Reliable but Not Right: Automated Evaluation of Generated R&D Ideas

Adrian Mladenić Grobelnik
adrian.m.grobelnik@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

Businesses regularly face the decision of which research ideas to fund, and an automated ranking of the candidate research and development ideas can provide decision support for it. As with any business intelligence tool, the ranking is worth acting on only if it reflects the judgment of the decision maker. Automated research systems generate more candidates than anyone can read, and large language models are commonly used to rank them. Such an automated evaluator is validated along several dimensions, accuracy on past work, self-consistency, position bias and agreement between models, none of which needs a human expert to look at the candidates it will rank. We propose a design for validating automated evaluators of generated research and development ideas against a human expert. We applied it in three studies: a citation-impact predictor choosing the starting concepts (seeds) for hypotheses, large language models ranking generated hypotheses, and language models ranking generated papers. We were surprised that in each study the evaluator passed its automated validation on every dimension, yet its ranking diverged from the human expert's: it was reliable but not right. Our experimental validation showed that recognising a good research idea involves something experts do that current models do not.

Keywords

LLM-as-judge, automated evaluation, human evaluation, research idea generation, self-consistency, position bias, citation impact prediction, research and development, decision support

1 Introduction

Research and development (R&D) runs on a choice no one can make well at scale: which ideas to fund before their outcome is known. Decision makers can read only a fraction of what is proposed. Automated research systems widen this gap because they generate hypotheses, experiments and papers faster than any reviewer can read them [1, 2]. A large language model is therefore asked to rank the candidates before a person sees them [3]. That ranking is a business intelligence tool [4] for the funding decision, but it helps only when the automated evaluator ranks the candidates the way the decision maker would have.

This paper asks whether an evaluator that passes validation without a human expert also ranks candidates the way a human expert would. We tested the question in three studies: (1) a citation-impact predictor, a model trained on past papers with known outcomes, chose the starting concepts for generating hypotheses; (2) large language model judges ranked generated hypotheses; and (3) large language model judges ranked generated papers. Each evaluator was validated along several dimensions before it ranked anything. By concept we mean a Wikipedia entry that a paper's title or abstract mentions.

The impact predictor was validated by its accuracy on held-out papers. The judges were validated on self-consistency, whether they return the same verdict when a pair of candidates is shown again. They were also validated on position bias, whether the verdict changes when the two candidates swap places [5], and on agreement between judges. None of these requires a human expert to examine the candidates being ranked.

We propose a design for validating an automated evaluator of generated R&D ideas against a human expert, and apply it in all three studies.

Contributions.

(1) *A design for validating automated evaluators of generated R&D ideas.*
(2) *Insights into the ability of current models to recognise high-relevance research.*

2 Related Work

Large language model judges. Earlier work validates a large language model judge against human labels for the task it judges, but the judges in our studies had no such labels when they were chosen. Zheng et al. [3] found a large language model agreeing with human preference on more than 80% of chatbot conversation pairs, yet found position, verbosity and self-enhancement biases alongside that agreement. Wang et al. [5] showed that swapping two candidates can reverse the verdict. Chen et al. [6] found judges misled by authority cues and presentation. Bavaresco et al. [7] found that agreement between judges and human annotators varies widely across 20 tasks.

Automated research and its evaluation. Automated research systems now incorporate such judges. The AI Scientist [1], the pipeline of Mladenić Grobelnik [8], Co-Scientist [2] and SciMON [9] generate hypotheses or whole papers. All four use a large language model to review them, rank them or assess their novelty. Si et al. [10] had more than 100 NLP researchers review generated and human-written research ideas: the generated ideas were judged more novel but slightly less feasible. That is the kind of human evaluation an automated evaluator stands in for.

Impact prediction. Citation-impact predictors, the first kind of evaluator we study, have been validated against what was later published or cited rather than against an expert's view of the ideas themselves. Uzzi et al. [11] link citation impact to unusual combinations of prior work. Gu and Krenn [12] forecast

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.82

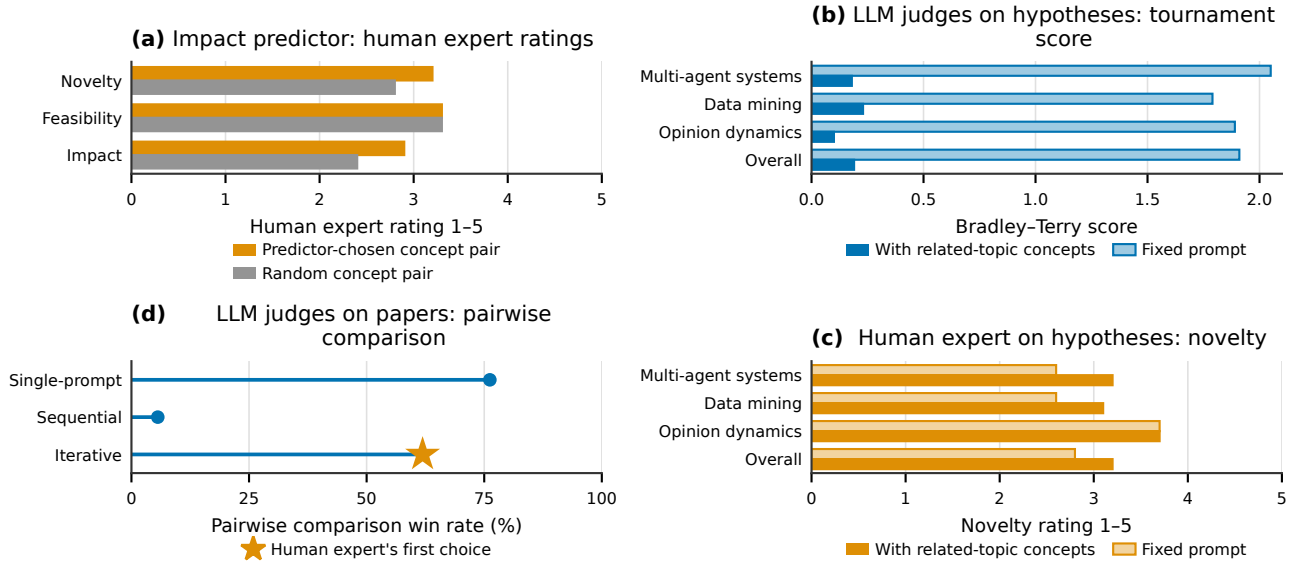


Figure 1: (a) The human expert's mean ratings of ten predictor-chosen and ten random hypotheses on novelty, feasibility and impact.(b) and (c) The 40 hypotheses by field: the judges preferred those without concepts from a related topic everywhere, the human expert those with them. (d) The three papers on the first hypothesis: the share of comparisons each strategy won when the judges saw only the paper. The star marks the human expert's first choice. Blue marks the automated evaluator's ranking, orange the human expert's.

the impact of concept pairs not yet published from an evolving knowledge graph. Vital et al. [13] predict citation impact from text embeddings. Outside research, business uses of large language models such as data preparation in analytics [14] and business process modelling [15] can be checked against a known answer.

3 The Proposed Design

Figure 2 shows the design: the automated evaluator first passes its validation without a human expert, and a human expert then assesses the candidates it ranked. Each study applies this design with a single expert: a domain expert, a domain researcher and a code reviewer respectively, each called the expert below.

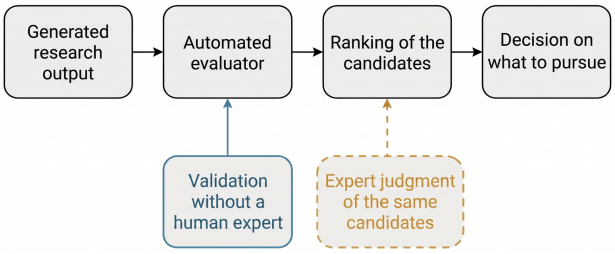


Figure 2: The proposed design for validating an automated evaluator of generated research output. The evaluator first passes its validation without a human expert, then a human expert assesses the candidates it ranked.

3.1 Citation-Impact Predictor for Generating Hypotheses

Can an accurate citation-impact predictor generate hypotheses that match a human expert's preferences? The idea is straightforward: train a model on papers with known citation outcomes, then let its predictions choose the starting concepts for generating hypotheses. We collected 1.71M distinct OpenAlex [16] papers from 2016 to 2026, the 200 highest-impact and 200 random papers in each of 4,516 topics. We trained nine models to predict field-weighted citation impact, a paper's citations divided by the average for its field and year. The inputs were an embedding of the title and abstract, metadata such as year and author count, and the venue's average impact. Further inputs were references shared with other papers, keywords and the concepts in the abstract. The impact predictor passed its validation without a human expert because it predicted the impact of held-out papers well. The strongest model, an AutoGluon ensemble [17], reached Spearman 0.73 with the true field-weighted citation impact on a held-out 25% over three runs. Its ROC-AUC at identifying the 10% most cited papers was 0.87. An FT-Transformer [18] followed at 0.72 and 0.86.

A domain expert then chose ten research directions in the field of multi-agent systems. For each direction a large language model generated two hypotheses, each from a concept pair. In the first, a concept derived from the research direction was paired with the highest-scoring concept, according to the impact predictor, among those not yet combined with it in any paper. In the second, the partner concept was chosen at random from those no paper had yet combined with the first. This gave a predictor-chosen set and a random set of ten hypotheses each, and the expert rated every hypothesis from 1 to 5 on novelty, feasibility and impact.

Result. The expert's ratings show no significant difference between the predictor-chosen set and the random set on novelty, feasibility or impact, as Figure 1a shows. The predictor-chosen set scored 3.2 against 2.8 on novelty, 3.3 against 3.3 on feasibility and 2.9 against 2.4 on impact, $p = 0.32$, 0.93 and 0.12. Ten ratings per set give little statistical power, so this is a null result, not evidence that the impact predictor chooses worse than random.

3.2 Large Language Model Judges on Generated Hypotheses

Can large language model judges that are self-consistent and positionally unbiased rank research hypotheses as a human expert would? Three large language models, none of which later served as a judge, generated 40 hypotheses across the domains of multi-agent systems, data mining and opinion dynamics from the same OpenAlex data [16]. Half the prompts named two concepts from a related topic, meaning concepts that papers in that topic use but papers in the target field do not. The other half used a fixed prompt without concepts from a related topic, giving 20 hypotheses with such concepts and 20 without.

We screened eight large language models as candidate judges on self-consistency and position bias, without a human expert. Table 1 reports the results: we kept the two models with the highest self-consistency, Claude Sonnet 4.5 [19] and GPT-5.2 [20], as the judges we compared with the expert. We expected their ranking to align with the expert's judgement.

Table 1: Validation of the eight candidate judges, without a human expert. Self-consistency is the share of pairs that return the same winner in both orders, higher is better. Position bias is how far the first-position win rate deviates from 50%, lower is better.

Model	Self-consistency	Position bias
Claude Sonnet 4.5	**100%**	**0%**
GPT-5.2	**89%**	**14%**
GPT-5-mini	80%	20%
GPT-5.2 Pro	79%	11%
Claude Opus 4.5	75%	25%
Gemini 3 Flash	75%	25%
gpt-oss-120b	75%	25%
Grok 4.1 Fast	65%	35%

The judges ranked all 40 hypotheses by pairwise comparison using the Bradley–Terry model [21], discarding any vote that reversed when the two candidates swapped places. The expert, a domain researcher, then independently rated every hypothesis from 1 to 5 on novelty and on feasibility.

Result. The judges preferred hypotheses without concepts from a related topic in every field, with a mean novelty Bradley–Terry score of 1.91 against 0.19. The expert rated hypotheses with concepts from a related topic more novel in multi-agent systems and data mining, equally novel in opinion dynamics, and 3.2 against 2.8 overall, $p = 0.055$, as Figure 1b and c show. Those same hypotheses were slightly less feasible, 3.5 against 3.8, $p = 0.16$. The judges and the expert agreed on feasibility, but on novelty the judges preferred hypotheses without concepts from a related topic while the expert preferred those with them. Within each field, the judges' novelty score and the expert's novelty rating had a Spearman correlation of 0.00 in multi-agent systems, 0.05 in data mining and 0.83 in opinion dynamics, over 17, 17 and six hypotheses.

3.3 Large Language Model Judges on Generated Papers

Can large language model judges that are self-consistent, positionally unbiased and in agreement with each other rank papers as a human expert would? Using the agentic pipeline of Mladenić Grobelnik [8], we compared three strategies for testing a hypothesis and writing it up as data, code and a paper, shown in Figure 3. The single-prompt strategy gives one agent the whole task. Fixed sequential execution splits it across three agents running one stage each, with no feedback. Iterative planning has agents propose strategies, pick the best, execute it and feed the results and what is still missing into the next round. Each strategy ran on two machine learning hypotheses, giving six runs.

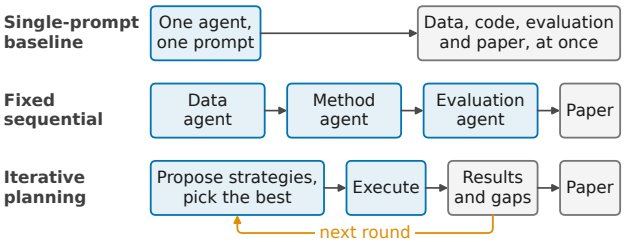


Figure 3: The three strategies for testing a hypothesis and writing it up as data, code and a paper. Only iterative planning feeds results back into the next round.

We screened candidate judges on self-consistency, position bias and cross-model agreement, without a human expert. Claude Sonnet 4.5 and GPT-5.2 passed and were used to rank the six runs by pairwise comparison [3], seeing only the paper, only the code, or both. The two judges agreed with each other on 89% of comparisons. A human expert independently assessed all six runs.[1]

Result. Figure 1d shows the pairwise comparison win rates of the judges. The judges ranked the three strategies six times: for each of the two hypotheses, once seeing only the paper, once only the code and once both. The expert put the iterative strategy first on both hypotheses, and the judges agreed with the expert in five of the six rankings. The exception was the first hypothesis judged on the paper alone, where the judges ranked the single-prompt strategy first. There the single-prompt strategy's paper claimed the hypothesis was confirmed by two small-scale tests. The expert found the claim unsupported by its code and preferred the iterative strategy, which ran 30 orderings and returned a negative finding. When the judges also saw the code they too preferred the iterative strategy.

4 Discussion and Limitations

Across the three studies, validation without a human expert did not replace the expert. Each automated evaluator passed its validation that required no human expert. Accuracy on past work measures whether the impact predictor has learned the patterns of published work. Self-consistency, position bias and agreement between judges measure whether the judges are stable. These checks are necessary, but not sufficient: only one of eight candidate judges returned the same verdict on every pair it saw twice. In each of the three studies, passing automated validation did not guarantee agreement with the human expert on generated R&D ideas.

[1]The six generated papers, their code and their outputs are available at tinyurl.com/24zx5bu7.

Limitations. There are several limitations to this work. Each study uses a single expert and a small sample: ten research directions, 40 hypotheses and six runs. The disagreement in each study could be specific to its expert, and ruling this out takes several experts per study, which is our next step. All three studies come from one pipeline and one research group. The claim is narrow: in three cases, passing validation without a human expert did not guarantee agreement with the expert.

Conclusion. We validated three automated evaluators of generated R&D ideas against a human expert: a citation-impact predictor on research directions, large language model judges on hypotheses, and language model judges on papers. Each passed its automated validation on every dimension, yet each diverged from the expert on part of what it ranked. The paper judges matched the expert on both hypotheses only once they saw the code. Recognising a good research idea involves something human experts do that current models do not. An organisation using such an evaluator to rank its R&D ideas should measure its agreement with a human expert before acting on it [22].

Acknowledgements

This research was financially supported by the Slovenian Research and Innovation Agency (ARIS) via the Gravity project AI for Science (GC-0001) and European HE projects DataPact (GA 101189771) and RAIDO (GA 101135800). The authors acknowledge the use of large language models to support paper organisation, drafting and figure preparation, and language and grammar refinement. All suggestions from the models were carefully reviewed and revised by the authors, who take full responsibility for the content of this paper.

References

[1] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: towards fully automated open-ended scientific discovery. arXiv: 2408.06292 [cs.AI].

[2] Juraj Gottweis et al. 2026. Accelerating scientific discovery with Co-Scientist. *Nature*, 655, 8122, 487–496. DOI: 10.1038/s41586-026-10644-y.

[3] Lianmin Zheng et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 46595–46623.

[4] Rimvydas Skyrius. 2021. *Business Intelligence: A Comprehensive Approach to Information Needs, Technologies and Culture. Progress in IS*. Springer. DOI: 10.1007/978-3-030-67032-0.

[5] Peiyi Wang et al. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. DOI: 10.18653/v1/2024.acl-long.511.

[6] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327. DOI: 10.18653/v1/2024.emnlp-main.474.

[7] Anna Bavaresco et al. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 238–255. DOI: 10.18653/v1/2025.acl-short.20.

[8] Adrian Mladenić Grobelnik. 2025. *Automating the Research Process with Agentic Large Language Models*. Bachelor's thesis. University of Ljubljana, Faculty of Computer and Information Science. Retrieved Sept. 5, 2026 from https://hdl.handle.net/20.500.12556/RUL-174624.

[9] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 279–299. DOI: 10.18653/v1/2024.acl-long.18.

[10] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *International Conference on Learning Representations*, 94003–94092.

[11] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science*, 342, 6157, 468–472. DOI: 10.1126/science.1240474.

[12] Xuemei Gu and Mario Krenn. 2025. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *Machine Learning: Science and Technology*, 6, 2, 025041. DOI: 10.1088/2632-2153/add6ef.

[13] Adilson Vital, Filipi N. Silva, Osvaldo N. Oliveira, and Diego R. Amancio. 2025. Predicting citation impact of research papers using GPT and other text embeddings. *Physica A: Statistical Mechanics and its Applications*, 674, 130789. DOI: 10.1016/j.physa.2025.130789.

[14] Mehran Nasseri, Patrick Brandtner, Robert Zimmermann, Taha Falatouri, Farzaneh Darbanian, and Tobechi Obinwanne. 2023. Applications of large language models (LLMs) in business analytics – exemplary use cases in data preparation tasks. In *HCI International 2023 – Late Breaking Papers* (Lecture Notes in Computer Science). Vol. 14059. Springer, 182–198. DOI: 10.1007/978-3-031-48057-7_12.

[15] Humam Kourani, Alessandro Berti, Daniel Schuster, and Wil M. P. van der Aalst. 2026. Evaluating large language models on business process modeling: framework, benchmark, and self-improvement analysis. *Software and Systems Modeling*, 25, 4, 1151–1186. DOI: 10.1007/s10270-025-01318-w.

[16] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv: 2205.01833 [cs.DL].

[17] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: robust and accurate AutoML for structured data. arXiv: 2003.06505 [stat.ML].

[18] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 18932–18943.

[19] Anthropic. 2025. Claude Sonnet 4.5 System Card. Retrieved Sept. 5, 2026 from https://www.anthropic.com/claude-sonnet-4-5-system-card.

[20] OpenAI. 2025. Update to GPT-5 System Card: GPT-5.2. Retrieved Sept. 5, 2026 from https://openai.com/index/gpt-5-system-card-update-gpt-5-2/.

[21] Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39, 3/4, 324–345. DOI: 10.2307/2334029.

[22] Matjaž Gams. 2025. The oath of researchers and developers. *Informatica*, 49, 1, 1–6. DOI: 10.31449/inf.v49i1.8149.