

Does Every AI Control-Plane Decision Need an LLM?

Evaluating Specialized Decision Models for Policy-Aware AI Routing

Debabrata Pruseth
Independent Researcher, Singapore

Abstract

Large language models (LLMs) are increasingly used not only to generate content but also to make control-plane decisions such as selecting models, tools, agents, and workflows. Existing routing research has demonstrated that heterogeneous model portfolios can improve cost-quality trade-offs, but a preceding systems question remains: what computational mechanism should perform the routing decision itself? We compare deterministic rules, weighted heuristics, supervised machine learning, a general-purpose LLM, and a specialized fast decision model in a synthetic routing environment designed to represent regulated-banking constraints. Deterministic governance first constructs a policy-eligible candidate set; routing intelligence then selects only among permitted candidates. The primary experiment contains 400 requests, of which 304 require live model-selection decisions and 96 are deterministically resolved as `NO_ROUTE`. Preferred Synthetic Reference-Route Agreement (SRRA) is 75.5%/71.5% for the general-purpose LLM and 74.5%/70.0% for the specialized decision model on IID/template-held-out benchmarks. Exact paired McNemar tests with Holm correction identify no statistically detectable agreement difference; this does not establish equivalence. Across actual live routing decisions, mean latency is 1423.75/1406.17 ms for the LLM execution path and 543.80/562.02 ms for the specialized-decision execution path, approximately 2.50–2.62 times lower for the latter. Total observed primary routing API cost is USD 0.223006 versus USD 0.029983, approximately 7.44 times lower under the evaluated providers and pricing. A separate 500-decision repeated-execution study shows greater observed selection consistency for the specialized decision path, descriptively. The results do not establish a universal winner; they indicate that general-purpose generative inference need not be the default mechanism for bounded, already-structured AI control-plane decisions.

1 Introduction

The proliferation of LLMs has transformed model selection from a static deployment choice into a runtime systems problem. Prior work has shown that routing and cascading can reduce serving cost while preserving quality [1, 2, 5, 6]. RouterBench formalizes multi-LLM routing evaluation [4], while LLMRouterBench re-evaluates routing at substantially larger scale and reports that several sophisticated methods do not reliably outperform a simple baseline under unified evaluation [7]. Where prior work primarily optimizes *which downstream model should execute a request*, we study the computational mechanism used to make the routing decision itself.

After normalization and policy filtering, many control-plane decisions are structured and bounded. They can be represented as

$$x, E(x), O \rightarrow a,$$

where x is structured request state, $E(x)$ the policy-eligible action set, O routing objectives, and a a discrete action. We compare deterministic logic, weighted optimization, supervised learning,

specialized decision inference, and general-purpose generative inference under a common policy-constrained routing architecture.

The primary research question is:

RQ1. How does a specialized fast decision model compare with a general-purpose LLM in Synthetic Reference-Route Agreement for a bounded, structured, policy-aware routing task?

The operational hypotheses are directional and were evaluated only within the frozen provider configuration:

H1a – Latency. The specialized-decision execution path will exhibit lower mean live routing latency than the general-purpose LLM execution path.

H1b – Cost. The specialized-decision execution path will exhibit lower external routing API cost than the general-purpose LLM execution path.

Agreement is not formulated as an equivalence or non-inferiority hypothesis because no equivalence margin was pre-specified. We therefore report paired differences, exact McNemar tests, and confidence intervals without interpreting non-significance as equivalence. Throughout the paper, *specialized decision model* denotes the evaluated Jev 1.13 decision model. TypeSafe describes Jev as its first “System One” model; we use that vendor terminology only when discussing the broader architectural interpretation and do not claim cognitive equivalence with human System-1 cognition.

The contribution is therefore not a new downstream model-routing algorithm. It is a controlled systems comparison of the *computational class used to make a bounded routing decision* under a shared structured input, a shared policy-eligible candidate set, and deterministic governance boundaries. The study further measures the operational overhead of that control-plane intelligence and compares it with simpler mechanisms that require no external inference.

2 Related Work

FrugalGPT studies adaptive selection and cascades for reducing LLM serving cost while preserving task performance [1]. Hybrid LLM predicts whether queries should be sent to smaller or larger models and exposes an inference-time quality–cost trade-off [2]. AutoMix similarly uses confidence-aware escalation between language models to improve cost–performance trade-offs [3]. RouteLLM learns routers from preference data and reports strong cost reductions without sacrificing benchmark response quality [5]. RouterBench provides a standardized benchmark with large collections of model outcomes [4]. More recent theoretical work unifies routing and cascading and identifies estimator quality as a central determinant of model-selection performance [6]. LLMRouterBench evaluates more than 400,000 instances across 21 datasets and 33 models and reports that many routing methods exhibit similar performance under unified evaluation [7].

Related learning-to-defer work studies when a predictive system should hand a decision to an external expert rather than act itself [8]; this provides adjacent conceptual context for selective delegation, although our action space consists of AI model endpoints rather than human experts. These studies primarily optimize downstream allocation, escalation, or quality–cost frontiers. Our focal variable is different: the *computational class of the control-plane mechanism that makes the routing decision*. In addition, we separate policy eligibility from optimization. This is consistent with the broader NIST AI RMF emphasis on governance as a cross-cutting risk-management function [9]; the deterministic eligibility architecture studied here is our engineering design rather than a NIST prescription.

Recent reasoning work also uses System-1/System-2 language to distinguish fast responses from more deliberative computation [10]. We cite this only as terminology precedent; our study concerns the architectural mechanism used for a control-plane decision, not adaptive reasoning inside an

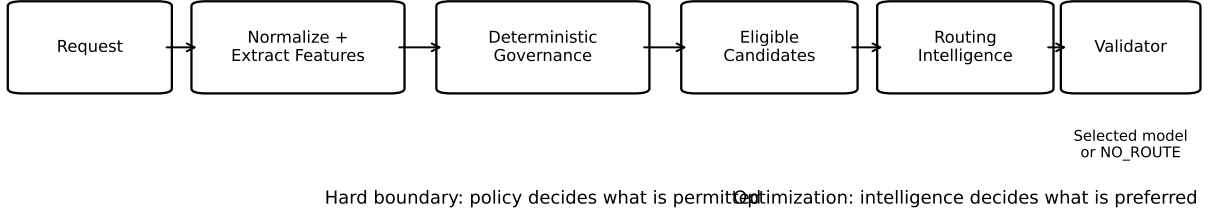


Figure 1: Policy-aware routing architecture. Deterministic governance establishes the eligible candidate set before optimization.

LLM.

Jev 1.13 is documented by OpenRouter as a structured decision model from TypeSafe. Its Decisions interface accepts application state and typed decision questions and returns structured choices or scores rather than free-form generated prose; routing and classification are listed as intended decision-point use cases [11]. TypeSafe calls this family “System One.” Public documentation describes the service interface and pricing but does not disclose sufficient internal architecture detail to attribute our observed operational differences to a particular model design. Accordingly, this paper treats Jev as a specialized decision service and attributes latency and cost findings only to the measured Jev-via-OpenRouter execution path.

3 System Architecture

For model catalog M , deterministic policy function $P(x, m)$ constructs

$$E(x) = \{m \in M : P(x, m) = 1\}.$$

If $E(x) = \emptyset$, the system returns `NO_ROUTE`. Otherwise routing strategy R selects $m^* = R(x, E(x))$, and validator V either accepts the eligible selection or rejects it:

$$M \xrightarrow{P(x)} E(x) \xrightarrow{R(x)} m^* \xrightarrow{V} \{m^*, \text{reject}\}.$$

We refer to this as a *constraint-before-optimization* architecture: deterministic governance establishes what is permitted before any probabilistic or learned mechanism determines what is preferred.

4 Decision-Mechanism Design Space

We compare Rules, a Weighted heuristic, a supervised Random Forest, a specialized decision model, and a general-purpose LLM. Figure 2 presents an illustrative two-dimensional design space rather than an ordinal intelligence or performance ranking. The horizontal axis reflects dependence on task-specific training or configuration; the vertical axis reflects inference generality and adaptability. These positions are conceptual and are not measured scores.

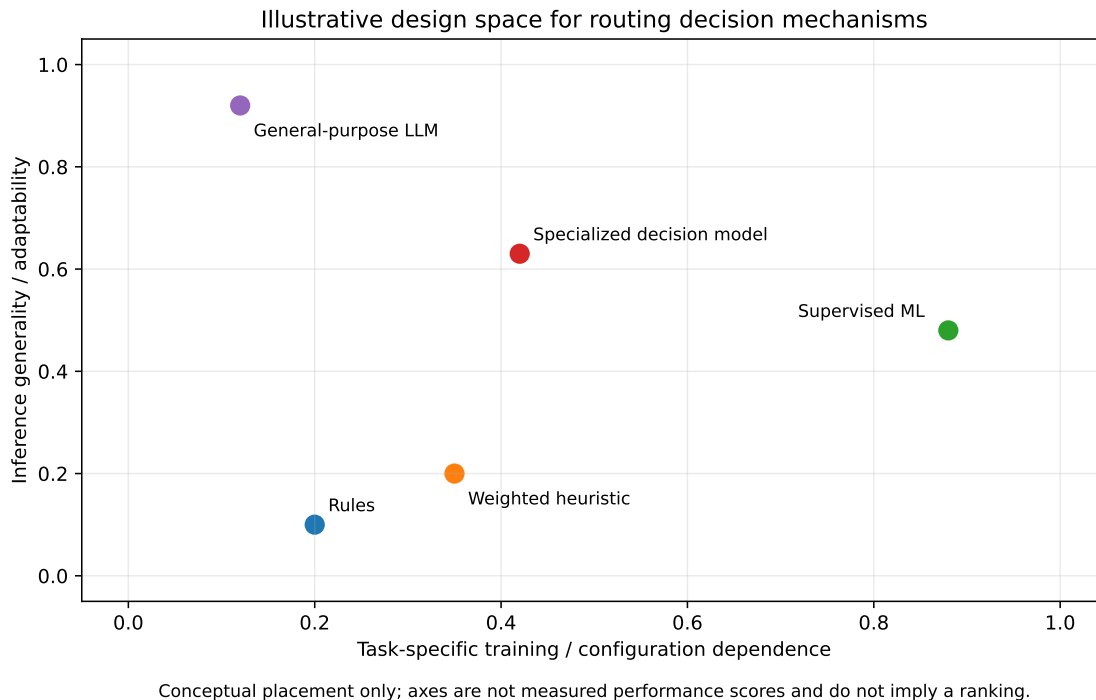


Figure 2: Illustrative design space for the evaluated routing mechanisms. Placement is conceptual and does not imply a universal ranking.

5 Experimental Methodology

The final primary sample contains 200 IID synthetic requests and 200 template-held-out requests. Respectively, 149 and 155 require routing; 51 and 45 are deterministically resolved as `NO_ROUTE`. Both live routers therefore perform 304 external decisions. The environment is synthetic and designed to represent regulated-banking constraints; it is not a production bank workload. Held-out evaluation tests unseen request templates within represented task categories, not unseen tasks or arbitrary production distributions.

5.1 Common canonical input

Raw prompts, request identifiers, preferred reference labels, acceptable reference labels, template identifiers, and synthetic evaluation labels are excluded from live provider payloads. Both live routers receive the same canonical structured state and the same policy-eligible candidates. This structured-only design improves information parity between mechanisms, but it may understate the potential advantage of a general-purpose LLM in routing problems where semantic interpretation of raw natural language is itself part of the routing task. The study therefore addresses bounded, already-structured control-plane decisions rather than model routing in general.

5.2 Live router configuration

The general-purpose path uses OpenAI structured response parsing. The specialized path uses the TypeSafe System One interface through OpenRouter with the eligible model IDs as the predefined choice set. The public documentation describes Jev as returning typed decisions rather than

Table 1: Live-router configuration. The two mechanisms receive the same canonical state and eligible candidate set but execute through different provider paths.

Property	General-purpose LLM	Specialized decision model
Model	GPT-4.1-mini-2025-04-14	Jev 1.13
Returned version	same pinned version	jev-1.13-20260917
Provider path	OpenAI direct	OpenRouter $\rightarrow$ TypeSafe
Input	canonical structured state	same
Eligible candidates	same	same
Raw prompt	excluded	excluded
Reference labels	excluded	excluded
Task-specific training	none	none
Decision output	selected model	selected model
Cost accounting	token usage + frozen price	provider-reported

free-form text [11]. Because the serving paths differ, operational results are attributed to these measured paths rather than to intrinsic model architecture alone.

5.3 Synthetic reference route

Evaluation uses *Synthetic Reference-Route Agreement (SRRA)*. For request x , task t , eligible set $E(x)$, configured task-quality score $q(m, t)$, estimated downstream cost $c(m, x)$, and tolerance $\delta = 0.12$,

$$q_{\max}(x, t) = \max_{m \in E(x)} q(m, t),$$

$$A(x) = \{m \in E(x) : q(m, t) \geq q_{\max}(x, t) - \delta\},$$

and the preferred synthetic reference route is

$$m_{\text{ref}}(x) = \arg \min_{m \in A(x)} (c(m, x), m_{\text{id}}),$$

where model ID provides deterministic tie breaking. If $E(x) = \emptyset$, the correct outcome is `NO_ROUTE`. Preferred SRRA measures exact agreement with m_{ref} (including correct abstentions); acceptable SRRA measures selection of any model in $A(x)$. SRRA is conformity to a configured synthetic objective, not human annotation, downstream response quality, or proof of an objectively optimal model.

Preferred and acceptable SRRA are compared with exact paired McNemar tests and Holm–Bonferroni correction. Confidence intervals use 10,000 request-level percentile bootstrap iterations at 95% confidence. Latency and cost are treated as descriptive operational measurements; no provider-independent latency superiority claim is made.

6 Primary Results

For preferred SRRA, specialized decision minus LLM is -1.0 percentage point on IID and -1.5 points on held-out evaluation. Holm-adjusted p values are 0.9165 and 0.7111. Acceptable SRRA differences are +1.0 and +0.5 points with adjusted $p = 1.0$ in both cases. These results show no statistically detectable difference, but do not establish equivalence or non-inferiority.

Positive table differences favor the specialized decision model. Bootstrap 95% confidence intervals for preferred SRRA are [69.5,81.5]% (LLM IID), [68.5,80.5]% (specialized decision IID), [65.0,77.5]%

Table 2: Synthetic Reference-Route Agreement (SRRA) including correct abstentions. The ML router is the only mechanism trained directly against the synthetic reference objective.

Router	Task-specific training	IID Pref.	Held Pref.	IID Acc.	Held Acc.
Rules	No	65.5%	63.0%	86.5%	88.5%
Weighted	No	78.0%	76.5%	84.0%	85.0%
ML	Yes: synthetic objective	99.0%	99.0%	99.5%	99.0%
General-purpose LLM	No	75.5%	71.5%	85.0%	83.5%
Specialized decision	No	74.5%	70.0%	86.0%	84.0%

Table 3: Paired specialized-decision versus general-purpose LLM comparison. Difference is specialized decision minus LLM. Exact McNemar tests use 200 paired observations per benchmark; adjusted p values use Holm–Bonferroni correction within benchmark and metric.

Benchmark	Metric	Absolute agreement difference (pp)	Discordant $(D+, L-)/(D-, L+)$	Adjusted p
IID	Preferred	-1.0	13 / 15	0.9165
IID	Acceptable	+1.0	8 / 6	1.0000
Held-out	Preferred	-1.5	13 / 16	0.7111
Held-out	Acceptable	+0.5	5 / 4	1.0000

(LLM held-out), and [63.5,76.5]% (specialized decision held-out). Acceptable-SRRA intervals are [80.0,89.5]%, [81.0,90.5]%, [78.0,88.5]%, and [79.0,89.0]%, respectively. Overlapping confidence intervals are descriptive and are not used as an equivalence test.

7 Operational Efficiency

Figure 3 compares preferred agreement with mean latency over requests that actually invoked the external live router.

Mean live routing latency is 1423.75 ms (IID) and 1406.17 ms (held-out) for the general-purpose LLM execution path versus 543.80 and 562.02 ms for the specialized-decision execution path. The observed mean-latency ratios are 2.62 and 2.50. Primary routing API cost across 304 calls per strategy is USD 0.223006 for the LLM path and USD 0.029983 for the specialized-decision path, a ratio of approximately 7.44. Cost accounting differs by provider: OpenAI cost is reconstructed from recorded token usage and frozen pricing, whereas Jev cost is reported by OpenRouter. These measurements are configuration-, provider-, network-, and time-specific.

Notably, the specialized-decision path records more total tokens (729,456) than the LLM path (540,145) while costing less under the frozen provider pricing. This reinforces that token count alone is insufficient for cross-provider economic comparison.

8 Repeated-Decision Stability

Fifty routable requests (25 per benchmark) were each executed five times by each live router, producing 500 successful live decisions. Full five-run selection consistency is shown in Figure 5. The result is descriptive because repeated observations are clustered within requests and the experiment was not designed as a statistical superiority test.

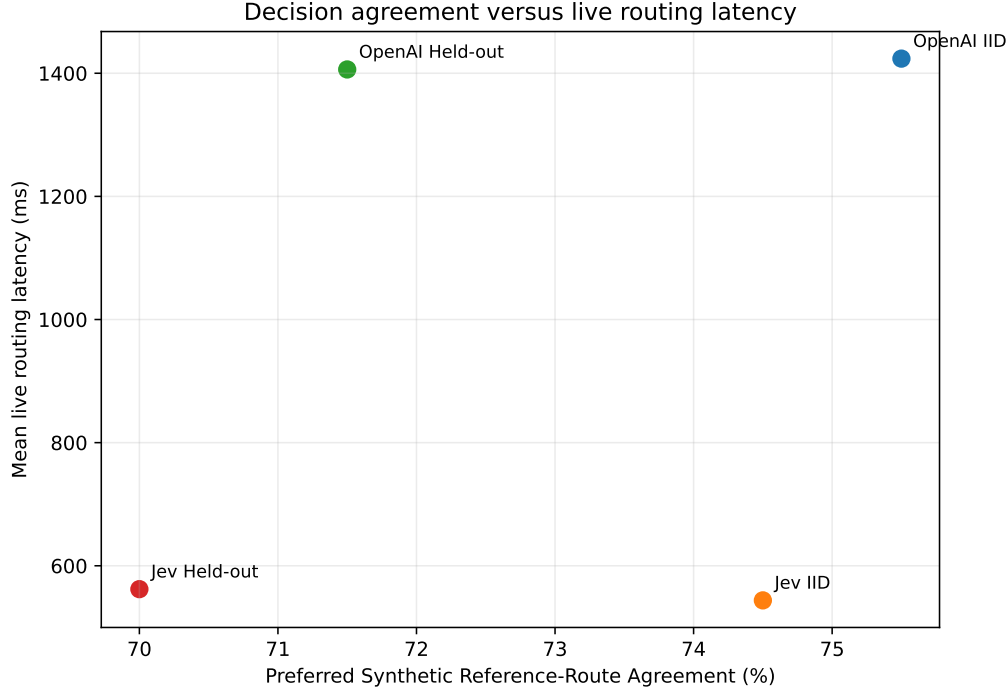


Figure 3: Preferred SRRA versus mean live routing latency. Latency is calculated over actual external routing decisions only.

9 Finding 2: Sophistication Was Not Monotonic

The Weighted heuristic achieves preferred SRRA of 78.0% IID and 76.5% held-out, numerically above both live routers. After Holm correction, neither live router shows a statistically detectable preferred- or acceptable-SRRA advantage over Weighted. The ML router reaches approximately 99% preferred SRRA with millisecond-scale local inference, but it was explicitly trained to reproduce the same synthetic routing objective. Its result demonstrates learnability and template-level generalization of that objective, not 99% real-world model-selection accuracy.

The broader observation is that increasing router sophistication did not monotonically improve the measured synthetic routing objective. This is directionally consistent with LLMRouterBench, which reports similar performance among many routing methods and failures of several sophisticated approaches to reliably beat a simple baseline under its own unified evaluation; that external result provides context, not validation of our synthetic benchmark [7]. The systems implication is not that simpler mechanisms are always better, but that the least-complex mechanism that adequately satisfies the decision requirements deserves explicit consideration.

10 Discussion: A Specialized Decision Layer

Most routing work begins with a model portfolio and asks how traffic should be allocated. Our results motivate a preceding question: what level and form of intelligence does the routing decision itself require? For bounded, high-frequency control-plane problems characterized by structured state, finite actions, explicit constraints, measurable objectives, repeated execution, latency or cost sensitivity, and limited need for open-ended generation, a specialized fast decision layer may be sufficient. TypeSafe calls Jev’s family “System One” models; we treat that phrase as a useful

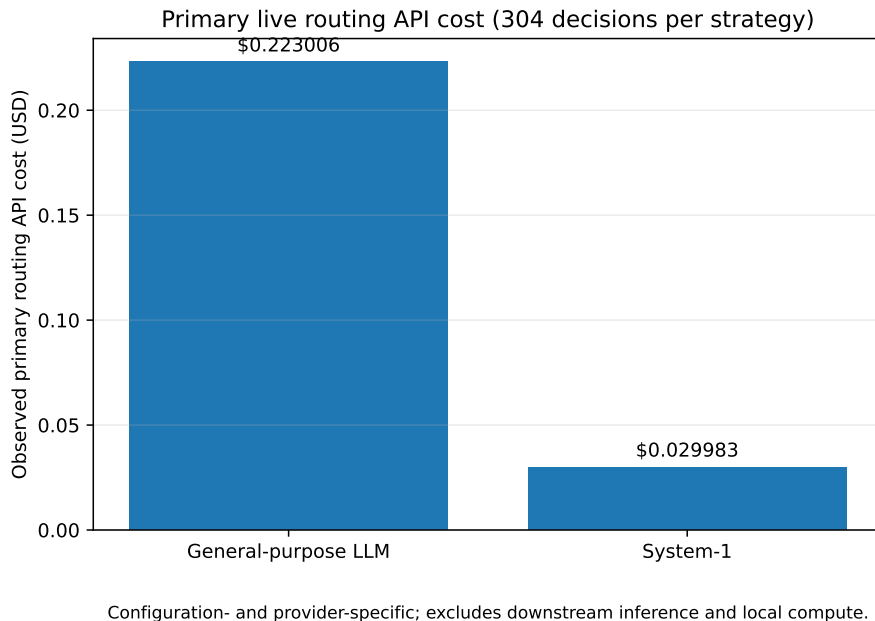


Figure 4: Observed external routing API cost for the primary live experiment. Each strategy made 304 external decisions. Values exclude downstream inference and local compute and should not be interpreted as total cost of ownership. OpenAI cost is reconstructed from recorded usage and frozen pricing; Jev cost is provider-reported by OpenRouter.

architectural shorthand rather than a universal scientific taxonomy.

The concept may extend to tool selection, agent selection, API selection, workflow dispatch, resource placement, and escalation, but these applications remain hypotheses for future evaluation.

A practical decision framework follows: use rules for explicit stable logic; weighted heuristics for explicit multi-objective scoring; supervised ML when a stable objective and representative labels exist; specialized decision inference for bounded adaptive decisions; and general-purpose LLMs when the decision itself requires substantial semantic interpretation, ambiguity resolution, planning, or open-ended reasoning. The core principle is to use the least-complex mechanism that adequately meets the decision’s quality, adaptability, latency, cost, and governance requirements.

11 Governance Implications

Across the primary and stability experiments, 1,108 live routing decisions completed successfully with no recorded ineligible selections or policy violations. This does not imply model-learned governance. Eligibility is enforced before routing and validated afterward. The design principle is therefore: *governance determines what is permitted; appropriately selected decision intelligence determines what is preferred.*

12 Threats to Validity and Claim Boundaries

Construct validity. Synthetic Reference-Route Agreement measures conformity to a configured routing objective, not downstream answer quality, user preference, safety, or an objectively optimal model. The preferred route is partly determined by configured quality and cost assumptions.

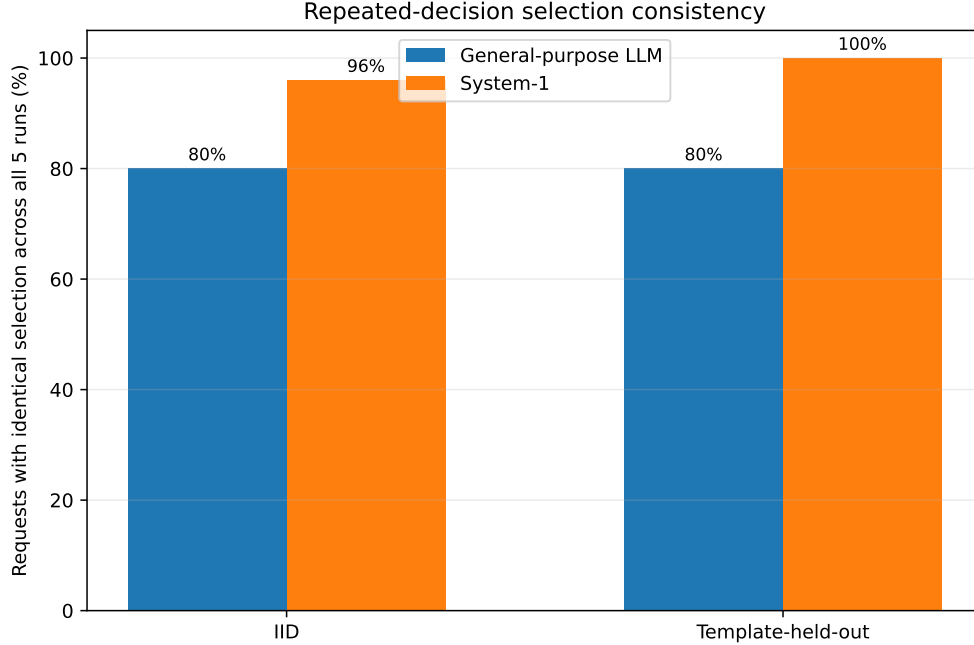


Figure 5: Percentage of requests for which all five repeated executions selected the same model.

Internal validity. Both live routers receive the same canonical structured state and eligible candidates, and prior live-pilot requests were excluded from final sampling. However, the two providers implement different model architectures, serving stacks, and pricing mechanisms; the experiment cannot isolate which component causes the observed latency or cost difference.

ML comparability. The Random Forest is trained directly against the synthetic objective. Its approximately 99% agreement is evidence that this objective is learnable and generalizes across held-out templates, not evidence that supervised ML is universally superior to zero-shot routing.

Generalization. Held-out evaluation is template-level within known task categories, not task-held-out or production-distribution evaluation. The synthetic banking environment does not reproduce the full organizational, regulatory, traffic, or data complexity of a production financial institution. The structured-only input design also intentionally removes raw-prompt semantics; this improves mechanism-level comparability but limits inference to already-structured control-plane decisions and may understate the advantage of a general-purpose LLM when semantic interpretation is itself required.

Statistical boundary. Failure to reject a difference between the specialized decision model and the LLM does not establish equivalence or non-inferiority. A future study should pre-specify a practically meaningful equivalence margin and power the sample accordingly. Stability comparisons are descriptive because repetitions are clustered within requests.

Operational boundary. Latency and cost depend on model versions, providers, network conditions, prompts, service tiers, and September 2026 pricing. Specialized-decision measurements are for Jev via OpenRouter, not direct Jev API execution; the experiment therefore measures execution paths rather than intrinsic architecture speed. External routing API cost excludes downstream inference and local compute and is not total cost of ownership.

Governance boundary. Zero recorded policy violations do not demonstrate learned policy compliance. Governance is enforced deterministically before routing and validated afterward.

13 Interpretation of RQ1 and Operational Hypotheses

For RQ1, observed preferred and acceptable SRRA are close and paired tests detect no difference after correction, but the experiment does not establish equivalence or non-inferiority. The strongest defensible agreement conclusion is therefore that *no SRRA advantage for the general-purpose LLM was detected in the evaluated structured routing task*. H1a and H1b are supported descriptively in the frozen configuration: the specialized-decision execution path has lower measured mean live routing latency and lower measured external routing API cost. These operational findings are configuration- and path-specific and are not attributed to model architecture alone.

14 Future Work

Future work should include equivalence or non-inferiority designs with pre-specified margins, independently human-labelled routing decisions, downstream execution for end-to-end quality/cost/latency measurement, production-derived regulated workloads, and experiments that explicitly vary decision complexity. A particularly useful direction is to learn the boundary

decision complexity $\rightarrow$ required intelligence class.

15 Conclusion

In this policy-aware routing experiment, a specialized decision model and a general-purpose LLM produced closely aligned Synthetic Reference-Route Agreement, with no statistically detectable agreement difference. Their measured execution paths differed substantially: the specialized-decision path showed approximately 2.50–2.62 times lower observed mean live routing latency and approximately 7.44 times lower observed routing API cost under the evaluated configuration. Repeated execution also showed greater observed selection consistency for the specialized-decision path, descriptively. Meanwhile, neither live router showed a statistically detectable SRRA advantage over a simple weighted heuristic after correction, and supervised ML reproduced the synthetic objective extremely closely when trained directly against it.

The results therefore do not support a universal winner. They support a systems-design principle: **general-purpose LLMs need not be the default intelligence mechanism for bounded, already-structured AI control-plane decisions**. The appropriate mechanism should be selected according to the structure, uncertainty, adaptability, latency, cost, semantic interpretation, and governance requirements of the decision.

Reproducibility Statement

Frozen release: v1.0.0-research; release commit 69386da90159d9171afaa8c2587973e162e6b6fb; primary run 20260929T152852-8d825a67; stability run 20260929T155310-839c6b13; final sample hash 70db720aadaa0322b2a9f29b9e9a796269d44d780c7d6c7023c11914eee7f7c0. The frozen repository (<https://github.com/debabratapruseth/Smart-Model-Router-for-Regulated-AI>) preserves the experimental inputs, configurations, decisions, statistical outputs, and provenance needed to inspect the reported evidence. Future re-execution of external provider calls is not guaranteed to reproduce identical selections, latency, availability, or pricing even when model identifiers are pinned.

References

- [1] L. Chen, M. Zaharia, and J. Zou. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *Transactions on Machine Learning Research*, 2024.
- [2] D. Ding, A. Mallick, C. Wang, R. Sim, S. Mukherjee, V. Rühle, L. V. S. Lakshmanan, and A. H. Awadallah. Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing. *International Conference on Learning Representations*, 2024.
- [3] P. Aggarwal et al. AutoMix: Automatically Mixing Language Models. *Advances in Neural Information Processing Systems 37*, 2024. doi:10.52202/079017-4164.
- [4] Q. J. Hu, J. Bieker, X. Li, N. Jiang, B. Keigwin, G. Ranganath, K. Keutzer, and S. K. Upadhyay. RouterBench: A Benchmark for Multi-LLM Routing System. arXiv:2403.12031, 2024.
- [5] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica. RouteLLM: Learning to Route LLMs from Preference Data. *International Conference on Learning Representations*, 2025.
- [6] J. Dekoninck, M. Baader, and M. Vechev. A Unified Approach to Routing and Cascading for LLMs. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:12987–13010, 2025.
- [7] H. Li et al. LLMRouterBench: A Massive Benchmark and Unified Framework for LLM Routing. *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37733–37754, 2026. doi:10.18653/v1/2026.findings-acl.1881.
- [8] H. Mozannar and D. Sontag. Consistent Estimators for Learning to Defer to an Expert. *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:7076–7087, 2020.
- [9] E. Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 2023. doi:10.6028/NIST.AI.100-1.
- [10] X. Wang, S. Wang, Y. Zhu, and B. Liu. System-1.5 Reasoning: Traversal in Language and Latent Spaces with Dynamic Shortcuts. *Advances in Neural Information Processing Systems 38*, 2025.
- [11] OpenRouter. TypeSafe: Jev 1.13 – API Pricing & Providers. Model documentation for `typesafe/jev-1.13`, accessed September 2026.