chairs.

Data Release: We authorize the release of our submission and author names to the public in the event of acceptance.

Submission Number: 9

Filter by reply type... ▾

Filter by author... ▾

Search keywords...

Sort: Newest First ▾

-

=

≡

Everyone Program Chairs Submission9 Reviewers

Submission9 Authors ✕

3 / 3 replies shown

-

=

≡

Paper Decision

Decision by Program Chairs  29 Sept 2026, 08:31 (modified: 30 Sept 2026, 07:03)

 Program Chairs, Reviewers, Authors  Revisions

Decision: Accept

Comment:

This paper studies recursive language models in depth and provides valuable insights. Accept.

-

=

≡

A rigorous, cost-aware, but specialization-dependent RLM evaluation.

Official Review by Reviewer eBjw  29 Sept 2026, 04:50 (modified: 30 Sept 2026, 07:02)

 Program Chairs, Reviewers, Authors  Revisions

Review:

The paper evaluates the released, standard RLM against operation-matched alternatives under same evaluation settings. On controlled families and standard-unlabeled Oolong, the matched routes match or outperform RLM at lower cost. Wrong-family and same-operator-library controls show that the advantage comes from knowing the operation family rather than from tool access alone. On BrowseComp-Plus, RLM significantly outperformed BM25 in an early experiment, but its numerical advantage was not statistically significant in either the fixed equal-cap comparison or different-seed repeat, and RLM remains costlier. The authors conclude that RLM evaluation should report a task-conditional accuracy–cost frontier rather than a single scaffold ranking.

Pros:

- The mechanism controls are well chosen to distinguish the effects of correct operation knowledge, access to typed tools, and answer finalization. The wrong-family route collapses from 36/45 to 2/45, while giving RLM the same operator library yields 31/45 or 29/45 at greater cost. A separate finalizer ablation shows that some failures arise from the answer contract rather than upstream reasoning. Together, these controls provide useful evidence about where the specialized route’s advantage comes from.
- The main contribution appears to be the evaluation protocol, which combines same-row comparisons of complete deployed routes under explicit information, scoring, and budget contracts with paired uncertainty, mechanism controls, and rollout-stability analysis.
- The paper is a timely to correct single-number scaffold rankings and provide solid cost evidence. The documented rollout variance is informative for anyone benchmarking RLM-style systems.

Cons:

- The information contract is unbalanced by design. The specialized routes receive task-family operators and schemas, whereas the primary standard-RLM route does not. The paper reports implementation surface and hypothetical amortization thresholds, but not actual engineering time or maintenance cost. It tests question-based schema inference within Oolong, yet does not evaluate the proposed two-tier policy with an end-to-end router over mixed task families.
- The main text is densely packed with acronyms, evidence tiers, dataset variants, and qualifications, making it difficult to distinguish the key claims from diagnostic and sensitivity analyses. Important methodological details are deferred to the long technical appendix.

Rating: 7: Good paper, accept

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct



A worthwhile audit of when recursive inference earns its overhead



Official Review by Reviewer 6QEC  29 Sept 2026, 04:49 (modified: 30 Sept 2026, 07:03)

 Program Chairs, Reviewers, Authors  Revisions

Review:

The question this paper asks is more useful than another scaffold ranking: does an RLM justify its cost when a simpler system already knows how to perform the required operation? The evidence is strongest on known-family aggregation, where specialized pipelines are both cheaper and more accurate. The open-search results are less decisive, but documenting that instability is itself valuable.

Strengths

1. The baselines are taken seriously. Fixed operators, semantic workers, and iterative retrieval offer credible alternatives to recursive inference. The same-operator and wrong-family controls help explain where their advantage comes from.
2. Failures count toward the result. Completion rates, latency, and cost are reported alongside correctness, avoiding an overly favorable picture based only on successful trajectories.
3. The paper does not hide reversals. It distinguishes prospective experiments from post-hoc repeats and reports when an apparent advantage fails to replicate. Clustered uncertainty estimates for Oolong and independent judge checks further strengthen the audit.

Weaknesses

1. The deployment recommendation assumes that the task family is known or can be identified reliably. The wrong-family collapse shows how consequential that assumption is, but no evaluated router accounts for family-selection errors. The results establish the value of specialization more directly than a deployable routing policy.
2. The open-search evidence remains underpowered: the strongest endpoint has 45 queries and one additional seed. Moreover, the move to equal caps also changes the query set, so the changed ordering cannot be attributed to budget matching alone. A same-query budget sweep would be more informative.

My suggestion is a weak accept. The contribution is an empirical audit rather than a new inference method, and it succeeds in showing why specialized baselines and execution failures belong in RLM evaluations.

Rating: 6: Marginally above acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

About OpenReview

FAQ

Sponsors

All Venues

Contact

Hosting a Venue

Terms of Use / Privacy Policy

Donate