Scientometrics
https://doi.org/10.1007/s11192-026-05792-x

Generative AI bias against scientific novelty: a cautionary tale from a small-sample evaluation of research proposals

Diogo Machado[1,2]

Received: 30 April 2026 / Accepted: 20 August 2026

Abstract
Generative AI (GenAI) tools, and large language models (LLMs) in particular, are being rapidly adopted across the scientific research process, including in funding evaluation and peer review. While their efficiency gains are well documented, their systematic biases remain comparatively understudied. A central concern is whether LLMs can recognise and value scientific novelty, an essential property for the long-run productivity of science. This paper examines 139 monodisciplinary research proposals submitted in 2021 to the Austrian Science Fund (FWF) in mathematics, physics and astronomy, chemistry, and biology. Each proposal carries expert-assigned scores for novelty, scientific quality, team capacity, and feasibility, along with the actual binary funding decision. Using GPT-4o, we generate AI-based funding decisions through within-field pairwise comparisons aggregated into field-level rankings, simulating a competitive selection process. AI-based funding decisions matched human funding decisions in 60% of cases. Conditional logistic regression with field fixed effects shows that, among the four review dimensions, novelty is the only one significantly associated with disagreement. Conditional and multinomial logit analyses further show that novelty systematically predicts the direction of disagreement. Proposals rated as more novel by human reviewers are substantially more likely to be funded by humans and not by the LLM. The observed pattern is consistent with a conservative tendency in LLM-based evaluation, potentially reflecting the statistical logic of next-token prediction trained on past scientific outputs. While the design does not allow us to identify the underlying mechanism, the results raise concerns that LLM-assisted evaluation may under-select proposals that human reviewers identify as highly novel.

Keywords Generative AI · Large language models · Research evaluation · Peer review · Novelty · Scientific funding

JEL Classification O30 · O38 · C25 · I23

✉ Diogo Machado
diogo.machado@technopolis-group.com
https://ror.org/017m81f09
https://ror.org/05f950310

[1] Technopolis-Group, Brighton, UK

[2] Katholieke Universiteit Leuven, Leuven, Belgium

Introduction

Generative artificial intelligence (GenAI) is moving with unusual speed from a topic of speculation into the everyday infrastructure of scientific research. Within a span of a few years, large language models (LLMs) have become embedded in literature search, text editing, code generation, data extraction, and the drafting of scientific arguments. Recent quantitative evidence suggests that LLMs are now used by a substantial share of researchers across disciplines, with applications ranging from writing and editing to literature synthesis and exploratory analysis (Ding et al., 2025; Liao et al., 2025). The next frontier, already partly crossed, is the integration of LLMs into research evaluation itself. A growing number of funders and journals are piloting AI-assisted triage of grant proposals and manuscripts, while individual reviewers are reportedly turning to LLMs to help structure their assessments (Pusdekar, 2025).

This trend raises a foundational question for the scientometrics community: can LLMs evaluate scientific work in a way that is consistent with the goals of science policy? Efficiency, reproducibility, and reduced reviewer fatigue are obvious potential benefits. The risks, however, are less well charted. A key concern is whether LLMs can recognise and value scientific novelty, a property widely regarded as essential to the long-run productivity and diversity of science. Novelty is intrinsically distributional. Truly novel work, by definition, departs from the patterns that dominate the existing corpus. LLMs, in turn, are trained to predict the most likely continuation of a sequence given everything previously written. Because LLMs are trained on large corpora of existing text, it is plausible that their evaluations may be influenced by recurring patterns in previously successful scientific writing. Whether this potential tendency toward familiar patterns translates into a systematic evaluation bias, and whether such a bias materially affects funding outcomes, is an empirical question that is becoming difficult to defer.

Several recent contributions have begun to map these dynamics. In a qualitative study of leading U.S. scientists in materials science, Nelson et al. (2025) report that AI tools are deployed primarily for data analysis and predictive modelling, helping researchers tackle complex, variable-rich problems. While AI may help explore under-theorised areas by revealing patterns or "filling in the gaps", it is rarely seen by interviewees as contributing genuinely new theoretical insights. Several respondents in fact warned that over-reliance on AI could suppress theory-building, the very domain that has historically underpinned fundamental breakthroughs. Complementary evidence comes from grant submissions. Qian et al. (2026) and Kozlov (2026) document that proposals drafted or edited with AI assistance are more likely to win NIH funding, but that they tend to resemble previously successful submissions and ultimately produce fewer high-impact outcomes. While these effects may partly reflect improved clarity and alignment with reviewer expectations, they also point to a homogenising mechanism. By implicitly steering applicants toward established formats, AI tools may compress the diversity of how ideas are articulated.

A separate but related body of literature has examined whether LLMs can reproduce expert judgment in peer review and research evaluation settings. Large-scale REF-related studies provide some of the clearest evidence. Thelwall (2024) evaluates whether ChatGPT can assess journal-article quality using REF2021 scoring criteria and finds that the model can produce plausible rationales, but that individual scores correlate only weakly with human quality judgments. Thelwall and Yaghi (2025) use journal articles, submitted to REF2021 and compare ChatGPT scores with expert-derived departmental quality scores across fields, finding generally positive but field-varying

correlations. Kousha et al. (2026) similarly use REF2021 environment statements and show that ChatGPT can generate plausible assessments of research environments, with positive correlations with expert scores in most Units of Assessment. At the same time, this literature also highlights important limitations. Kousha and Thelwall (2026), using 99,277 REF2021-submitted journal articles, show that ChatGPT research-quality scores are associated with stylistic features such as linguistic complexity and abstract length, suggesting that LLM-based evaluation may partly reflect textual presentation as well as substantive quality. More directly related to grant evaluation, Sandström and Thelwall (2026) revisit the Wennerås and Wold peer-review data and find that LLM-generated grant proposal scores correlate positively but weakly with expert scores, suggesting some potential for triage or tie-breaking but limited suitability as stand-alone evaluators. Other studies report that LLMs produce reviews that overlap substantially with those of human reviewers, although typically with a tendency toward generic, surface-level critiques (Liang et al., 2024; Robertson, 2023).

What remains less well understood is whether disagreements between LLMs and humans, when they occur, are random or cluster systematically along specific dimensions of evaluation. From a science policy perspective, this distinction is critical. Random divergence would be costly but tractable. Systematic divergence associated with novelty would raise the possibility that LLM-based evaluation could compound conservative tendencies already present in human peer review (Boudreau et al., 2016; Stephan et al., 2017).

This paper addresses that question with a small-sample but information-dense empirical exercise. We use 139 monodisciplinary proposals, submitted in 2021 to the Austrian Science Fund (FWF), each with expert-assigned scores on four standard evaluation dimensions (novelty, scientific quality, team capacity, feasibility) and an actual binary funding decision. We construct an LLM-based counterfactual funding decision by asking GPT-4o, in pairwise comparisons within each scientific field, to choose which of two proposals should be funded. The pairwise outputs are aggregated into a field-level ranking, and the top N proposals in each field are designated as "AI-funded", where N is the number actually funded by FWF. The two sets of decisions are then compared at the proposal level, and the determinants of agreement and disagreement are modelled as a function of the four expert-assigned dimension scores using conditional logistic regression with field fixed effects.

We make three contributions. First, we provide empirical evidence that LLM-based evaluation does not diverge from human peer review at random. Rather, divergence is systematically associated with the novelty dimension, with proposals that human reviewers rated as more novel are more likely to be funded by humans but not by the LLM. Second, by separating disagreement (whether the two systems differ) from direction (which way they differ), we show that this difference reflects an asymmetric pattern of disagreement around novelty. Third, we discuss implications for the design of LLM-assisted evaluation tools that are central to the LLM4SCIM agenda, arguing that the standard frame of “LLM as efficient reviewer” misses the more consequential question of “LLM as non-neutral selector”. Given the restricted but unusually detailed dataset, the study should be read as an exploratory contribution. It identifies a structured pattern of disagreement within a specific funding context and model configuration, rather than providing a definitive demonstration of generalised LLM conservatism. The remainder of the paper is structured as follows. Sect. "Conceptual background" develops the conceptual background and motivates the empirical strategy. Sect. "Data and methodology" describes the data and methodology. Sect. "Results" presents the empirical results. Sect. "Discussion" discusses implications, limitations, and an agenda for future work, and Sect. "Conclusion" concludes.

Conceptual background

Why novelty is hard for statistical learners

Most LLMs in current use are autoregressive models trained to minimise next-token prediction error on very large corpora. Their parameters encode, at scale, the joint distribution of words, ideas, and rhetorical moves that have co-occurred in the training data. As a consequence, the texts they produce, the judgments they articulate, and the comparisons they make are systematically pulled toward configurations that are statistically common in the training corpus. This is not a bug. It is the optimisation target. But it has well-known consequences for the tails of distributions. Rare, unusual, or unprecedented combinations of ideas are precisely those for which the model has the weakest signal and toward which it has the least pull (Bender et al., 2021; Birhane et al., 2022).

In scientific evaluation, this translates into a recognisable preference for proposals whose framing, methods, and conceptual structure resemble previously successful work. The most cited illustration of this logic is AlphaFold (Jumper et al., 2021), which transformed protein structure prediction not by exploring speculative folding mechanisms but by identifying configurations with the highest expected accuracy, given the configurations of previously solved structures. The system's remarkable performance is grounded in a tight statistical coupling to the past. When such a logic is generalised to research evaluation, the natural consequence is a preference for proposals that look like winners, rather than for proposals that look unlike anything that has won before.

This is not the only relevant pressure. Human peer review is itself well known to penalise unconventional work, both in funding and in publication outcomes (Boudreau et al., 2016; Lane et al., 2022; Stephan et al., 2017). Reviewers face cognitive and reputational costs in endorsing ideas that depart from established consensus. The expected utility of a "safe" proposal often dominates that of a high-variance bet, especially in regimes of limited budgets and short evaluation cycles. The relevant question, therefore, is not whether LLMs have a bias against novelty in some absolute sense, but whether they introduce a bias on top of, and in the same direction as, the conservative tendencies already present in human review. If so, the marginal effect of substituting or augmenting human review with LLM review is not neutral. It compounds.

This study examines research proposals, not completed scientific outputs and novelty at the proposal stage should be distinguished from novelty in eventual research findings. Incremental projects may ultimately produce highly impactful novel results, even if their initial framing, methods, and expected contributions remain close to established lines of work. The argument developed here therefore applies most directly to proposals whose novelty is visible ex ante, that is, those that combine ideas, methods, or questions in unusual ways, depart from established disciplinary conventions, or embody a higher-risk, higher-reward profile. It should not be read as implying that all forms of innovation, including incremental advances that later prove consequential, are equally difficult for LLM-based evaluation to recognise.

From scoring to selection: the role of forced choice

A pragmatic complication for empirical work is that when asked to score proposals on absolute scales, LLMs tend to produce highly compressed distributions. Most proposals receive similar high scores, regardless of dimension. This is a familiar pathology of LLM

evaluation more broadly (Liang et al., 2024) and reflects, in part, the polite, calibrated tone that contemporary chat-tuned models default to. For ranking and selection, however, score compression is fatal. A homogeneous score distribution provides no signal for differentiating proposals.

Pairwise comparison is a standard remedy. By asking the model to choose between two specific proposals, we recast the task as a relative judgment. Pairwise comparisons impose a structural forced choice that prevents tied scores and produces directly informative ordinal data. Aggregating pairwise choices into a global ranking is itself a classical problem in social choice and bibliometrics (Bradley & Terry, 1952; Saaty, 1980). Several aggregation approaches are possible. In the main analysis, we use a simple Borda-style win count, described in detail in Sect. "Data and methodology", because it is transparent and directly interpretable for the scale of our dataset.

Disagreement vs direction: two empirical questions

Once an LLM-based funding decision is constructed for each proposal, the relevant inferential targets become more precise. Two distinct questions need to be separated. First, is the probability that the LLM and human decisions disagree associated with proposal characteristics, in particular with the novelty score? Second, conditional on disagreement, is the direction of divergence (human-funds-not-AI vs AI-funds-not-human) systematically related to those same characteristics? The first question speaks to the magnitude of the bias; the second speaks to its sign. A model that disagrees randomly, even at non-trivial rates, may still produce unbiased selections in expectation. A model that disagrees systematically with respect to novelty, in a consistent direction, by contrast, would represent a meaningful threat to the use of LLMs in research evaluation.

Data and methodology

Data

We use a dataset of 139 research proposals, submitted in 2021 to the FWF, the Austrian Science Fund. Of these, 51 were funded. The 139 proposals represent all submissions in 2021 that were fully classified by FWF's field assignment system into one of four non-overlapping scientific fields, consisting of mathematics, physics and astronomy, chemistry, and biology. Interdisciplinary proposals were excluded to ensure clear within-field comparisons and to control for differences in evaluation cultures across disciplines.

For each proposal, we observe: (i) the full proposal text, submitted to FWF; (ii) human peer-review scores along the four official FWF evaluation dimensions, namely novelty/originality, scientific quality, capacity of the research team, and feasibility of the proposed approach; and (iii) the actual binary funding decision. These four dimensions are used by FWF reviewers under the official evaluation guidelines and are described in more detail in Sect. "Empirical strategy", where they serve as the explanatory variables of interest in the regression analysis.

In the 2021 FWF review procedure, external reviewers provided both written assessments and formal ratings for the programme's specific evaluation questions. According to FWF's decision-making principles, reviewers were asked to address specific questions regarding the proposal and to provide a formal assessment for each question on a five-point scale. In the

administrative data, this scale is coded as 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Average, and 5 = Poor. For ease of interpretation in the empirical analysis, we invert the original scale so that higher values always indicate a more favourable assessment: 1 = Poor, 2 = Average, 3 = Good, 4 = Very Good, and 5 = Excellent.

FWF's description links the five rating categories to substantive standards and explicitly anchors the assessment in the proposal's field of research. The scores are applied separately to each evaluation dimension. A proposal can therefore be rated, for example, as excellent in novelty but only good in feasibility or team capacity. The field-relative interpretation applies to these dimension-specific judgements rather than only to an overall proposal score. An "Excellent" rating indicates that, on the assessed dimension, the proposal is among the best 5% in its field worldwide. A "Very Good" rating indicates that it is among the best 15% in its field worldwide and at the forefront of the research area internationally, though with minor potential improvements. A "Good" rating indicates international competitiveness but some weaknesses; an "Average" rating indicates that the proposal is expected to provide some new insights but has significant weaknesses; and a "Poor" rating corresponds to a weak assessment and rejection. This field-relative wording is important for our design because it indicates that reviewers were not asked to compare proposals against a single cross-disciplinary standard, but to evaluate each dimension relative to international standards within the relevant field.

The novelty/originality score is therefore an expert-assigned evaluative judgement embedded in FWF's standard review rubric, rather than an algorithmic or bibliometric measure. It captures the extent to which a proposal is judged to be innovative and to make an original contribution to its field, rather than merely extending established lines of work. The remaining dimensions capture scientific quality, the research team's capacity, and the feasibility of the proposed approach.

An important advantage of the FWF scoring structure is that reviewers were asked to assign separate scores to each dimension rather than to collapse their assessment into a single holistic judgement. This reduces, though does not eliminate, the risk that novelty/originality is absorbed into the overall impression of the proposal's quality. It also allows us to examine the association between novelty/originality and funding outcomes while controlling for the other dimensions of evaluation. In the regression analysis, we therefore estimate models that include novelty/originality alongside scientific quality, team capacity, and feasibility, which allows us to assess whether novelty/originality is associated with the outcome after accounting for the other official review criteria.

Because such judgements may still partly reflect field-specific evaluation cultures, levels of epistemic risk, and disciplinary norms about what counts as original, we restrict the sample to monodisciplinary proposals and conduct all comparisons within each field. Although FWF's scale is field-relative, we do not assume that scores are perfectly interchangeable across disciplines. We therefore include field fixed effects in all regression models. The number of proposals per field, after the monodisciplinary filter, ranges from approximately 30 to 40, which is sufficient for within-field pairwise comparison while constraining the precision of regression estimates. We treat the small sample explicitly as a feature of the design, not a bug, and discuss its implications in Sect. "Discussion".

LLM evaluation pipeline

Why pairwise, not absolute

We initially prompted GPT-4o, with the temperature parameter set to zero to ensure deterministic outputs, to assign a single overall evaluation score to each proposal. The prompt was constructed to mirror the official FWF reviewer guidelines, including the four evaluation criteria and the scoring logic. This direct-scoring design produced highly compressed score distributions, with most proposals receiving similar high scores and limited differentiation across submissions. The signal-to-noise ratio was insufficient to support ranking-based selection.

We therefore moved to a pairwise comparison framework. Within each field, proposals were grouped into pairs, and the model was asked to choose which of the two should be funded for each pair, being reframed as a forced binary choice. This is functionally equivalent to a tournament between proposals within a field. Crucially, the human peer review scores on the four dimensions were not provided to the model. They were retained as held-out variables and used ex post as explanatory variables in the analysis.

Pairwise comparisons and ranking

Every proposal was compared pairwise against every other proposal assigned to the same monodisciplinary field. The tournament was therefore exhaustive. No cross-field comparisons were made, and no within-field pairs were sampled or omitted. Each pair was presented to the model in randomised order to mitigate position effects, and the same prompt template was used across all comparisons. The model was forced to return a single choice.

The resulting tournament provides a complete within-field preference matrix, in which each proposal's ranking position is based on its performance against all other proposals in the same field rather than against a subset of randomly selected competitors. This design was chosen to avoid sampling variability in constructing the LLM-based ranking and to keep the comparison aligned with the field-specific structure of the human funding process.

The set of binary outcomes was then aggregated into a field-specific ranking using a Borda-style win count. Each proposal's overall score was the number of pairwise comparisons it won within its field. The Borda-style win count provides a simple, transparent, and directly interpretable aggregation of the pairwise outcomes. Robustness checks using Bradley-Terry and Elo-style aggregation produced highly similar rankings, so the main analysis retains the Borda-style win count as the primary specification.[1]

To simulate a competitive funding decision, we then selected, in each field, the top N proposals according to this ranking, where N is the actual number of proposals that FWF funded in that field. This procedure preserves the field-specific selectivity rate of the human process and ensures that the LLM "funds" exactly the same number of proposals as the human reviewers, both overall and within each field. Differences between the two sets of funding decisions are therefore not driven by differences in budget or selectivity, but by differences in which specific proposals are selected.

[1] The Borda-style rankings were compared with Bradley-Terry and Elo-style rankings estimated from the same exhaustive pairwise comparison matrices. The three approaches produce highly similar rankings: approximately 70% of the proposals differ by no more than one rank, 30% have identical ranks across all methods and pairwise rank correlations are high (Spearman ≥ 0.97).

The prompt

Each pairwise comparison was structured as a two-message exchange with the model, comprising a system message that fixed the role and the task and a user message that delivered the two proposals to be compared. The system message asked the model to act as a scientific expert evaluating two research proposals and to decide which of the two should be funded, without specifying which evaluation criteria to emphasise. The user message then provided the full text of both proposals, labelled as Proposal A and Proposal B, and required a single binary answer (A or B) without any accompanying justification, free text, or qualifying language. The same template was used for every pair within every field. The simplified structure of the exchange is shown below.

System message:

"You are a scientific expert evaluating two research proposals, submitted to a competitive funding agency. Your task is to read both proposals carefully and decide which one should be funded. You must choose exactly one of the two proposals. Respond with a single letter, A or B, and nothing else."

User message:

"Proposal A:
[full text of the first proposal]
Proposal B:
[full text of the second proposal]
Which proposal should be funded? Answer with A or B only."

The forced single-letter answer was a deliberate design choice. Free-text rationales tend to elicit hedged language from chat-tuned models and reduce the discriminative power of the comparison. By restricting the answer space to two, the prompt maximises the information content of each comparison while keeping outputs trivially parseable. Importantly, the prompt itself is intentionally general. It instructs the model to act as a scientific expert and to choose between the two proposals, but it does not enumerate a preferred evaluation criterion. This avoids steering the model toward any particular dimension of merit through prompt design and instead lets it apply whatever implicit notion of scientific merit it has internalised from training. The four FWF-assigned dimension scores are therefore not used to construct or condition the model's choice. They are held out and used only as explanatory variables in the regression analysis described next.

The model was not asked to provide an explanation or justification for its choices. This was a deliberate design choice. First, the main objective was to construct a clean counterfactual selection outcome rather than to analyse the model's verbal rationales. Requiring a single-letter response kept the output directly comparable across all pairwise contests and avoided additional variation introduced by differences in explanation length and style. Second, the proposal texts are highly confidential. Asking the model to generate open-ended explanations would have created practical and ethical difficulties for reporting and

interpreting the results, since such explanations would reveal proposal-specific content. For this reason, the study focuses on the structure of the model's choices rather than on its stated reasons for those choices.

Empirical strategy

We assess three outcomes of interest. The first is a binary indicator of disagreement, taking the value one if the LLM-based decision differs from the human funding decision and zero otherwise. The second, defined only on the subset of disagreement cases, is a binary indicator of direction, taking the value one if humans funded the proposal but the LLM did not, and zero if the LLM funded the proposal but humans did not. The third is a three-category variable defined on the full sample, distinguishing agreement, human-funds-not-LLM, and LLM-funds-not-human cases.

For the first two outcomes, we estimate conditional logistic regressions with field-level fixed effects. Conditional logit is appropriate because it allows us to compare proposals only with others in the same field, while controlling non-parametrically for field-specific differences in scoring practices, evaluation cultures, or selectivity. For the three-category outcome, we estimate a multinomial logit model on the full sample, using agreement as the reference category and including field indicators. This complementary specification allows us to test whether novelty is associated differently with the two directions of disagreement relative to cases where human and LLM decisions agree.

The regressors are the four FWF-assigned dimension scores that human reviewers used to evaluate the proposals. These four dimensions form the official FWF rubric and capture the criteria that human reviewers were instructed to weigh when reaching their funding recommendations. Specifically:

- Novelty/originality. The extent to which the proposal advances ideas, methods, or questions that depart from established approaches and address open problems. This dimension captures whether the proposal introduces something genuinely new or is an incremental refinement of existing work.
- Scientific quality. The technical and conceptual soundness of the proposed research, including the clarity of the research questions, the rigour of the methodology, the alignment between aims and means, and the appropriateness of the analytical framework.
- Team capacity. The qualifications, prior track record, and institutional setting of the principal investigator and the research team, including their demonstrated ability to carry out the proposed work.
- Feasibility of the approach. The realism of the work plan, the adequacy of the proposed resources and timeline, and the likelihood that the proposed methods can be successfully executed within the project envelope.

These four scores are the dimensions that human reviewers applied to the proposals in this dataset. Using them as explanatory variables turns the regression into a direct test of whether any of the funding agency's criteria are systematically associated with disagreement between human and LLM funding decisions, and, if so, in which direction. Crucially, these scores were not provided to the LLM at any point. The model produced its pairwise choices based solely on the proposal text, under a deliberately general prompt, while the human-assigned scores were held out and used only by the analyst to examine the structure of agreement and disagreement ex post.

We estimate four single-dimension specifications, in which each of the four dimensions enters individually, and one combined specification in which all four enter jointly. This allows us to inspect both marginal associations (does any single criterion predict disagreement on its own?) and partial associations (does any criterion predict disagreement conditional on the others?). Single-dimension specifications are informative because the four dimensions are correlated within proposals. In the joint specification, partial coefficients capture the residual signal of each dimension once the others are absorbed, which is the more demanding test.

Formally, for proposal i in field f, the conditional logit specification for disagreement is:

$$Pr\left(Disagree_i = 1 | field f\right) = exp\left(X_i'\beta\right)/\Sigma_j \in f\, exp\left(X_j'\beta\right)$$

where X_i is the vector of dimension scores. The same specification is estimated on the disagreement subsample for the direction outcome. Standard errors are robust. We report odds ratios (OR) throughout to facilitate interpretation. We also report p-values associated with two-sided tests of the null that each coefficient equals zero.

Results

Overall agreement

At the proposal level, the LLM-based and human funding decisions coincide in 87 of 139 cases, an agreement rate of 62.6%. By construction, this corresponds to 52 cases of disagreement. The agreement rate exceeds what would be expected from randomly selecting 51 proposals out of 139 within each field, confirming that the LLM is far from a random selector. At the same time, the residual disagreement is substantial enough to make the question of its structure non-trivial. The relevant policy concern is not whether the LLM agrees with humans on average, but whether it disagrees systematically and in policy-relevant ways.

Predicting disagreement

We ran conditional logistic regression estimates for disagreement between LLM and human funding decisions. The estimates include four single-dimension specifications, in which each of the four evaluation dimensions enters separately, and a joint specification including all four dimensions simultaneously.

The results are largely consistent. Novelty is the only evaluation dimension that systematically predicts disagreement between the two evaluation systems. In the single-dimension specification, a one-point increase in the human-assessed novelty score is associated with a 2.64-fold increase in the odds that the LLM and human reviewers reach different funding decisions (OR = 2.64, $p < 0.01$). After controlling simultaneously for scientific quality, team capacity, and feasibility, the estimated association becomes even larger (OR = 4.80, $p < 0.05$). By contrast, none of the remaining evaluation dimensions exhibits a statistically significant association with disagreement. Scientific quality, team capacity, and feasibility all have odds ratios close to one in the joint specification, and their confidence intervals include the null value. Whether considered in isolation or conditional on the other evaluation criteria, proposals judged by human reviewers to be more novel are substantially more likely to produce disagreement between human and LLM funding decisions. This suggests

that divergence is concentrated on a single evaluation dimension rather than reflecting general differences in how proposals are assessed (Table 1).

Figure 1 summarises the conditional logit estimates for the disagreement outcome, reporting both the single-dimension specifications and the joint specification in which all four dimensions enter together. The visual evidence makes the asymmetry across dimensions clear. The point estimate for novelty lies clearly to the right of zero in both specifications, and the confidence intervals exclude zero. For the other three dimensions, point estimates are smaller in magnitude, sometimes change sign across specifications, and have confidence intervals that comfortably include zero.

Direction of disagreement

The previous section established that novelty is the only proposal characteristic systematically associated with disagreement between human and LLM funding decisions. We now examine whether novelty also predicts the direction of that disagreement. Specifically, we distinguish between cases in which humans funded a proposal that the LLM rejected ("Human funds, LLM does not") and cases in which the LLM funded a proposal that humans rejected ("LLM funds, Human does not"). To do so, we combine two complementary analyses. First, we estimate a conditional logit model on the 52 disagreement cases only. Second, we estimate a multinomial logit model on the full sample of 139 proposals, treating agreement as the reference outcome and testing whether novelty has differential effects on the two disagreement categories. Table 2 reports the results.

The conditional logit model confirms that, among disagreement cases, novelty strongly predicts which side prevails. A one-point increase in the human-assigned novelty score is associated with an 8.8-fold increase in the odds that the proposal is funded by humans rather than by the LLM (OR = 8.82, $p = 0.007$). The multinomial logit yields a consistent picture. Relative to agreement, novelty increases the likelihood

Table 1 Conditional logistic regressions predicting disagreement between human and LLM funding decisions

Predictor	Novelty only	Quality only	Team only	Feasibility only	All dimensions
Novelty/originality	2.64***				4.8***
	(0.91)				(2.68)
Scientific quality		1.63			0.62
		(0.52)			(0.34)
Team capacity			1.37		0.68
			(0.53)		(0.35)
Feasibility				1.5	0.88
				(0.47)	(0.39)
Observations	139	139	139	139	139
Field fixed effects	Yes	Yes	Yes	Yes	Yes

The table reports odds ratios (OR) from conditional logistic regressions. The dependent variable is an indicator equal to one when the human and LLM funding decisions differ. All models are estimated on the monodisciplinary proposal subset and include field fixed effects through conditioning on field of science. Standard errors are reported in parentheses

$p < 0.10$, p < 0.05, $p < 0.01$

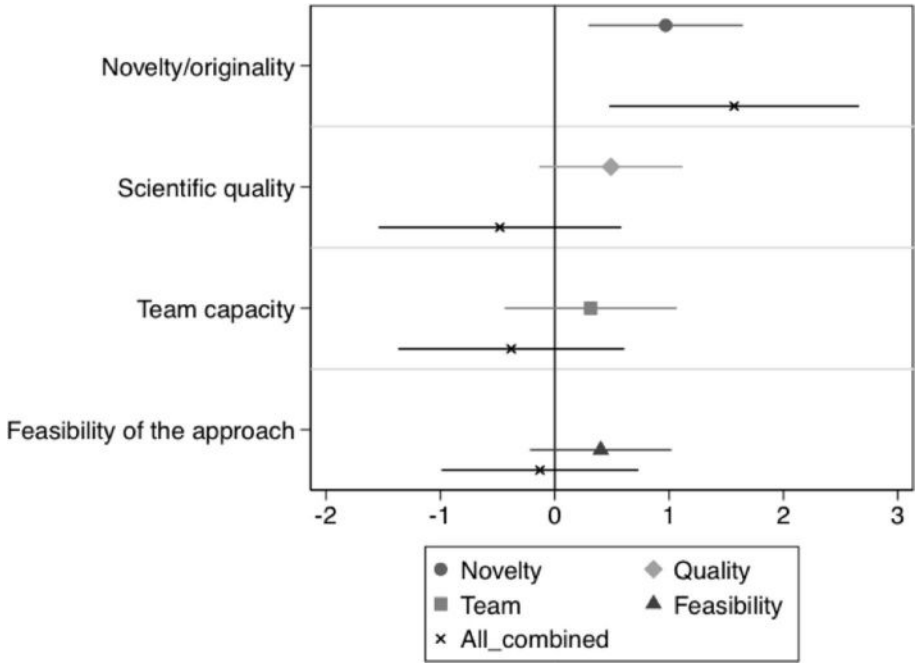

Fig. 1 Coefficient plot for the conditional logit model of disagreement between LLM-based and human funding decisions. Each marker reports the estimated log-odds (centred on zero) for the corresponding evaluation dimension, with horizontal bars showing 95% confidence intervals. Single-dimension specifications are shown by the shaded markers (circle: novelty; diamond: scientific quality; square: team capacity; triangle: feasibility). The "All combined" specification is shown by the cross. Field fixed effects are included in all specifications. Estimates are based on 139 proposals submitted to the Austrian Science Fund in 2021. Novelty is the only dimension whose confidence interval excludes zero, both in the single-dimension and in the all-combined specification

Table 2 Novelty predicts both disagreement and the direction of disagreement between human and LLM funding decisions

Panel	Model	Outcome/comparison	N	Estimate (OR)	Std. error	*p*-value
A	Conditional logit	Disagreement: Human ≠ LLM	52	8.82***	7.06	0.007
B	Multinomial logit	Human funds, LLM does not vs agreement	139	9.76***	6.13	<0.001
		LLM funds, Human does not vs agreement		1.38	0.52	0.396
		Wald test: βHuman-only = βLLM-only		$\chi 2 = 8.32$	—	0.0039

Panel A reports odds ratios (OR) from a conditional logit model estimated on the subset of disagreement cases. Panel B reports relative risk ratios (RRR) from a multinomial logit model estimated on the full sample, with agreement as the reference outcome. All estimates refer to a one-point increase in the human-assigned novelty/originality score and include field-of-science fixed effects. Standard errors correspond to the exponentiated estimates reported by Stata (eform). $p < 0.10$, $p < 0.05$, $p < 0.01$

of the "Human funds, LLM does not" outcome by almost a factor of ten (RRR = 9.76, $p < 0.001$), while showing no statistically significant association with the opposite outcome (RRR = 1.38, $p = 0.396$). A Wald test rejects equality of the two coefficients ($\chi^2 = 8.32$, $p = 0.0039$), providing further evidence that novelty shifts disagreement systematically in one direction rather than the other.

Figure 2 makes the asymmetry visually concrete. In the lower-novelty bucket (novelty< 4), the disagreement cases are dominated by LLM-funds-not-human outcomes. When proposals are rated lower in novelty by human reviewers, the LLM is more likely to select them nonetheless. In the medium-novelty bucket ($4 \leq$ novelty < 5), disagreement is more balanced. In the high-novelty bucket (novelty = 5), disagreement is overwhelmingly of the human-funds-not-LLM type. Among proposals rated with the maximum novelty score, the LLM is markedly less willing to fund them than human reviewers are. The pattern is monotonic in novelty and is consistent with conservative selection in this evaluation exercise. However, the figure should be interpreted descriptively; it does not, by itself, establish that the pattern is inherent in LLM architectures.

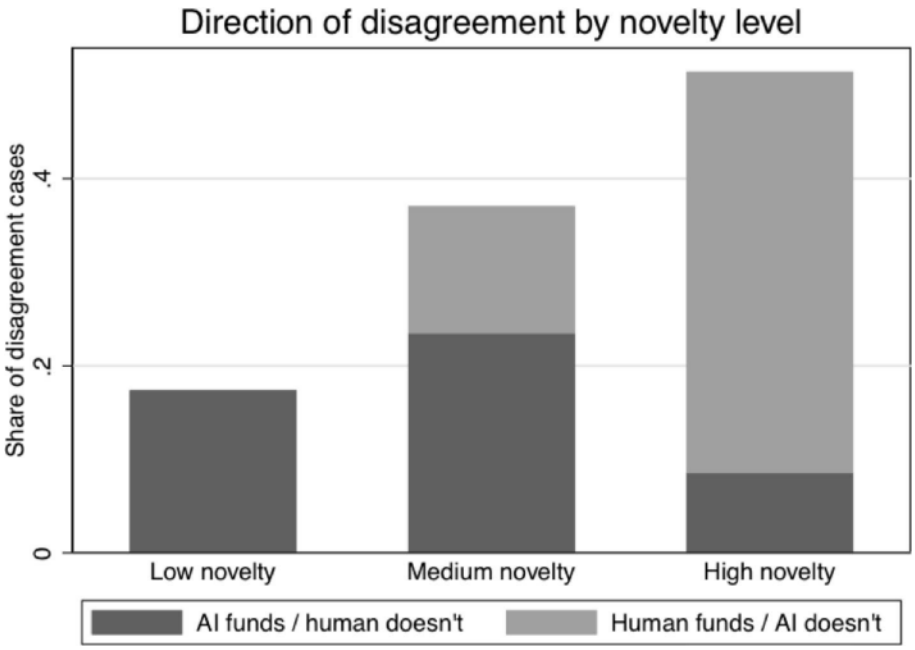

Fig. 2 Direction of disagreement by novelty level. The vertical axis reports the share of cases in each novelty bucket that fall into one of the two disagreement categories. The dark portion of each bar represents cases where the LLM funded the proposal, but humans did not; the light portion represents cases where humans funded the proposal, but the LLM did not. Novelty buckets are defined using the human-assigned novelty score: low novelty corresponds to novelty < 4, medium novelty to $4 \leq$ novelty < 5, and high novelty to novelty = 5. The figure shows that the share of human-funds-not-AI cases rises sharply with novelty, while the share of AI-funds-not-human cases is largest in the low-novelty bucket

Discussion

What the results imply

Three implications follow from the empirical pattern. First, the LLM-based selection process is not a noisy approximation to human peer review. It is a structured one. The 60% agreement rate is substantial, but the residual 40% of disagreement is concentrated on a single dimension, and that dimension is novelty. Second, the direction of disagreement is asymmetric. In this sample, higher human-assessed novelty is associated with a greater likelihood that a proposal is funded by humans but not by the LLM. The LLM does not simply have higher variance around the human assessment of novelty; it systematically pulled selection away from novel proposals. This pattern is consistent with a conservative selection tendency in this evaluation setting, although the design does not establish the causal mechanisms underlying the finding. Third, the magnitude is large enough to matter for funding outcomes.

Interpretation in light of LLM training logic

The pattern is consistent with the underlying training logic of contemporary LLMs. Models trained to predict the most likely next token, on a corpus dominated by previously published and previously successful science, will encode statistical regularities about what successful proposals look like. When asked to evaluate two proposals, they may favour the one whose framing, methods, and conceptual structure most closely resemble those regularities. Truly novel proposals, by definition, occupy the tails of the empirical distribution and thus enjoy weaker statistical support in the model's representation space. One possible interpretation is that the model's choices are influenced by such textual and conceptual familiarity. Because LLMs are trained on large bodies of existing scientific and scientific-adjacent text, proposals that resemble established forms of successful research may be easier for the model to recognise as fundable. Conversely, proposals that human reviewers rate as highly novel may depart from these familiar patterns. The present design cannot directly test this mechanism, but the observed association between novelty and asymmetric disagreement is compatible with it.

This interpretation aligns with parallel findings on AI-assisted grant writing (Kozlov, 2026; Qian et al., 2026), which document that AI-edited proposals are more likely to be funded but also more homogeneous. Both phenomena point to a homogenisation effect. AI tools, on both the writing and evaluation sides, can narrow the set of proposals that are plausibly successful. The risk is not that any single proposal is unfairly treated, but that the aggregate distribution of funded science is shifted toward the centre, and that the long tail of high-risk, high-reward work is systematically thinned.

Implications for the LLM4SCIM agenda

For the scientometrics community, and specifically for the LLM4SCIM agenda, the results contribute to several design questions. Evaluation studies of LLMs as reviewers should not be limited to overall agreement metrics. The relevant question is not only whether the LLM agrees with humans most of the time, but whether disagreements concentrate on policy-relevant dimensions. Standard evaluation practice currently underweights this distinction. The design of LLM-based evaluation tools should therefore explicitly account for potential

conservatism in their outputs. Possible mitigations include calibration on novelty-rich training corpora, prompt engineering that explicitly instructs the model to weight novelty, fine-tuning on labelled novelty assessments, and ensembling LLM scores with human or bibliometric novelty indicators (Lee et al., 2015; Uzzi et al., 2013). Transparency is also needed about which evaluation dimensions are most affected by the model. Otherwise, the appearance of efficiency gains may mask a structural shift in the composition of funded research.

An additional implication concerns validation. AI-assisted evaluation should be assessed not only against contemporaneous human judgments, but also against downstream indicators of scientific value. Earlier work on machine learning in grant review shows that project efficiency can be modelled algorithmically, while also illustrating the dependence of such models on measurable and imperfect proxies for scientific contribution (Sikimić & Radovanović, 2022). For LLM4SCIM, this suggests that the relevant validation question is not only whether LLMs reproduce panel decisions, but whether the proposals they favour generate the type of scientific outputs, novelty, diversity, and longer-term impact that funding systems are intended to support.

A further implication concerns auditability. If LLMs are incorporated into funding workflows, evaluation systems should make it possible to verify what was evaluated, under which model and rubric configuration, and with what output, without creating opportunities for strategic gaming. Bicakci (2026) proposes one such approach for auditable AI-assisted grant evaluation, based on verifiable links between the submission, model-and-rubric configuration, and evaluation output. Such procedural auditability would not by itself establish fairness or validity, but it would provide a necessary infrastructure for detecting, documenting, and contesting systematic biases such as the novelty effects identified here.

Limitations

Several limitations should be acknowledged. The sample is small, both in absolute terms and within each field, which limits the precision of the regression estimates and constrains the granularity of subgroup analyses. The analysis covers a single funder, the Austrian Science Fund (FWF), and a single submission year, 2021. Patterns of agreement and disagreement may therefore differ in other agencies, countries, disciplines, years, or panels with different evaluation cultures. The LLM used, GPT-4o with deterministic settings, is also only one model configuration. Results may differ across model families, model versions, training corpora, prompting strategies, and temperature settings.

The LLM evaluation pipeline also involves several design choices. The pairwise comparison procedure, while useful for avoiding compressed absolute scores, introduces decisions about pair construction, proposal ordering, forced-choice prompting, and ranking aggregation. These choices were held constant in the present analysis but not systematically varied. Future work could test whether the observed pattern is robust to alternative prompts, ranking methods, model settings, and aggregation procedures. Similarly, the four evaluation dimensions enter the regressions linearly. Given the discrete nature of the underlying review scores, alternative functional forms and non-linear specifications could be explored in larger samples.

A further caveat concerns the interpretation of novelty itself. Proposals rated as highly novel by human reviewers may differ systematically in other respects. Possible examples include readability, uncertainty, speculative framing, methodological risk, rhetorical style,

or distance from established disciplinary conventions. Some of these concerns are partly addressed in the empirical design. First, the regressions control for human-assessed scientific quality, team capacity, and feasibility. Second, comparisons are restricted to monodisciplinary proposals within the same field, reducing the possibility that the observed pattern is driven by broad disciplinary differences or by interdisciplinarity. Nevertheless, other unobserved conceptual features may still mediate or confound the relationship between novelty and disagreement.

Another design choice concerns the information provided to the LLM. The model was not asked to score proposals separately on the four FWF evaluation dimensions, nor were these dimensions explicitly included in the pairwise prompt. This was deliberate because the aim was to examine whether a general LLM-based selection process, based solely on the proposal texts, diverged from human funding decisions along dimensions that were held out from the model and used only ex post in the analysis. Including the four FWF dimensions in the prompt, especially novelty/originality, could have changed the estimand by explicitly steering the model toward the criteria later used to explain disagreement. At the same time, this choice limits what the study can say about rubric-guided LLM evaluation and may place the model at a disadvantage relative to human reviewers operating within the FWF evaluation framework. Future work should therefore compare the present text-only selection design with rubric-guided prompting, including prompts that ask the model to score novelty, quality, team capacity, and feasibility separately.

For these reasons, the results should be interpreted as evidence that disagreement between human and LLM-based funding decisions is systematically associated with human-assessed novelty, rather than as proof that novelty itself is the sole causal mechanism. The quantitative magnitudes and the underlying mechanisms should be interpreted with appropriate caution. These limitations motivate future replications with larger samples, multiple funders, alternative model families, richer measures of novelty and proposal characteristics, and alternative prompt designs.

Future research

Several extensions follow naturally. First, the analysis should be replicated with larger and more diverse samples, ideally spanning multiple funders, multiple years, and multiple disciplines. Second, the comparison across LLMs is a first-order priority. It is not currently known whether the bias documented here generalises across model families, or whether it is specific to a class of training regimes. Third, the analysis should be extended to alternative measures of novelty, including bibliometric novelty indicators based on combinations of cited references (Lee et al., 2015; Uzzi et al., 2013), to triangulate the human-assigned novelty scores. Fourth, intervention studies that explicitly aim to debias LLM evaluation through prompt design, fine-tuning, or retrieval augmentation would help assess the malleability of bias and inform the design of LLM-assisted evaluation tools.

Future work could also examine the model's stated reasons for its choices. In the present study, the LLM was not asked to provide open-ended justifications because the underlying proposal texts are highly confidential, making such explanations difficult to analyse and report without risking disclosure of proposal-specific content. In settings where confidentiality constraints are less severe, explanations could provide useful additional evidence about whether the observed selection patterns are reflected in the model's stated evaluation criteria. However, such explanations should not be treated as fully transparent evidence of the decision's underlying basis, since some decision patterns

and biases, including those related to novelty, may remain hidden or implicit. Finally, downstream impact analysis, linking AI-funded versus human-funded proposals to ex post bibliometric and policy outcomes, would help assess whether the bias documented here translates into measurable losses in scientific productivity.

Conclusion

LLMs are entering the scientific evaluation pipeline at speed. The natural framing of this transition has been efficiency, including faster reviews, lower reviewer burden, and more consistent rubrics. The findings of this paper suggest that a different framing is also necessary. In a small but information-dense sample of FWF research proposals, an LLM-based selection process agreed with human funding decisions in 60% of cases, but the residual disagreement was not random. It concentrated on the novelty dimension, and it concentrated in a specific direction. The LLM was systematically less willing than human reviewers to fund proposals rated highly for novelty. The pattern is consistent with the concern that LLM-based evaluation may favour proposals that resemble established forms of successful science over proposals that human reviewers judge to be highly novel. It also aligns with parallel evidence on the homogenising effects of AI-assisted grant writing. However, the present design does not identify the causal mechanisms underlying this pattern, and the findings should be interpreted as an empirical indication from a specific funding context, year, prompting strategy, and model configuration rather than as definitive evidence of generalised LLM conservatism.

For the LLM4SCIM agenda, the implication is that LLMs should not be evaluated only on their overall agreement with human reviewers. They should be evaluated on the structure of their disagreements, particularly along dimensions that matter for science policy and for the long-run productivity of science. For science policy, the implication is that the integration of LLMs into research evaluation should be accompanied by explicit safeguards, ranging from prompt design and fine-tuning to ensembling and human oversight, that protect the space of unconventional, exploratory, and high-uncertainty ideas. Without such safeguards, AI-assisted evaluation could risk narrowing the range of science that receives support, particularly if similar patterns are found in larger and more diverse settings. The technical path forward is open. The policy choice is not.

Acknowledgements The author would like to thank the Austrian Science Fund (FWF) for providing the proposal and peer-review data used in this study. The author also gratefully acknowledges the constructive comments of the two anonymous peer reviewers and the feedback received from participants at the 2026 European Forum for Studies of Policies for Research and Innovation (Eu-SPRI).

Author contributions D.M conceived and designed the study, collected and curated the data, developed the methodology, conducted the analysis, and wrote the manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The data used in this study were provided by the Austrian Science Fund (FWF) under a confidentiality agreement and consist of anonymised research proposal records. The raw proposal texts and reviewer scores are confidential and cannot be shared by the author. Aggregated data underlying the figures and regression tables, together with the analysis code, can be made available upon reasonable request and subject to FWF's data sharing policies.

Declarations

Competing interests The author has no relevant financial or non-financial interests to disclose. The author is employed by Technopolis Group, which has, in other engagements unrelated to this work, advised research funders on the use of evaluation tools. No funder of the present work has had any role in the design, analysis, or reporting of the study.

Ethical approval The data used in this study were provided by the Austrian Science Fund (FWF) under a confidentiality agreement and consist of anonymised research proposal records. No human participants were involved beyond the standard peer review process, which was conducted independently by the funder. Ethical approval was therefore not required under the relevant institutional guidelines.

Use of generative AI GPT-4o was used as the LLM under study to generate counterfactual evaluation decisions, as described in Sect. "Data and methodology". Additionally, generative AI was used for routine copy editing.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bicakci, K. (2026). Making AI-assisted grant evaluation auditable without exposing the model. *arXiv*. https://doi.org/10.48550/arXiv.2604.25200

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 173–184. https://doi.org/10.1145/3531146.3533083

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science, 62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*(3/4), 324–345. https://doi.org/10.2307/2334029

Ding, L., Lawson, C., & Shapira, P. (2025). Rise of generative artificial intelligence in science. *Scientometrics, 130*(9), 5093–5114. https://doi.org/10.1007/s11192-025-05334-4

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kousha, K., & Thelwall, M. (2026). Which abstract stylistic features fool ChatGPT research evaluations?. *Scientometrics*. https://doi.org/10.1007/s11192-026-05756-1

Kousha, K., Thelwall, M., & Gadd, E. (2026). Can ChatGPT evaluate research environments? Evidence from REF2021. *Scientometrics*. https://doi.org/10.1007/s11192-026-05633-x

Kozlov, M. (2026). Grant proposals drafted with AI help more likely to win NIH funding. *Nature*. https://doi.org/10.1038/d41586-026-00369-3

Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E. C., & Lakhani, K. R. (2022). Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation. *Management Science, 68*(6), 4478–4495. https://doi.org/10.1287/mnsc.2021.4107

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy, 44*(3), 684–697. https://doi.org/10.1016/j.respol.2014.10.007

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*. https://doi.org/10.1056/AIoa2400196

Liao, Z., Antoniak, M., Cheong, I., Cheng, E. Y. Y., Lee, A. H., Lo, K., Chang, J. C., & Zhang, A. X. (2025). LLMs as research tools: A large-scale survey of researchers' usage and perceptions. In *Proceedings of the Second Conference on Language Modeling (COLM 2025)*. https://openreview.net/forum?id=p0BwJk3R1p

Nelson, J. P., Olugbade, O., Shapira, P., & Biddle, J. B. (2025). Can Artificial Intelligence Accelerate Technological Progress? Researchers' Perspectives on AI in Manufacturing and Materials Science. *arXiv*. https://doi.org/10.48550/arXiv.2511.14007

Pusdekar, S. (2025). AI enters the grant game, picking winners. Science.

Qian, Y., Wen, Z., Furnas, A. C., Bai, Y., Shao, E., & Wang, D. (2026). The rise of large language models and the direction and impact of US federal research funding. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2601439123

Robertson, Z. (2023). GPT4 is slightly helpful for peer-review assistance: A pilot study. *arXiv*. https://doi.org/10.48550/arXiv.2307.05492

Saaty, T. L. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. McGraw-Hill.

Sandström, U., & Thelwall, M. (2026). Can large language models evaluate grant proposal quality? Revisiting the Wennerås and Wold peer review data. *arXiv*. https://doi.org/10.48550/arXiv.2603.14565

Sikimić, V., & Radovanović, S. (2022). Machine learning in scientific grant review: Algorithmically predicting project efficiency in high energy physics. *European Journal for Philosophy of Science, 12*, 50. https://doi.org/10.1007/s13194-022-00478-6

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature, 544*(7651), 411–412. https://doi.org/10.1038/544411a

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science, 9*(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M., & Yaghi, A. (2025). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. *Trends in Information Management, 13*(1), 1–29.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468–472. https://doi.org/10.1126/science.1240474