ROUTE-Bench: Evaluating Selective Routing Under Composition

J. W. Miller
j.w.miller@curv.institute
Founder & Research Director
CURV Institute

Abstract

We introduce ROUTE-Bench, a synthetic benchmark for evaluating how language models integrate evidence from multiple sources under compositional task demands. The benchmark measures routing quality across retrieval-augmented generation (RAG), structured memory access, and tool use in long-context settings. We evaluate three control strategies—baseline inference, hard-constrained prompting, and Harmonized Hyper-Connections—and find that the harmonized variant doubles evidence-grounded accuracy (0.040 vs. 0.020), albeit from a low baseline typical of compositional evidence integration tasks, while exhibiting 35% lower position sensitivity compared to baseline. Hard constraints improve evidence recall but reduce raw accuracy. Tool utilization remains a limitation across all variants.

1 Introduction

Language models increasingly operate in compositional settings where they must integrate information from multiple sources, including retrieved documents, structured memory stores, and computational tools. Prior work has shown that naïve approaches to multi-source reasoning suffer from position-dependent accuracy variation, poor evidence grounding, and suboptimal tool utilization.

This paper introduces ROUTE-Bench, a fully synthetic and deterministically verifiable evaluation suite for studying selective routing under composition. Unlike existing benchmarks that isolate individual capabilities, ROUTE-Bench requires models to simultaneously retrieve relevant evidence from large document collections, query structured memory, invoke tools when computation is required, and integrate evidence across variable context positions. A synthetic benchmark allows precise control over evidence placement, difficulty, and verification, which is necessary to isolate routing behavior under composition.

This paper is a direct evaluation follow-on to Harmonized Hyper-Connections [2], which introduced a mechanism for stabilizing residual transport by regulating applied composite gain. Here, we extend that work by focusing on evaluation rather than mechanism design, examining how harmonized regulation affects evidence grounding, routing behavior, and long-context robustness in compositional language model settings. This work is conducted within the broader research program of the CURV Institute, which studies how failures of representational control give rise to instability in complex systems, and how such failures can be diagnosed and corrected through empirical validation.

2 ROUTE-Bench Design

2.1 Task Structure

Each ROUTE-Bench example contains a document heap (50–200 synthetic documents with one evidence-bearing document), a structured memory table (64–512 key–value pairs with one relevant entry), a set of available tools (2–4 simple tools such as a calculator or string processor), and a query that requires document lookup, memory access, and optional tool use. Evidence is placed at varying positions within the context to induce long-context retrieval challenges.

2.2 Skill Types

The benchmark covers eight skill categories requiring distinct reasoning patterns: arithmetic operations, string transformations, logical entailment, table joins, date arithmetic, pattern matching, unit conversion, and comparison.

2.3 Metrics

We report AnswerAcc (exact-match answer accuracy), GroundedAcc (accuracy conditioned on correct evidence citation, requiring both correct answer and correct evidence), and ToolTraceAcc (tool-call correctness). Routing behavior is evaluated using EvidenceRecall, ToolUnderuseRate, and ToolOveruseRate. Long-context robustness is measured via Acc@Position and PositionSensitivity, defined as the standard deviation of accuracy across evidence position quintiles (lower indicates greater stability).

3 Model Variants

We evaluate three control strategies using `Mistral-7B-Instruct-v0.2`.

Baseline. Standard inference without additional control.

Hard-Constrained (mHC-style). Strict prompting constraints requiring exactly one document citation, one memory citation, and mandatory tool invocation for computational queries. This mirrors hard constraint strategies that favor conservative mixing [1].

Harmonized Hyper-Connections (applied-gain prompting). A prompt-based instantiation of the Harmonized Hyper-Connections mechanism introduced in prior work [2], where guidance is used to regulate the effective influence of routed evidence under composition. In this benchmark, harmonization is implemented via applied-gain prompting rather than architectural modification. *While implemented via prompting here, the goal is to evaluate the behavioral effects of harmonization rather than propose a new prompting strategy.*

4 Results

4.1 Overall Performance

Metric	Baseline	Hard-Constrained	Harmonized
AnswerAcc	**0.255**	0.150	0.225
GroundedAcc	0.020	0.025	**0.040**
ToolTraceAcc	0.410	**0.435**	0.420
EvidenceRecall	0.333	**0.403**	0.365
ToolUnderuseRate	0.686	**0.347**	0.839
ToolOveruseRate	0.000	0.000	0.000

Harmonized Hyper-Connections improve evidence-grounded accuracy by a factor of two relative to baseline. Hard constraints increase evidence recall but reduce raw answer accuracy, consistent with conservative routing that suppresses decisiveness.

4.2 Position Sensitivity

Position sensitivity is measured as the standard deviation of AnswerAcc across evidence position quintiles.

Variant	Std. Dev.	Range
Baseline	0.060	0.164
Hard-Constrained	0.056	0.143
Harmonized	**0.039**	**0.111**

The harmonized variant reduces position sensitivity by 35% relative to baseline, indicating more stable performance across context positions.

4.3 Early-Position Grounding

At early evidence positions (0–20% of context length), Harmonized Hyper-Connections improve GroundedAcc by 3× relative to baseline (0.075 vs. 0.025), where evidence integration is most difficult.

4.4 Tool Usage

Hard-constrained prompting significantly reduces tool underuse through explicit format enforcement. Harmonized Hyper-Connections do not by themselves encourage tool invocation and increase tool underuse relative to baseline. This indicates that stabilizing routing under composition and incentivizing tool use are complementary but distinct problems. *This separation allows tool-use incentives to be studied independently of routing stability.*

5 Limitations and Null Results

ROUTE-Bench is synthetic, and real-world tasks may exhibit different routing patterns. Results are reported for a single model family. Harmonization is implemented via prompting rather than architectural modification. Tool utilization remains a limitation, and arithmetic and comparison

skills remain challenging across all variants. GroundedAcc remains low in absolute terms, highlighting the difficulty of compositional evidence integration.

6 Future Work

An important direction for future work is the application of harmonized routing to agentic systems that interleave internal reasoning with external tool use over multiple steps. Recent agent architectures rely on repeated routing decisions—such as when to trust tool outputs, how to reuse intermediate results, and how to balance earlier commitments against later evidence—which are composed across long reasoning traces. Our results suggest that many failures in such systems may stem not from incorrect decisions, but from unstable composition of their influence over time. Applying harmonized regulation of *applied residual transport* in these settings could help correct tool outputs and intermediate conclusions remain influential without dominating subsequent reasoning. We emphasize that this would not replace explicit planning or tool-selection policies, but rather provide a stable substrate on which such policies operate. Evaluating this hypothesis in multi-step agent benchmarks and coding-oriented agent loops is a promising avenue for extending the insights of ROUTE-Bench to real-world agentic workflows.

7 Conclusion

ROUTE-Bench provides a controlled framework for evaluating selective routing under composition. Harmonized Hyper-Connections improve evidence grounding and reduce sensitivity to evidence position while preserving routing capacity. Hard constraints trade raw accuracy for recall. Together with prior work on Manifold-Constrained and Harmonized Hyper-Connections, these results suggest that stabilizing applied residual transport is a prerequisite for reliable routing under composition, and that ROUTE-Bench offers a controlled setting for evaluating such effects.

Acknowledgment

This work was conducted at the CURV Institute as part of an ongoing research program on the diagnosis and regulation of representational instability in complex systems. The author acknowledges prior work by Xie et al. [1] for establishing composite residual gain as the governing instability mechanism in widened residual architectures, which directly motivated the evaluation focus of ROUTE-Bench.

Reproducibility

All code, data, and results are available at: `https://github.com/curv-institute/route-bench`

References

[1] Z. Xie et al. Manifold-Constrained Hyper-Connections. arXiv preprint arXiv:2512.24880, 2025.

[2] J. W. Miller. Harmonized Hyper-Connections: Stabilizing Residual Transport Through Feedback on Geometry. arXiv preprint, 2026.

[3] DeepSeek-AI. Tool-augmented and agentic reasoning with large language models. arXiv:2512.24601, 2025.