DOI: 10.2507/37th.daaam.proceedings.xxx

EVIDENCE-GATED SELECTION OF LARGE LANGUAGE MODEL CONFIGURATIONS FOR TASK-SPECIFIC DEPLOYMENT

Kemal Zahirovic & Kresimir Rakic

This Publication has to be referred as: Zahirovic, K[emal] & Rakic, K[resimir] (2026). Evidence-Gated Selection of Large Language Model Configurations for Task-Specific Deployment, Proceedings of the 37th DAAAM International Symposium, pp.xxxx-xxxx, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-xx-x, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/37th.daaam.proceedings.xxx

Abstract
Large language model (LLM) selection is commonly reduced to leaderboard scores or a single cost-quality objective, although deployment decisions often contain hard constraints that cannot be compensated by higher average performance. This paper proposes an evidence-gated selection framework that separates feasibility, minimum acceptance, evidence sufficiency, and comparative objectives. A candidate is evaluated as a deployable configuration, not only as a base model, including inference settings, retrieval, tools, permissions, runtime, and human-review controls. To ground the framework empirically, a secondary evidence audit was conducted over recent 2025–2026 studies covering real-world instruction constraints, benchmark contamination, RAG faithfulness, tool use, LLM judging, and cost-aware routing. The reviewed results show gaps between generic benchmark capability and workflow-specific requirements: TRUEBench reports a 69.07% overall pass rate for o1, BFCL identifies persistent weaknesses in memory and long-horizon tool use, and LiveBench addresses benchmark staleness and contamination through frequently refreshed evaluation tasks. Recent routing studies report cost reductions under explicit constraints, while the 2026 LLMRouterBench study shows that several routing methods do not reliably outperform simple baselines. The evidence supports the proposed decision structure, but does not validate a universal model ranking. The framework treats missing evidence as unresolved eligibility and requires scenario-specific testing before deployment.
Keywords: large language models; model selection; LLM routing; evaluation; evidence gating; tool use; RAG; deployment

1. Introduction

Selecting an LLM for an operational system is not equivalent to selecting the model with the highest public benchmark score. A deployment has a defined task, data boundary, latency or cost envelope, tool permissions, reliability requirements, and failure consequences. These conditions create different kinds of evidence. A model may be strong on general reasoning yet fail a required instruction constraint, produce unsupported RAG content, or use tools incorrectly. Conversely, a cheaper model may be adequate for a restricted task even when it is weaker on broad benchmarks. Recent research on LLM routing formalizes this cost-quality trade-off, but it also shows that routing performance depends on how the candidate set and evaluation are constructed [1][2][3].

Working Paper of 37th DAAAM Symposium

The practical problem is therefore a decision problem under heterogeneous evidence. Existing benchmark suites measure valuable capabilities, but no single benchmark establishes feasibility for a particular deployment. LiveBench was designed to reduce contamination and benchmark staleness through frequently refreshed tasks, while BFCL separates single-turn function calling from stateful and multi-step agent behaviour [4][5], Studies of realistic instruction following and deployed RAG systems likewise report failure modes that are not visible in generic leaderboards [6][7].

This paper develops an evidence-gated framework with two principles. First, mandatory requirements are non-compensatory: a severe violation cannot be cancelled by a higher average score elsewhere. Second, evidence is assessed against the claim it supports. A benchmark may support a capability claim without supporting a deployment claim. The intended result is not a universal ranking of models, but a traceable path from requirements to admissibility and then to comparative choice.

2. Related work and empirical basis

LLM routing directly addresses model choice under cost and quality constraints. RouteLLM reports more than 2× cost reduction without sacrificing response quality in its evaluated settings [1]. BEST-Route extends this idea by selecting both the model and the number of samples at inference time and reports up to 60% cost reduction with less than a 1% performance drop on its real-world datasets [2]. OctoSelector formulates batch-aware selection as an optimization problem with explicit accuracy, latency, and cost constraints and reports up to 67.7% cost reduction on NL2SQL workloads [3]. These results support a comparative stage that treats cost and quality jointly, but they do not imply that a routing policy is valid for an unseen application.

The 2026 LLMRouterBench study provides an important counterweight. Its unified evaluation contains more than 400,000 instances, 21 datasets, and 33 models. It confirms model complementarity but reports that several recent routing methods, including commercial routers, do not reliably outperform a simple baseline; it also finds diminishing returns from simply enlarging the candidate ensemble [8]. This supports evidence gating for the selector itself: a sophisticated selection algorithm must be evaluated, not assumed to improve decisions.

The evidence needed before comparison also depends on workflow type. TRUEBench was designed around implicit constraints, multilingual instructions, and multi-turn interactions; the reported overall pass rate for OpenAI o1 was 69.07%, illustrating that strong general models can remain far from perfect under realistic constraints [6]. BFCL evaluates serial and parallel calls, abstention, and stateful multi-step tool use; its results show that single-turn calling is substantially easier than memory, dynamic decisions, and long-horizon interaction [5]. For RAG, recent EMNLP work reports persistent unsupported or contradictory generations and introduces FaithJudge to evaluate faithfulness separately from ordinary answer quality [7].

Finally, evidence about an LLM can itself be unreliable when the evaluator is an LLM. RobustJudge studies 15 attacks, 8 defenses, and 13 judge models and finds that robustness is sensitive to prompt templates and judge-model choice [9]. A 2026 survey likewise identifies reliability, bias, and standardization as central unresolved issues in LLM-as-a-judge systems [10]. Evaluation evidence therefore needs its own sufficiency check rather than being accepted solely because it produces a numerical score.

3. Evidence-gated selection framework

The study is guided by three questions: RQ1, how can mandatory deployment constraints be kept distinct from comparative preferences; RQ2, how can evidence be judged as sufficient for a specific claim; and RQ3, how should recent empirical findings alter the selection path? The framework is a design-science artefact: requirements are derived from the decision problem, formalized as decision rules, and checked against an external evidence base. The present paper does not claim prospective causal validation of the artefact.

The framework uses a non-compensatory gate followed by a compensatory comparison. This ordering is deliberate. A weighted score is useful when alternatives are already acceptable, but it is unsuitable as the first filter when one condition is mandatory. A candidate that violates a data-location rule, exceeds a hard latency limit, or fails a safety-critical tool restriction should not remain eligible simply because it performs well on unrelated tasks.

The framework evaluates a candidate as a configuration C rather than as a bare model name. C includes the model and version, inference parameters, system instructions, context and retrieval strategy, memory, tools and permissions, validators, human-review controls, deployment runtime, and relevant hardware. Evidence is transferable only when these configuration differences do not invalidate the claim being transferred.

Working Paper of 37th DAAAM Symposium

The procedure has five decision stages.

- *Stage 1* defines the workflow and records users, inputs, outputs, operating conditions, failure consequences, and measurable service requirements.
- *Stage 2* converts these requirements into explicit decision roles. Hard feasibility constraints establish whether the configuration is permissible; acceptance thresholds establish minimum task performance; evidence requirements establish whether the available measurements are sufficiently relevant; comparative objectives are applied only among configurations that pass the mandatory gates.
- Each requirement should be assigned one role before evaluation: hard feasibility, minimum acceptance, evidence sufficiency, or comparative preference. The same measurable variable can occupy different roles at different levels. For example, latency may be a maximum service-level constraint for eligibility and a preference for further optimization below that limit. Recording this distinction before viewing candidate results reduces the risk of moving the acceptance boundary to favour a particular model.
- *Stage 3* builds an evidence record for each material claim. The record should identify the source and date, configuration evaluated, task population, test conditions, sample size, repetitions, metric, result, and known limitations. Evidence is considered adequate only when it answers three questions: does it address the required capability under relevant conditions; is the evaluation method valid for the intended interpretation; and is the coverage sufficient for the consequence of the decision? Absence of evidence is not treated as evidence of poor performance.
- *Stage 4* applies predefined acceptance tests. Aggregate scores are not permitted to hide failure modes with independent consequences. For an RAG system, retrieval quality, grounding/faithfulness, answer correctness, and unanswerable-question handling should be tested separately. For an agent, tool selection, argument validity, permissions, final system state, and recovery behaviour are separate claims. Repeated trials estimate execution variability, whereas repeated runs do not create new independent tasks.
  The outcome is one of four statuses: infeasible (a mandatory constraint is violated); below acceptance (adequate evidence demonstrates failure of a minimum threshold); evidence insufficient (the available evidence cannot resolve eligibility); or eligible for comparison (all mandatory conditions are adequately supported). These statuses intentionally prevent an uncertain result from becoming a false pass or a false fail.
- *Stage 5* compares only eligible configurations. Dominance can remove candidates that are no better on any selected objective and worse on at least one, subject to measurement uncertainty. When trade-offs remain, a documented multi-criteria objective may combine cost, latency, quality, and human-correction burden. Related DAAAM research demonstrates preference-based evaluation among efficient alternatives [11]. Sensitivity analysis should be used when weights or thresholds are uncertain. The selected configuration then enters pilot validation and monitoring; a material change to the model, prompt, retrieval, tools, permissions, or runtime triggers a review of evidence validity.

4. Empirical evidence audit (2025–2026)

This study conducts a structured secondary audit of ten published studies from 2025–2026 rather than a primary experimental comparison. Studies were selected for their relevance to the proposed framework's decision stages and for reporting explicit evaluation methods, measurable outcomes, or documented limitations. The reported findings are drawn from the cited publications and are not measurements produced by the present authors. Results were not statistically pooled because the studies differed in datasets, metrics, candidate models, and operating conditions. Instead, the audit maps published evidence to the selection decisions it can support:

- **TRUEBench** [6]. A reported 69.07% overall pass rate for o1 on realistic constraint-heavy tasks shows why constraint compliance needs direct testing.
- **LiveBench** [4]. Frequent refresh and contamination control support freshness and task-relevance checks.
- **BFCL** [5]. Single-turn tool use is stronger than memory, dynamic decisions and long-horizon agentic use; these capabilities should be tested separately.
- **RAG faithfulness** [7]. Relevant retrieved context can still yield unsupported or contradictory answers; grounding and faithfulness require separate checks.
- **RouteLLM/BEST-Route** [1],[2]. Reported >2× and up to 60% cost reductions show that cost-quality routing can matter after mandatory gates are satisfied.
- **LLMRouterBench/OctoSelector** [3],[8]. Routing does not uniformly beat simple baselines; the selector itself must therefore be validated.

Across the reviewed studies, no single public benchmark supplies all evidence needed for deployment. The audit therefore supports a decision rule: general benchmark scores alone cannot establish deployment eligibility. A configuration becomes eligible for comparison only when sufficient scenario-relevant evidence demonstrates that mandatory constraints and acceptance thresholds have been satisfied. This is a methodological result rather than a claim that one model is superior.

Working Paper of 37th DAAAM Symposium

The audit also shows why evidence gating should be applied to evaluation mechanisms. When an LLM judge is used to score candidates, the judge model, prompt, and evaluation protocol become additional configuration variables. RobustJudge reports sensitivity to judge-model and prompt-template choices [9], while a 2026 survey identifies reliability, bias, and standardization as continuing concerns in LLM-as-a-judge systems [10]. A selection system should therefore retain raw task outcomes and, where stakes justify it, use programmatic checks or independent validation rather than relying on a single automated judge score.

5. Discussion and limits

The contribution is a structured decision method, not a new universal benchmark or a new routing model. Its main benefit is traceability: each rejection, unresolved case, or progression decision is tied to a requirement and an evidence record. This distinction is important because model selection, routing, and deployment validation answer different questions. The framework also makes explicit that application constraints can change the decision boundary: latency may be a hard service limit, a soft preference, or both at different levels.

The principal limitation is that the study relies on a secondary literature audit rather than prospective validation in a production environment. Since the reviewed studies differ in tasks, configurations, and evaluation conditions, their results cannot establish a universal model ranking or demonstrate that the proposed framework improves selection outcomes. Future work will evaluate the framework on an end-to-end industrial RAG system, comparing evidence-gated selection with leaderboard-based and weighted-score baselines using the same candidate pool, representative tasks, and predefined requirements [8][11]. After controlled offline testing, configurations passing all feasibility, acceptance, and evidence-sufficiency gates will undergo randomized A/B testing, where feasible, with primary measurements of answer quality, constraint violations, p95 latency, and actual inference cost [3][7]. Repeated trials, independent quality checks, and sensitivity analysis will assess reliability and decision stability [9]. Deployment monitoring will support re-evaluation following significant performance or configuration changes [12].

6. Conclusion

Recent 2025–2026 evidence supports a shift from score-based model choice to evidence-gated configuration selection. Realistic constraint benchmarks, tool-use evaluations, RAG faithfulness studies, contamination-resistant benchmarks, routing experiments, and judge-reliability research all show that the decision cannot be reduced to one aggregate capability score. The proposed framework responds by separating feasibility, acceptance, evidence sufficiency, and comparative objectives, and by treating unresolved evidence as unresolved eligibility. Its practical conclusion is deliberately limited but actionable: a model should become a candidate for comparative optimization only after the evidence required by the deployment scenario has been demonstrated. Future work (industrial validation) should prospectively test whether this procedure reduces unsuitable selections, evaluation effort, or deployment failures in real systems.

7. References

[1] Ong, I.; Almahairi, A.; Wu, V.; Chiang, W.-L.; Wu, T.; Gonzalez, J. E.; Kadous, M. W. & Stoica, I. (2025). RouteLLM: Learning to route LLMs from preference data, Proceedings of the 13th International Conference on Learning Representations (ICLR 2025). https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html

[2] Ding, D.; Mallick, A.; Zhang, S.; Wang, C.; Madrigal, D.; Hipolito Garcia, M. D. C.; Xia, M.; Lakshmanan, L. V. S.; Wu, Q. & Rühle, V. (2025). BEST-Route: Adaptive LLM routing with test-time optimal compute, Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 267, pp. 13870–13884, PMLR. https://proceedings.mlr.press/v267/ding25d.html

[3] Zhang, G.; Lin, Y. & Mehrotra, S. (2026). OctoSelector: Efficient and effective batch-aware model selection for large language models, Proceedings of the ACM on Management of Data, Vol. 4, No. 3, Article 218, pp. 1–26. https://doi.org/10.1145/3802095

[4] White, C.; Dooley, S.; Roberts, M.; Pal, A.; Feuer, B.; Jain, S.; Shwartz-Ziv, R.; Jain, N.; Saifullah, K.; Dey, S.; Shubh-Agrawal; Sandha, S.; Naidu, S.; Hegde, C.; LeCun, Y.; Goldstein, T.; Neiswanger, W. & Goldblum, M. (2025). LiveBench: A challenging, contamination-limited LLM benchmark, Proceedings of the 13th International Conference on Learning Representations (ICLR 2025). https://proceedings.iclr.cc/paper_files/paper/2025/hash/e4a46394ba5378b3f9a186a5b4c650d1-Abstract-Conference.html

[5] Patil, S. G.; Mao, H.; Yan, F.; Ji, C. C.; Suresh, V.; Stoica, I. & Gonzalez, J. E. (2025). The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models, Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 267, pp. 48371–48392, PMLR. https://proceedings.mlr.press/v267/patil25a.html

Working Paper of 37th DAAAM Symposium

[6] Park, J.; Song, J.; Choi, M.; Heo, K.; Huh, T. & Kim, J. W. (2025). TRUEBench: Can LLM response meet real-world constraints as productivity assistant?, Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 14383–14409, Association for Computational Linguistics, Suzhou, China. https://doi.org/10.18653/v1/2025.findings-emnlp.776

[7] Tamber, M. S.; Bao, F. S.; Xu, C.; Luo, G.; Kazi, S.; Bae, M.; Li, M.; Mendelevitch, O.; Qu, R. & Lin, J. (2025). Benchmarking LLM faithfulness in RAG with evolving leaderboards, Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 799–811, Association for Computational Linguistics, Suzhou, China. https://doi.org/10.18653/v1/2025.emnlp-industry.54

[8] Li, H.; Zhang, Y.; Guo, Z.; Wang, C.; Tang, S.; Zhang, Q.; Chen, Y.; Qi, B.; Ye, P.; Bai, L.; Wang, Z. & Hu, S. (2026). LLMRouterBench: A massive benchmark and unified framework for LLM routing, Findings of the Association for Computational Linguistics: ACL 2026, pp. 37733–37754, Association for Computational Linguistics, San Diego, USA. https://doi.org/10.18653/v1/2026.findings-acl.1881

[9] Li, S.; Xu, C.; Wang, J.; Gong, X.; Chen, C.; Zhang, J.; Wang, J.; Lam, K.-Y. & Ji, S. (2025). LLMs cannot reliably judge (yet?): A comprehensive assessment on the robustness of LLM-as-a-judge, arXiv preprint, arXiv:2506.09443. https://doi.org/10.48550/arXiv.2506.09443

[10] Gu, J.; Jiang, X.; Shi, Z.; Tian, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; Wang, S.; Zhang, K.; Lin, Z.; Zhang, B.; Ni, L. M.-S.; Gao, W.; Wang, Y. & Guo, J. (2026). A survey on LLM-as-a-judge, The Innovation, Vol. 7, No. 6, Article 101253. https://doi.org/10.1016/j.xinn.2025.101253

[11] Zhao, Q.; Balugani, E.; Gamberini, R. & Lolli, F. (2025). A value-based decision framework for evaluating the integrated efficiency of smart grids, Proceedings of the 36th DAAAM International Symposium, pp. 0131–0140, B. Katalinic (Ed.), DAAAM International, ISBN 978-3-902734-47-1, ISSN 1726-9679, Vienna, Austria. https://doi.org/10.2507/36th.daaam.proceedings.017

[12] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, USA. https://doi.org/10.6028/NIST.AI.100-1

Working Paper of 37th DAAAM Symposium