Should a national research funding agency use AI to triage grant proposals before human peer review?

THE VERDICT

No.

The Court answered the question in the negative.

Strongly supported

3 of 3 lenses · read from the Court record

WHAT TO DO

Refrain from deploying AI triage for grant proposals unless a domain-specific system with human oversight demonstrably matches or exceeds human triage in recall and fairness under controlled conditions.

THE WEIGHT OF THE JUDGES

17 of 18 judge votes against (94%)

Lens 1 · 6-0 · for this lens · supports No

Lens 2 · 0-6 · against

Lens 3 · 1-5 · against

Record 20 · Sealed on 16 November 2026

Fingerprint: cfd87ff7986ae726b55ac20478d24d1f8db9fc5d3a15711b5329265f390a3f94

REVISION 2 OF RECORD 19 · WHAT CHANGED AND WHY

Lens 1 asks whether AI triage would wrongly exclude good proposals, so its for side argued against triage and its 6-0 result supports the verdict No. The first file counted it as dissent. The brief is also written again: it gives each source its kind (the arXiv study is a preprint) and names organisations as the sources name them.

Before: Supported, with dissent, 2 of 3 lenses · read from the Court record · 11 of 18 judge votes against (61%)

Now: Strongly supported, 3 of 3 lenses · read from the Court record · 17 of 18 judge votes against (94%)

The original Record 19 stays as it was sealed, and its fingerprint still verifies: b40c5878fbd028326c581c20cdd44f817ece5cd8b8396321dddd9041adac51ae

The original file: https://fasil.quotefather.com/debate/52/result/record/19/pdf/

The decision's context, as the institution stated it

A national research funding agency receives a fast-growing number of grant applications, many of them now drafted with the help of generative AI tools, while the number of qualified reviewers has not grown at the same pace. The agency is considering adding an AI step before human peer review that would flag or screen out proposals judged weak or out of scope, so that reviewers can focus on the rest. The decision concerns only this pre-review triage stage, not final funding decisions. Any option must protect the confidentiality of applications, allow each decision to be explained to applicants, and treat all applicants fairly.

Why this verdict

The grounds

1. AI triage misses good proposals, prioritizes surface writing signals

The same sources are decisive in lenses 1 and 2. In one national funding programme of 959 proposals, an LLM screener achieved recall of only 45.82%, excluding more than half of proposals the committee later selected. Peer-reviewed research on LLM-based essay scoring finds that LLMs operationalize writing quality through surface linguistic signals rather than meaningful content.

Lens 1

2. Productivity gains do not transfer to merit review

In the citizen reimbursement claims initiative described by the source, reviewer productivity rose by 20.1%, but that programme is not a competitive research grant context, and no established evidence in the record supports extending that gain to merit-based screening.

Lens 2

3. AI triage gave scores without reasons, raised confidentiality and bias concerns

The same source is decisive in lenses 2 and 3. In the one documented operational deployment, reported by Science, applicants received only scores and no justifications. Applicants also feared their unpublished ideas had entered the AI systems, and separately questioned whether names, genders, or affiliations had influenced decisions.

Lens 3

The warrant

When an automated filter risks systematically excluding meritorious work before expert scrutiny, the burden of proof for equivalence must be met before adoption, not after harm occurs.

The backing

• Recall, Risk, and Governance in Automated Proposal Screening for Research Funding: Evidence from a National Funding Programme · arxiv.org (Institution · preprint)

• AI-assisted prescreening of biomedical research proposals: ethical considerations and the pilot case of "la Caixa" Foundation · cambridge.org (Institution · research)

• Governing AI-assisted evaluation in public innovation funding: where efficiency ends and judgement begins · technopolis-group.com (Institution · web)

• Leveraging LLMs to streamline the review of public funding applications · aclanthology.org (Institution · research)

• Responses to the AI grant flood must prioritize fairness as part of excellence · nature.com (Institution · press)

• Opening the blackbox of LLM-based automated essay scoring: Insights into feature weighting patterns and score validity - ScienceDirect · sciencedirect.com (research)

• U.K. funder uses ‘AI triage’ to reject grant proposals—leaving researchers dismayed | Science | AAAS · science.org (press)

What weighs against it

Human peer review already excludes meritorious proposals at scale

A registered report protocol finds that single rater intraclass correlations for NSF grant reviews ranged from 0.17 to 0.37, showing the incumbent process is already highly inconsistent. A well-governed AI triage layer that informs rather than determines decisions, with mandatory human verification of every exclusion, may add less incremental harm than the verdict assumes.

Holds unless

• a well-designed AI triage system with domain-specific keywords and human oversight is shown to achieve recall and fairness outcomes comparable to or better than unassisted human triage under controlled conditions.

• evidence shows that AI triage models found to be flawed are corrected promptly before applicants are harmed, with documented remediation rates or timelines in deployed screening contexts.

• a peer-reviewed study finds that non-native English speakers face no measurably higher false-negative rate from AI triage systems specifically, compared to human review.

Open questions the judges recorded

• It is unclear how the error rates of AI triage compare with the low agreement found among human reviewers. (Raised by 18 judges)

• No evidence shows whether confidence-based escalation reliably catches wrongful exclusions in grant screening. (Raised by 12 judges)

• It is unresolved whether audit logs or model changes actually lead to bias being fixed in practice. (Raised by 6 judges)

• Whether calibrated uncertainty quantification methods (ensemble disagreement, conformal prediction) can reliably flag misclassified grant proposals for human escalation, rather than assigning false confidence to wrongful rejections, is asserted by the case for but unsupported by any empirical evidence in the record. (Raised by 1 judge)

• No evidence in the record establishes what error-detection rate human reviewers achieve when operating a confirmation layer over AI triage outputs, leaving the central safety-net claim on both sides unresolved. (Raised by 1 judge)

• The correlation between aggregate LLM rankings and expert scores (WAME newsletter) is insufficient to establish that AI triage reliably identifies weak proposals at the level of individual decisions required for fair exclusion with explanation. (Raised by 1 judge)

The judges

Lens 1: Fairness and error: would AI triage wrongly exclude good proposals, or reward fluent writing and resemblance to past winners over the quality of the research?

6-0 · for this lens · supports No

Judge 1 · for

Strongest point · for: The la Caixa Foundation study explicitly warns that readability measures used in automated assessment could disadvantage non-native English speakers, whose texts are more likely to be perceived as badly written and rejected without thorough review, providing a documented, source-grounded fairness risk.

Weakest point · for: The inference that replacing human triage with AI makes exclusion 'permanent and invisible' is marked insufficient in the record and rests on an assumption about accountability that is not empirically grounded in the evidence; human triage decisions are also often unrecorded and unreviewable in practice.

Strongest point · against: The la Caixa Foundation study itself acknowledges that AI achieving human-level performance replicates human biases, which supports the case against's argument that the bias problem originates in the incumbent human review system rather than being uniquely introduced by AI triage.

Weakest point · against: The case against's central claim that AI triage produces 'logged, versioned, inspectable decisions' correctable across the entire applicant pool is rated insufficient in the record and contradicted by the CRANE deployment, where the AI systems used were not even disclosed; this claim is thus an unverified assertion rather than established evidence.

Judge 2 · for

Strongest point · for: The case for opposing AI triage correctly identifies that false negatives at the triage gate are irrecoverable: the arXiv study explicitly states that proposals excluded during initial screening do not re-enter the evaluation pipeline, making this category of error categorically worse than false positives that can be caught by subsequent human review.

Weakest point · for: The case for opposing AI triage relies partly on the CRANE deployment as the sole documented real-world case, but the record also contains evidence of successful AI-assisted review pilots (the Portuguese Environmental Fund and IExp initiatives) that achieved a 20% productivity gain with negligible false-positive rates, showing the worst case is not the only plausible case.

Strongest point · against: The case for AI triage correctly notes that human review is itself riddled with inconsistency: NSF inter-rater correlations of 0.17-0.37 and a 45-study meta-analytic correlation of 0.34 show the incumbent process already excludes meritorious proposals at scale, so the incremental downside of adding a well-governed AI triage layer may be smaller than claimed.

Weakest point · against: The case for AI triage argues that AI produces logged, inspectable, correctable decisions, but the only documented real-world deployment (CRANE, reported by Science/AAAS) did not disclose which AI systems were used and shared only scores without justifications, the very transparency and explainability benefits invoked remain undemonstrated in practice.

Judge 3 · for

Strongest point · for: The record includes a real deployment that failed the explainability and confidentiality constraints. The UK funder shared only scores, not justifications, did not disclose which AI systems it used, and applicants feared their unpublished ideas had entered the AI knowledge base.

Weakest point · for: The claims on non-native-writer bias and on how long flawed models persist are unverified. The supporting facts are speculative ('could disadvantage'), come from other domains such as essays, or rest on web pages.

Strongest point · against: Human review has real capacity limits and costs, and the opponents are right that the status quo is imperfect. The record shows NSF reviewing tens of thousands of proposals at hours per review, and notes that low reliability makes score differences near the funding cutoff hard to resolve.

Weakest point · against: The opponents' own cited evidence undercuts them: it says human-level AI performance may just replicate historical biases, and that bias in training data is likely to be reproduced and amplified. The inference that the AI only reproduces rather than adds harm is marked unverified.

Judge 4 · for

Strongest point · for: Several independent sources agree that AI struggles with fine-grained merit and tends to reproduce previously funded proposal types, which hurts novel work. The agency's own historical-review data would carry its biases forward.

Weakest point · for: The recall evidence comes from one study of 959 proposals in a single national programme, and it also shows a simpler, transparent method with much lower false-negative counts. The result therefore indicts a particular model rather than all triage.

Strongest point · against: Institutional guidance acknowledges that human review is influenced by bias and prestige. A well-governed system that keeps AI advisory could in principle add logging and consistency.

Weakest point · against: Unreliable human review does not show AI is better. The proponents' own source says AI that replicates human performance may scale up existing bias, which cuts against the opponents' view, and no evidence in the record shows AI error rates matching or beating human ones.

Judge 5 · for

Strongest point · for: The proponents' concerns about legitimacy are borne out by the Science report: the UK funder shared only AI scores, not the reasons for them, with rejected applicants and did not say which AI tools it used. That conflicts with the agency's duty to explain each decision.

Weakest point · for: The claim that non-native writers are harmed relies on a warning about a possible effect, not on a measured result in grant triage. A research study of grant applications found that non-native English writers do not face a language disadvantage under human review.

Strongest point · against: The opponents rightly point out that the la Caixa analysis traces AI bias back to biased historical human reviews, so the incumbent system is not neutral either.

Weakest point · against: The reliability evidence is mixed: a research study of NIH data reports reviewer agreement above 0.6. A noisy benchmark also does not make an even lower AI recall acceptable to the applicants wrongly screened out.

Judge 6 · for

Strongest point · for: The case for answers the reliability objection head-on. Its point is that copying biases already present in past human reviews, as the la Caixa Foundation paper warns, spreads the harm further rather than cancelling it out.

Weakest point · for: The claim that AI triage disadvantages non-native English writers rests on a caution in the la Caixa paper, not on measured results. A research study on writing style found that non-native writers did not have a language disadvantage in grant success.

Strongest point · against: The case against rightly points out that, by the la Caixa paper's own reasoning, any bias an AI copies started in past human reviews. So the current human system is not a neutral baseline.

Weakest point · against: The claims that AI triage produces logged, versioned decisions that can be corrected across the whole applicant pool have no supporting facts. The only real deployment discussed did not even disclose which models it used.

Lens 2: Capacity and speed: would AI triage relieve reviewer overload from the surge in applications without lowering the quality of funding decisions?

0-6

Judge 1 · against

Strongest point · for: A study summarised by the WAME newsletter (but the underlying finding is referenced across multiple sources) found that average LLM rankings correlate with expert consensus scores, suggesting AI can at least rank proposals in a direction consistent with expert judgment at the aggregate level.

Weakest point · for: The core claim that AI triage 'reliably' identifies weak submissions rests primarily on aggregate LLM-expert rank correlation from a web source (WAME newsletter), while the only study with individual-level data (arXiv preprint) shows the LLM system excluded more than half of proposals the committee ultimately approved, directly contradicting the reliability claim.

Strongest point · against: The la Caixa Foundation paper warns that an algorithm achieving human-level performance may simply be replicating biases already present in historical human reviews, and that biases related to gender and seniority in training data are likely to be reproduced and amplified in subsequent iterations, a documented structural risk, not a speculative one.

Weakest point · against: The opponents' claim that AI triage errors are 'irreversible' while human peer review errors are 'correctable through appeal' is asserted without verified evidence of appeal success rates or the actual proportion of AI-triage false negatives that would remain permanently uncorrected under a human-overseen architecture.

Judge 2 · against

Strongest point · for: A Portuguese public-funding pilot (the ReClaim and IExp initiatives) demonstrated a 20.1% increase in reviewer productivity and a reduction of more than two months in total evaluation time, with a negligible false-positive rate on the test set, showing that AI assistance can produce measurable workload relief in real administrative settings.

Weakest point · for: The case for relies centrally on the finding that LLM average rankings correlate with expert

scores, but the same body of evidence shows that individual LLM scores are inaccurate and that the LLM-based system excluded more than half of proposals ultimately approved by the selection committee, a catastrophic false-negative rate that the proponents do not resolve with verified countermeasures.

Strongest point · against: The la Caixa Foundation study warns that an algorithm achieving human-level performance may simply be replicating biases already present in historical reviews, and that readability measures could put non-English speakers at a systematic disadvantage; these biases are structural and amplify with each iteration, making the fairness harm both predictable and self-reinforcing.

Weakest point · against: The case against does not fully account for the fact that the status quo, unaided human peer review, itself produces documented low inter-rater reliability (intraclass correlations of 0.17-0.37 at NSF, 0.26 at the Austrian Science Fund), meaning the baseline against which AI triage is being judged is itself error-prone and not a clean alternative.

Judge 3 · against

Strongest point · for: A funder-sector report from Technopolis records a roughly 20% reviewer productivity gain with a negligible false-positive rate. That is a real efficiency precedent, though it concerns reimbursement-claim checking rather than merit triage.

Weakest point · for: The case for says mandatory verification of every exclusion is unnecessary, yet the La Caixa pilot it cites rests on human verification of every exclusion. Verification cost under that design is never quantified.

Strongest point · against: The case against uses data from the actual use case, proposal screening for a national funding programme, rather than analogies from other settings. It also shows a simpler transparent method did better than the LLM, so tool choice matters.

Weakest point · against: The claim that human verification of every exclusion cancels the efficiency gain is asserted without any measured verification cost. The inference is rated insufficient, and the 20% productivity precedent points the other way.

Judge 4 · against

Strongest point · for: The case for correctly identifies that the choice is not binary. Funder guidance accepts AI in a supporting role, and AI adds clear value on rule-based eligibility checks within controlled systems.

Weakest point · for: The case for says the La Caixa pilot used mandatory human verification of every exclusion, but the pilot material only says human oversight made it AI-assisted rather than AI-powered. The viability and efficiency conclusions are rated insufficient.

Strongest point · against: The record shows applicants in a UK funder's AI triage received only scores and no justifications. They also worried about confidentiality and hidden bias, which conflicts with the explainability and confidentiality requirements and could damage funder credibility.

Weakest point · against: The circularity argument about uncertainty flags and the claim that human verification erases efficiency gains are asserted, not evidenced, and are rated insufficient. The unverified claims on appeal and silent errors also add nothing.

Judge 5 · against

Strongest point · for: The proponents offer a specific design remedy for the linguistic bias documented in the la Caixa Foundation paper: drop readability metrics and audit outcomes across applicant groups. This takes the fairness concern seriously instead of ignoring it.

Weakest point · for: The proponents point to the la Caixa pilot's mandatory human verification of every exclusion as proof that the approach is viable. They then shift to tiered oversight that reviews only uncertain cases, which weakens the very accountability safeguard they had cited.

Strongest point · against: The opponents correctly point out that a correlation in average rankings can coexist

with serious errors on individual proposals. That matters for each excluded applicant even when the aggregate figures look acceptable.

Weakest point · against: The opponents' claim that human verification of exclusions would wipe out the efficiency gain is asserted rather than shown. Its supporting inference is rated insufficient.

Judge 6 · against

Strongest point · for: Some evidence in the record does show capacity gains. A Portuguese public-funding deployment raised reviewer productivity by about 20% with negligible false positives and cut total evaluation time by more than two months.

Weakest point · for: The Portuguese productivity gains came from reimbursement and eligibility checks. Technopolis says AI adds clear value in rule-based eligibility checks but struggles to tell apart proposals of similar merit, which is the judgement that quality triage needs.

Strongest point · against: The opponents correctly point out that population-level correlation can coexist with serious errors on individual proposals, which the newsletter source itself concedes.

Weakest point · against: The opponents claim that mandatory human verification wipes out all efficiency gains, but no evidence is offered. The Portuguese deployment recorded real productivity gains while humans stayed in the loop.

Lens 3: Trust and accountability: can a funder keep applications confidential and explain each exclusion to applicants if a model screened it first?

1-5

Judge 1 · against

Strongest point · for: The la Caixa Foundation paper (a peer-reviewed study) documents a pilot in which every AI-flagged exclusion was reviewed by a human officer before applicants were notified, demonstrating that a human-confirmation layer is at least operationally feasible in a controlled biomedical funding context.

Weakest point · for: The la Caixa pilot was a single foundation's controlled biomedical portfolio; no evidence in the record establishes that its human-confirmation protocol has been replicated at national scale across heterogeneous disciplines, leaving the scalability claim unverified.

Strongest point · against: The la Caixa Foundation paper warns that an AI achieving human-level performance may simply replicate historical biases (e.g. gender, seniority), and that readability measures could systematically disadvantage non-English speakers, both documented fairness risks grounded in peer-reviewed evidence.

Weakest point · against: The opponents' inference that human reviewers demonstrably fail to catch AI triage errors is asserted but not established: no fact in the record measures the actual error-detection rate of human reviewers operating in a human-confirmation protocol over AI triage outputs.

Judge 2 · against

Strongest point · for: The research on automated proposal screening shows that a well-designed TF-IDF system achieved 78.95% recall and produced only 68 false negatives from 323 committee-selected proposals, demonstrating that some automated screening tools can substantially reduce wrongful exclusions compared to LLM-based alternatives, supporting the case for a carefully calibrated triage step.

Weakest point · for: The UKRI episode, the only documented national-scale AI triage deployment in the record, saw applicants receive only AI-generated scores with no justifications at all, which is precisely the explainability failure the case for AI triage claims governance can prevent; this real-world precedent directly undermines the claim that institutional safeguards will function at operational scale.

Strongest point · against: The la Caixa Foundation paper warns that an AI achieving human-level assessment performance may simply be replicating biases already present in historical peer-review data, including gender and seniority biases, and that these are likely to be amplified in subsequent iterations, meaning AI triage risks systematically entrenching existing inequities at national scale.

Weakest point · against: The case against AI triage does not engage with the counterfactual downside of the status quo: the research record shows that human peer review itself has low inter-rater reliability (intraclass correlations of 0.17 to 0.37 at NSF) and is subject to cognitive bias and professional-network effects, meaning the alternative to AI triage is not a perfectly fair human process.

Judge 3 · against

Strongest point · for: AI looks clearly workable for rule-based formal eligibility checks in controlled administrative systems. A Portuguese pilot processed about 80,000 applications in under three weeks with a reported ~20% reviewer productivity gain, though that pilot was reimbursement claims rather than research merit.

Weakest point · for: The claim that aggregate LLM rankings reliably surface weak proposals rests on a low newsletter summary that itself says individual LLM scores are inaccurate. It gives no error rates, so it cannot show that exclusions are safe.

Strongest point · against: The record supports the concern that human checking is not a dependable backstop. Automation bias is acknowledged, and merit differentiation in the middle of the distribution is where AI is weakest and where decisions are actually made.

Weakest point · against: The claim that humans demonstrably fail to catch AI errors has no direct measurement of human catch rates. The step from benchmark false negatives to human-check failure is an unverified inference.

Judge 4 · against

Strongest point · for: The proponents' distinction between AI that informs a decision and AI that determines it fits a conditional design with human sign-off. Aggregate LLM rankings correlating with expert scores gives some basis for using AI to surface weak proposals.

Weakest point · for: The proponents' key claims, that human-confirmation protocols catch AI errors and that explainability compliance is measurable in audited deployments, are unverified. No fact in the record shows measured error-detection rates or applicant satisfaction.

Strongest point · against: Independent sources point to systematic fairness risks. Readability-based screening may disadvantage non-English speakers, historical reviewer bias may be amplified, and AI may favour polished writing, which would turn writing capacity into unequal access to funding.

Weakest point · against: The claim that human reviewers fail to catch AI errors at meaningful rates is unverified. The cited study measures model false negatives against a committee, not human failure to catch them, so the inference overreaches.

Judge 5 · against

Strongest point · for: The record contains workable ways to reduce harm to applicants: a transparent keyword-based screener had far fewer wrongful exclusions than a general-purpose language model, and the Research Council of Finland runs its own tools without putting applications into generative AI.

Weakest point · for: The case for leans on the WAME newsletter's claim that average language-model rankings track expert scores. That is a web source, and it concerns ranking, not confidentiality or reasons given to applicants.

Strongest point · against: Major public funders' practice backs the opponents' concern about confidentiality: the National Institutes of Health and the Australian Research Council ban generative AI in peer review, and the

Research Council of Finland bans putting application data into such tools.

Weakest point · against: The opponents misread the recall study, which compares automated screeners with the selection committee's decisions and does not test whether human checkers miss errors. The claim that human confirmation cannot catch mistakes therefore remains unverified.

Judge 6 · for

Strongest point · for: The national screening study found that a transparent keyword-based system outperformed a general-purpose large language model. This suggests interpretable triage is technically possible rather than ruled out.

Weakest point · for: The claim that aggregate model rankings reliably surface weak proposals rests on a web newsletter summary, so it is unverified. The la Caixa pilot was also a single-foundation biomedical programme, and nothing on record shows the same results at national scale.

Strongest point · against: The la Caixa paper itself names the pressure on the human in the loop to agree with the algorithm, which weakens the idea that human confirmation is enough. The NIH and the Australian Research Council have banned generative AI in peer review over confidentiality concerns.

Weakest point · against: The opponents' central point, that human checkers fail to catch triage errors at meaningful rates, rests on an unverified claim with no measured detection rates. The screening study compared automated tools with committees; it did not test humans confirming model output.

Sources

1. Recall, Risk, and Governance in Automated Proposal Screening for Research Funding: Evidence from a National Funding Programme (Institution · preprint)

arxiv.org

https://arxiv.org/html/2602.07869v1

2. Opening the blackbox of LLM-based automated essay scoring: Insights into feature weighting patterns and score validity - ScienceDirect (research)

sciencedirect.com

https://www.sciencedirect.com/science/article/pii/S2666920X26000305

3. Registered report protocol: Factors associated with inter-rater agreement in grant peer review (research)

pmc.ncbi.nlm.nih.gov

https://pmc.ncbi.nlm.nih.gov/articles/PMC12124745

4. The present and future of peer review: Ideas, interventions, and evidence (research)

pmc.ncbi.nlm.nih.gov

https://pmc.ncbi.nlm.nih.gov/articles/PMC11804526

5. AI-assisted prescreening of biomedical research proposals: ethical considerations and the pilot case of "la Caixa" Foundation (Institution · research)

cambridge.org

https://www.cambridge.org/core/services/aop-cambridge-core/content/view/D17CC86C884D4F9E3A6DD9ABF3F4605E/S2632324924000415a.pdf/ai-assisted-prescreening-of-biomedical-research-proposals-ethical-considerations-and-the-pilot-case-of-la-caixa-foundation.pdf

6. Governing AI-assisted evaluation in public innovation funding: where efficiency ends and judgement begins (Institution · web)

technopolis-group.com

https://technopolis-group.com/governing-ai-assisted-evaluation-in-public-innovation-funding-where-efficiency-ends-and-judgement-begins/

7. U.K. funder uses ‘AI triage’ to reject grant proposals—leaving researchers dismayed | Science | AAAS (press)

science.org

https://www.science.org/content/article/u-k-funder-uses-ai-triage-reject-grant-proposals-leaving-researchers-dismayed

8. Leveraging LLMs to streamline the review of public funding applications (Institution · research)

aclanthology.org

https://aclanthology.org/2025.emnlp-industry.143.pdf

9. Artificial Intelligence/Large Language Models (AI/LLM) Newsletter Compilation || WAME (web)

wame.org

https://wame.org/page2.php?id=151

10. Living guidelines on the responsible use of generative AI in research (Institution · official)

research-and-innovation.ec.europa.eu

https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf

11. Lost in the middle? examining scoring reliability and position bias in LLM-based automated essay scoring | Education and Information Technologies | Springer Nature Link (research)

link.springer.com

https://link.springer.com/article/10.1007/s10639-026-14019-8

12. Identifying Limitations and Bias in ChatGPT Essay Scores: Insights from Benchmark Data - The Learning Agency (web)

the-learning-agency.com

https://the-learning-agency.com/the-cutting-ed/article/identifying-limitations-and-bias-in-chatgpt-essay-scores-insights-from-benchmark-data

13. Nonstandard English and the Automated Scoring of Open-Ended Math Problems (research)

educationaldatamining.org

https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.long-papers.195/index.html

14. AI Shows Racial Bias When Grading Essays — and Can’t Tell Good Writing From Bad – The 74 (web)

the74million.org

https://www.the74million.org/article/ai-shows-racial-bias-when-grading-essays-and-cant-tell-good-writing-from-bad

15. Practices for Assessing Intellectual Merit and Other Impacts ... (official)

nsf-gov-resources.nsf.gov

https://nsf-gov-resources.nsf.gov/files/Broader-Impacts-Literature-Review.pdf?VersionId=pEEdnFPn5aISEsH5NMaA4JqhFlCKqDMG

16. A Cautionary Note on the Use of Inter-Rater Reliability in Evaluating ... (research)
academic.oup.com
https://academic.oup.com/jrsssa/article/184/3/904/7068416

17. What do we know about grant peer review in the... | F1000Research (research)
f1000research.com
https://f1000research.com/articles/6-1335

18. Grant Peer Review Reliability: What the Data Says | Proposia (web)
proposia.ai
https://proposia.ai/blog/peer-review-reliability-evidence

19. Distrust in grant peer review—reasons and remedies (research)
academic.oup.com
https://academic.oup.com/spp/article/51/1/28/7247582

20. Experimenting with Grant Peer Review: A Mixed Methods Case Study of the Effects on Time Use and the Quality of Reviewing (research)
mdpi.com
https://www.mdpi.com/2304-6775/14/2/33

21. Dealing with the limits of peer review with innovative approaches to ... (web)
embo.org
https://www.embo.org/documents/science_policy/peer_review_report.pdf

22. The effect of writing style on success in grant applications (research)
sciencedirect.com
https://www.sciencedirect.com/science/article/pii/S1751157722000098

23. Language Barriers in Research Funding | Proposia (web)
proposia.ai
https://proposia.ai/blog/language-barrier

24. Non Native English Paper Rejection Data | Quillcademia | Quillcademia (web)
quillcademia.com
https://quillcademia.com/non-native-english-paper-rejection

25. More Versus Better: Artificial Intelligence, Incentives, and the ... (web)
pubsonline.informs.org
https://pubsonline.informs.org/doi/10.1287/orsc.2026.ed.v37.n3

26. The Daily Multi-Layered Barriers Faced by Non-Native English Researchers – Debating Development Research (web)
developmentresearch.eu
https://www.developmentresearch.eu?p=1895

27. Revising and Resubmitting Rejected Proposals (research)
research.temple.edu
https://research.temple.edu/sites/research/files/Revising-and-Resubmitting-Rejected-Proposals.pdf

28. Research Funding: the Case for a Modified Lottery (research)

pmc.ncbi.nlm.nih.gov

https://pmc.ncbi.nlm.nih.gov/articles/PMC4959526

29. Research Funding Rejection Timing: 2 Days to 14 Months (web)

grantsights.com

https://grantsights.com/blog/common-grant-rejection-reasons

30. Federal Grant Proposal Guide: Sections & Scoring (web)

grantsights.com

https://grantsights.com/blog/how-to-write-a-federal-grant-proposal

31. What to Do When a Grant Is Declined | OpenGrants (web)

opengrants.io

https://opengrants.io/encyclopedia/proposal-craft/responding-to-rejection

Method

Argument: cross-examination debate, with an assigned devil's advocate. Adjudication: independent aggregated judgment, as in the Good Judgment Project (IARPA ACE programme). Report: ICD 203 analytic standards; the recommendation is set out in Toulmin's structure of argument.