GrantReview-Bench: A Confound-Controlled Benchmark for LLM-Based Grant Proposal Review

Anonymous ACL submission

Abstract

We introduce GrantReview-Bench, a confound-controlled benchmark to evaluate LLM-based grant proposal review. Existing public grant-review datasets are limited by awarded-only collections, synthetic declined proposals, single-funder scope, or non-reproducible access, making reliable evaluation difficult. GrantReview-Bench addresses these limitations with 455 real proposals spanning 8 sources and 10 funder families, including 194 confirmed declined proposals—a 4.1× increase over the largest prior open multi-funder corpus with declined examples. The benchmark additionally includes 61 paired NIH peer-review summary statements for future per-criterion alignment evaluation.

We further propose a confound-controlled evaluation protocol that separates content-quality discrimination from superficial signals, such as the length of the proposal and the data source. The protocol combines length-stratified evaluation, per-source metrics, source-held-out cross-validation, and inter-source transfer analysis.

Across multiple LLM-based and classical baselines, naive aggregate evaluation substantially overestimates model capability. Under source-held-out evaluation, performance remains limited, and across 20 inter-source transfer evaluations, AUCs cluster near chance, with no result surviving multiple-comparison correction. These findings suggest that awarded/declined labels are highly source-dependent and that reliable evaluation of automated grant review requires explicit control of dataset confounds and distribution shift.

We release the corpus, protocol, canonical splits, and baseline implementations to support reproducible evaluation of the review of LLM-based grant proposals.

1 Introduction

Large language models (LLMs) are increasingly being explored as assistants for scientific evaluation tasks, including manuscript review, educational evaluation, and grant proposal review. In the context of research funding, the appeal is clear: grant peer review allocates tens of billions of dollars annually in public research funding and operates under substantial workload and cost constraints (Gallo et al., 2014). At the same time, inter-reviewer agreement remains only moderately reliable (Pier et al., 2018), motivating interest in computational support systems for triage, reviewer assistance, and consistency analysis.

Despite this growing interest, progress on automated grant-review evaluation is constrained by the lack of a shared and reliable benchmark. Existing public corpora are limited in several important ways. Many contain only awarded proposals, preventing discrimination-based evaluation. Others rely on synthetic declined examples (Schulz et al., 2022), which may not reflect realistic review boundaries. Several datasets are restricted to a single funder, while many proprietary corpora cannot be redistributed or independently reproduced. As a result, current evaluations are difficult to compare and often measure dataset-specific artifacts rather than genuine review capability.

In this paper, we introduce **GrantReview-Bench**, a confound-controlled benchmark for evaluating LLM-based grant proposal review. The benchmark combines real proposals from multiple funders and proposal repositories, including both awarded and declined applications, together with paired NIH peer-review summary statements for future alignment analysis. Code, metadata, and evaluation scripts are available in an anonymized repository[1].

Beyond the dataset itself, we show that evaluation in this domain is fundamentally complicated by strong dataset confounds. Two factors are partic-

[1] <https://github.com/grantagent-bench/grantreview-bench>

ularly dominant: **proposal length** and **data source**. Declined proposals in existing public corpora are often substantially shorter than awarded proposals because many are voluntarily shared drafts or summary forms rather than complete submissions. At the same time, proposal length is strongly correlated with source, creating cross-source distribution effects that can produce misleading aggregate metrics. On GrantReview-Bench, a simple length-only logistic regression achieves AUC 0.87 under standard StratifiedKFold evaluation and 0.65 under source-held-out evaluation, demonstrating that naive aggregate performance can substantially overestimate true content-quality discrimination.

To address these issues, we propose a **confound-controlled evaluation protocol** designed to separate genuine review capability from superficial dataset signals. The protocol combines (1) length-quartile-stratified evaluation, (2) per-source metrics, (3) source-held-out GroupKFold evaluation, and (4) inter-source label-transfer analysis. Together, these evaluations provide a more reliable assessment of whether models generalize across funders rather than exploiting source-specific artifacts.

Using this protocol, we evaluate multiple LLM-based review systems and classical baselines. Our central empirical finding is that cross-source generalization remains limited: across 20 inter-source transfer evaluations, AUCs cluster near chance, and no result remains significant after multiple-comparison correction. These results suggest that awarded/declined labels are highly source-dependent and that reliable evaluation of automated grant review requires explicit control of confounds and distribution shift.

More broadly, GrantReview-Bench is intended as an infrastructure for rigorous and reproducible evaluation in automated scientific review. While the benchmark focuses on grant proposals, the methodological issues it exposes - including source leakage, length bias, and non-portable supervision labels - are relevant to evaluation settings across NLP and LLM-as-judge research.

Our contributions are as follows:

- **GrantReview-Bench.** We introduce and release a multi-funder benchmark for evaluating LLM-based grant proposal review across heterogeneous proposal sources and funding agencies.
- **A confound-controlled evaluation protocol.** We identify proposal length and data source as dominant confounds and propose an evaluation framework combining length-stratified analysis, per-source metrics, and source-held-out evaluation.
- **An inter-source generalization analysis.** Through cross-source label-transfer experiments, we show that awarded/declined labels transfer poorly across funders, highlighting the importance of source-aware evaluation.
- **Reproducible benchmark infrastructure.** We release the corpus, canonical splits, baseline implementations, and analysis code to support reproducible evaluation and future extensions.

2 Related work

Open corpora for grant-review evaluation. Publicly available datasets for grant proposal review remain limited in ways that complicate reliable evaluation of automated systems. Many existing repositories, including NIH sample-application collections and NSF/SERC proposal repositories, contain only awarded proposals and are intended primarily as exemplar materials rather than evaluation benchmarks (National Institute of Allergy and Infectious Diseases, 2026; Science Education Resource Center, 2026). Other resources are restricted to single funders, such as the Wellcome Open Research Fund release,[2] or rely on voluntary sharing platforms such as Open Grants (Open Grants, 2026). Synthetic proposals have also been used for evaluation (Schulz et al., 2022), though they may not reflect realistic review boundaries.

Among existing open resources, Open Grants provides the largest publicly accessible multi-funder collection containing both awarded and declined proposals, although declined examples remain limited and voluntarily shared. The Wellcome release substantially expands declined-proposal coverage but is restricted to a single funding program. These limitations make it difficult to study cross-funder generalization or robustness under distribution shift.

GrantReview-Bench extends this landscape by aggregating proposals across multiple funders and repositories within a unified, confound-aware evaluation framework.

[2] https://wellcome.org/research-funding/open-research-fund-project-proposals-submitted-wellcome

Automated grant-review evaluation. Research on automated grant-review assessment remains limited compared to adjacent evaluation domains. Prior work has explored funding-outcome prediction, reviewer agreement, and LLM-assisted scoring. Heyard et al. (2022) show that even informed human reviewers struggle to reliably predict outcomes for borderline proposals, highlighting the ambiguity of the task. Schulz et al. (2022) evaluate LLM-based proposal scoring using synthetically degraded declined proposals, while other studies rely on proprietary or institution-specific datasets that are not publicly reproducible.

More broadly, recent work on LLM-based evaluation has examined conference peer review, educational grading, and LLM-as-a-judge settings (Liang et al., 2024). Grant review differs from these domains in both document structure and institutional heterogeneity: proposals are substantially longer, vary across funders, and contain strong source-specific formatting cues.

Our work differs from prior studies by focusing on benchmark design and evaluation methodology rather than maximizing predictive performance.

Evaluation under distribution shift and confounding. Our evaluation protocol is related to research on domain generalization and robustness under distribution shift. Benchmarks such as WILDS (Koh et al., 2021) and DomainBed (Gulrajani and Lopez-Paz, 2021) evaluate models across heterogeneous domains with systematically different train and test distributions.

GrantReview-Bench exhibits similar cross-domain structure: each proposal source corresponds to a distinct domain with different formatting conventions, length distributions, and review practices. However, the failure mode differs from standard robustness settings. Cross-source transfer frequently collapses toward chance, suggesting that awarded/declined labels may themselves be source-dependent rather than reflecting a portable notion of proposal quality.

Our work is also related to studies of verbosity and confounding in LLM evaluation (Wang et al., 2023; Saito et al., 2023). Prior work shows that LLM evaluators are sensitive to superficial properties such as response length and formatting. We show that similar effects arise in grant-review evaluation, where proposal length and source can substantially inflate naive aggregate metrics.

Dataset documentation and reproducibility. The benchmark documentation follows Datasheets for Datasets (Gebru et al., 2018), with additional guidance from data statements (Bender and Friedman, 2018) and model cards (Mitchell et al., 2019). To support reproducibility, we release canonical splits, preprocessing pipelines, evaluation scripts, and versioned metadata.

3 Dataset construction

GrantReview-Bench aggregates publicly available and permissioned grant proposals into a unified benchmark for evaluating LLM-based grant proposal review. The corpus contains 455 real proposals spanning 8 sources and 10 funder families, including 261 awarded and 194 declined applications (Table 1). In addition, the benchmark includes 61 paired NIH peer-review summary statements for future evaluation of alignment between model-generated assessments and human reviewer feedback. The benchmark combines proposals from multiple types of sources, including voluntary sharing platforms, government sample-application repositories, publicly released funder datasets, and FOIA-accessible materials.

Source	N	Aw.	Dec.	Format
Open Grants	166	119	47	PDF
Wellcome ORF 2018/19	155	17	138	TXT
NIAID Sample Apps	45	45	0	PDF
NCI Sample Apps	33	33	0	PDF
SERC (NSF Geo)	22	22	0	PDF
NIDCD Sample Apps	15	15	0	PDF
Declined extras	10	1	9	PDF
NHGRI ELSI (OCR'd)	9	9	0	TXT
Total apps	**455**	**261**	**194**	mixed
NIH summary statements	61	—	—	TXT

Table 1: Per-source composition of GrantReview-Bench. Full licensing and distribution details are provided in Appendix F.

Source aggregation and preprocessing. Proposal text is extracted from PDF and plain-text sources using standardized preprocessing pipelines, including OCR for scanned documents. Documents with insufficient or corrupted text are excluded. Source provenance and metadata are preserved throughout preprocessing to support reproducibility and source-aware evaluation.

Awarded and declined proposal coverage. A central goal of GrantReview-Bench is to support evaluation using real declined proposals, which

remain rare in existing public corpora. Open Grants provides multi-funder diversity with both funding outcomes, while the Wellcome release substantially expands declined-proposal coverage. Additional declined examples are drawn from publicly shared and FOIA-accessible materials.

Length as a dominant confound. Shown in Figure 1, proposal length differs substantially between awarded and declined applications: awarded proposals have a median length of 10,229 words compared to 992 for declined proposals. This imbalance arises because many declined examples originate from shorter drafts, summary forms, or voluntarily shared materials.

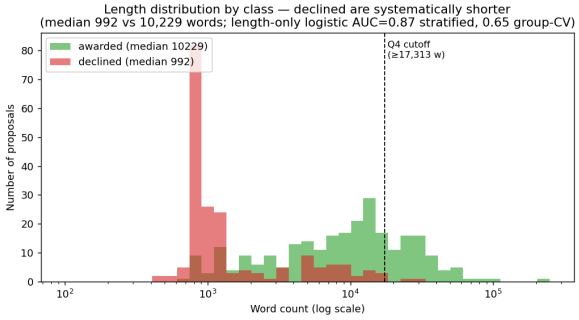

Figure 1: Proposal length distribution by funding outcome on a log-scale x-axis. Declined proposals are substantially shorter than awarded proposals, creating a strong evaluation confound.

As a result, proposal length alone becomes highly predictive under naive evaluation settings. A length-only logistic regression baseline achieves AUC 0.87 under StratifiedKFold evaluation and 0.65 under source-held-out GroupKFold evaluation, indicating that aggregate metrics can substantially overestimate content-quality discrimination by exploiting superficial distributional cues.

Because proposal length is also correlated with source, GrantReview-Bench explicitly preserves source metadata and supports confound-controlled evaluation protocols described in Section 4.

Wellcome label validation. Funding labels for the Wellcome Open Research Fund release are derived from the `decision_raw` field using a keyword-based normalization rule. Validation against both the published manifest and a two-rater human gold standard shows perfect agreement on a uniformly sampled subset (50/50 matches; $\kappa = 1.00$). Full validation details are provided in Appendix B.

Paired NIH peer-review summary statements. A subset of NIH applications is paired with publicly available NIH Summary Statements containing reviewer scores and narrative assessments across the NIH review criteria. These pairs span NIAID, NIDCD, and NHGRI applications and are released as a separate benchmark configuration.

While this paper does not report criterion-level alignment experiments, the paired data support future work on reviewer-alignment evaluation and structured assessment generation.

Distribution and reproducibility. The benchmark is distributed using a tiered release structure reflecting source-specific redistribution conditions. We release canonical splits, preprocessing scripts, metadata, and evaluation utilities to support reproducible evaluation. Full licensing and distribution details are provided in Appendix F.

4 Confound-controlled evaluation protocol

GrantReview-Bench exposes several evaluation confounds that can substantially distort performance estimates if not explicitly controlled. In particular, proposal length and proposal source are both strongly correlated with funding outcome, creating opportunities for models to exploit superficial distributional cues rather than proposal quality.

We therefore propose a confound-controlled evaluation protocol designed to separate content-quality discrimination from source-specific artifacts. The protocol combines length-aware evaluation, source-aware cross-validation, and inter-source transfer analysis to provide a more reliable assessment of model behavior under heterogeneous proposal distributions.

Why naive aggregate evaluation is misleading. Proposal length and proposal source are both strongly correlated with funding outcome on GrantReview-Bench. As a result, aggregate metrics can substantially overestimate content-quality discrimination by exploiting superficial distributional cues. For example, a length-only logistic regression baseline achieves AUC 0.87 under standard StratifiedKFold evaluation despite using no proposal content beyond document length. Similarly, train/test splits constructed without source awareness permit substantial source leakage through formatting, vocabulary, and length distributions.

4.1 Length-aware evaluation

To reduce the impact of proposal-length confounds, we recommend reporting performance stratified by proposal length quartile in addition to aggregate metrics.

On GrantReview-Bench, quartile boundaries are: Q1 $<$ 4,415 words, Q2 4,415–9,450, Q3 9,450–17,313, and Q4 $\geq$ 17,313 words.

Length-only baselines perform strongly in shorter quartiles but collapse toward chance in Q4, where proposal lengths become substantially more comparable across funding outcomes. Q4 therefore provides the cleanest available evaluation setting for assessing content-quality discrimination after controlling for length effects.

We recommend reporting:

- overall AUC,
- per-quartile AUC with sample counts,
- and Q4-specific bootstrap confidence intervals.

Bootstrap confidence intervals are computed using 2,000 paired iterations with fixed random seeds.

Length-stratified evaluation should be interpreted as a diagnostic rather than a replacement for aggregate metrics. The goal is not to eliminate length information entirely, but to measure how strongly model behavior depends on superficial length cues.

4.2 Source-aware evaluation

Because proposal source is strongly correlated with both proposal structure and funding outcome, source-aware evaluation is essential for reliable assessment.

We therefore recommend two complementary source-aware metrics:

1. **Within-source evaluation.** Report AUC separately for sources containing both awarded and declined proposals at sufficient scale.
2. **Source-held-out evaluation.** Use GroupKFold cross-validation with proposal source as the grouping variable.

Under source-held-out evaluation, all proposals from a given source are assigned exclusively to either train or test folds, preventing source leakage between splits.

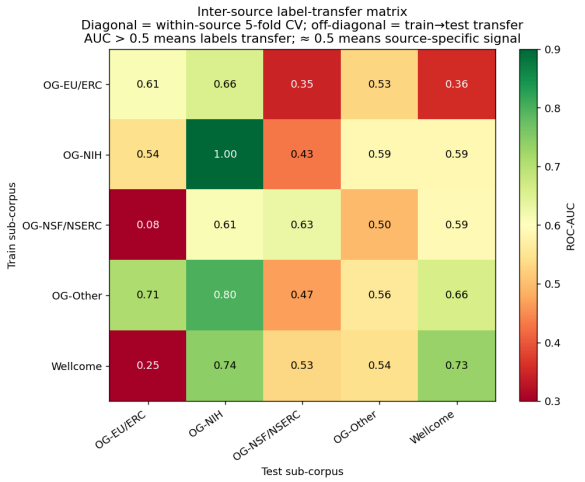

Figure 2: Inter-source transfer matrix across five source-specific sub-corpora. Diagonal entries correspond to within-source cross-validation; off-diagonal entries correspond to train→test transfer AUC. Cross-source transfer performance clusters near chance.

This evaluation setting differs substantially from standard StratifiedKFold evaluation. In preliminary experiments, performance estimates under GroupKFold are consistently lower and substantially more stable across systems, indicating that much of the apparent performance under aggregate evaluation derives from source-specific distributional cues.

For GrantReview-Bench, we recommend GroupKFold performance as the primary headline metric.

4.3 Inter-source transfer analysis

To further evaluate whether awarded/declined labels encode a portable notion of proposal quality, we introduce an inter-source label-transfer analysis.

In this setting, a classifier is trained on one source-specific sub-corpus and evaluated on another. If funding labels reflect a shared underlying concept of proposal quality, transfer performance should remain substantially above chance across sources. Conversely, if labels are source-dependent, cross-source transfer performance should collapse.

We compute transfer evaluations across five sub-corpora containing sufficient representation of both funding outcomes: OG-EU/ERC, OG-NIH, OG-NSF/NSERC, OG-Other, and Wellcome.

For each source pair, we train a TF-IDF logistic regression classifier on source-A labels and evaluate on source B. The resulting transfer matrix is shown in Figure 2, with full confidence intervals and permutation-test statistics reported in Appendix C.

Across 20 off-diagonal transfer evaluations,

AUCs cluster near chance and no result remains significant after multiple-comparison correction ($\alpha = 0.05$). These findings suggest that awarded/declined labels are highly source-dependent and that evaluation based solely on aggregate prediction accuracy can substantially overestimate cross-source generalization.

More broadly, the transfer analysis functions as a diagnostic for label portability under heterogeneous supervision settings. The goal is not to optimize transfer performance directly, but to measure whether evaluation labels encode a stable cross-domain signal.

4.4 Recommended reporting protocol

Based on these observations, we recommend that future work on GrantReview-Bench report:

- overall aggregate AUC,
- per-quartile AUC,
- source-held-out GroupKFold performance,
- within-source metrics,
- and calibration metrics such as ECE.

Where cross-source transfer experiments are performed, we additionally recommend reporting:

- bootstrap confidence intervals,
- permutation-test statistics,
- and multiple-comparison correction procedures.

Aggregate AUC alone is insufficient for reliable evaluation on this benchmark because models may exploit superficial correlations between proposal length, proposal source, and funding outcome. Confound-controlled evaluation is therefore necessary to meaningfully assess cross-source generalization in automated grant-review systems.

5 Reference baselines and evaluation analysis

We provide a set of reference baselines intended to illustrate the behavior of GrantReview-Bench under the proposed confound-controlled evaluation protocol. The goal of these experiments is not to establish state-of-the-art performance, but to demonstrate how evaluation conclusions change under different validation settings and to provide reproducible reference points for future work.

Table 2 summarizes results across classical baselines and representative LLM-based review systems. We report both standard StratifiedKFold evaluation and source-held-out GroupKFold evaluation, together with length-controlled Q4 evaluation and within-source metrics for Open Grants and Wellcome.

Aggregate evaluation substantially overestimates performance. The strongest effect observed across the baseline panel is the discrepancy between standard stratified evaluation and source-aware evaluation. Under StratifiedKFold evaluation, the length-only logistic regression baseline achieves AUC 0.87, substantially exceeding all LLM-based systems. However, under source-held-out GroupKFold evaluation, performance drops to 0.65.

This behavior illustrates the central confound in the benchmark: proposal length and funding outcome are strongly correlated, while proposal length is itself correlated with source. Under naive evaluation settings, models can therefore achieve high aggregate AUC by exploiting source-specific formatting and length cues rather than proposal content.

Cross-source evaluation remains limited. Across the evaluated systems, source-held-out GroupKFold performance remains relatively modest. Most systems cluster between AUC 0.58 and 0.65, including both classical baselines and LLM-based review pipelines. Notably, the multi-agent system achieves approximately the same GroupKFold performance as the length-only baseline despite substantially higher inference cost.

These results suggest that cross-source generalization remains difficult under confound-controlled evaluation. Improvements observed under standard aggregate evaluation do not consistently transfer to source-held-out settings.

Length-controlled evaluation. In the longest-length quartile (Q4), where proposal length becomes less predictive, most systems exhibit wide confidence intervals overlapping chance performance. Although the multi-agent system achieves the highest point estimate, the small number of declined proposals limits statistical power. Q4 should therefore be interpreted primarily as a diag-

System	Strat	Group	Q4	OG	Wel	ECE	Cost
Random (stratified)	.46 ± .04	—	.27 [.00,.59]†	—	—	.29	$0
Length-only logistic	.87 ± .04	.65 ± .13	.50 [.21,.79]†	.52 [.42,.61]	.60 [.45,.76]	.14	$0
GPT-4 (single call)	.39 ± .05	.60 ± .07	.33 [.05,.66]†	.55 [.46,.64]	.55 [.45,.63]	.28	$.01
GPT-4 + structured	.45 ± .06	.62 ± .09	.31 [.02,.61]†	.49 [.40,.59]	.71 [.57,.84]	.21	$.01
GPT-4 + RAG	.47 ± .04	.58 ± .03	.20 [.00,.51]†	.54 [.46,.62]	.61 [.50,.72]	.25	$.02
Multi-agent + gpt-4o-mini	.55 ± .06	.65 ± .19	.69 [.34,1.0]†	.43 [.34,.54]	.63 [.49,.78]	.09	$.50–1.00

Table 2: Reference baselines on GrantReview-Bench. Strat = StratifiedKFold AUC; Group = source-held-out GroupKFold AUC; Q4 = longest-length quartile evaluation; OG = Open Grants; Wel = Wellcome; ECE = expected calibration error. ± denotes mean ± standard deviation across 5 folds. [lo, hi] denotes 95% bootstrap confidence intervals (2,000 paired iterations). † Q4 contains only 4 declined examples and should be interpreted cautiously.

nostic for length sensitivity rather than a standalone leaderboard metric.

Per-source variability is substantial. Within-source evaluation reveals large performance variation across proposal sources. For example, structured prompting improves performance on Wellcome proposals but not on Open Grants, reflecting substantial differences in proposal structure, length distribution, and review conventions across sources.

These source-dependent effects reinforce the importance of reporting source-aware metrics rather than relying solely on aggregate evaluation.

Calibration and cost tradeoffs. Calibration quality varies substantially across systems. The multi-agent baseline achieves the lowest expected calibration error (ECE), but this improvement does not correspond to clear gains in cross-source discrimination performance. At the same time, multi-agent pipelines incur substantially higher inference cost than simpler prompting strategies.

These results highlight that calibration, inference cost, and aggregate AUC may behave independently under confound-controlled evaluation.

Interpretation. Overall, the baseline analysis supports the central motivation for GrantReview-Bench: evaluation conclusions depend strongly on whether confounds such as proposal length and source leakage are controlled. Under naive aggregate evaluation, superficial signals can dominate performance estimates. Under source-aware and length-controlled evaluation, performance differences between systems become substantially smaller and cross-source generalization remains limited.

The released baselines are therefore intended primarily as reproducible reference points for future benchmark development rather than as competitive system rankings.

6 Distribution and falsifiability invitation

Falsifiability. The claim that GrantReview-Bench is the largest open multi-funder corpus with real declined proposals is intended as a *falsifiable statement*. We conducted an extensive search across public literature, agency repositories, voluntary-share archives, and FOIA channels and did not identify a larger open corpus of this type. If such a resource exists, we invite the community to contact the (anonymous-during-review) maintainers; we will apply the evaluation protocol in §4 and report a direct comparison.

Distribution. The corpus is distributed in a tiered manner reflecting source-specific licensing constraints. Open Grants (CC BY 4.0) and Wellcome ORF are directly redistributed via HuggingFace, while other sources (NIH sample applications and SERC materials) are provided as link-only access through a build script. In total, approximately 321 of 455 documents are directly downloadable, with the remainder accessible via source links.

Documentation and licensing. Full per-source license review, Datasheet for Datasets documentation, and Croissant 1.0 metadata are provided in Appendix F.

7 Conclusion

We introduce GrantReview-Bench, an open multi-funder benchmark for evaluating LLM-based grant proposal review, together with a confound-controlled evaluation protocol and an inter-source label-transfer diagnostic.

Our central finding is that cross-funder generalization is limited: across 20 off-diagonal transfer evaluations, no result survives multiple-comparison

correction, and transfer performance is largely indistinguishable from chance. This indicates that awarded/declined labels in existing open corpora do not encode a portable notion of proposal quality.

We release the corpus, evaluation protocol, reference baselines, and analysis code as a community resource and a falsifiability invitation. Progress on this problem will require broader access to declined proposals and more careful treatment of evaluation confounds.

Acknowledgments

We acknowledge the use of AI assistants (ChatGPT) during the preparation of this manuscript for stylistic copyediting, section polishing, and checklist formatting verification. All core ideas, experimental designs, and data interpretations remain entirely the work of the authors.

8 Limitations

GrantReview-Bench is intended as infrastructure for confound-aware evaluation of automated grant-review systems. However, several limitations affect both the benchmark and the interpretation of reported results.

Representativeness bias. Declined proposals are drawn primarily from voluntary sharing platforms and the Wellcome Open Research Fund release, while several awarded-proposal sources consist of curated exemplar applications from NIH and NSF/SERC repositories. As a result, neither awarded nor declined proposals should be assumed to represent the full distribution of real-world funding submissions.

Source heterogeneity and confounding. Only a subset of benchmark sources contain both funding outcomes at substantial scale, limiting reliable within-source evaluation. In addition, proposal length, formatting conventions, and review practices vary substantially across funders, creating heterogeneous label semantics and strong superficial correlations with funding outcome. Although the benchmark supports confound-controlled evaluation, no protocol can fully eliminate these effects.

Limited statistical power. Some confound-controlled subsets, particularly the longest-length quartile (Q4), contain relatively few declined proposals, resulting in wide confidence intervals. Similarly, the paired NIH reviewer summary statements are intended primarily as infrastructure for future reviewer-alignment studies rather than fully validated supervision for criterion-level evaluation.

Benchmark scope. GrantReview-Bench focuses on proposal-level outcome prediction and review evaluation under heterogeneous funding sources. It does not capture the full institutional context of grant review, including panel deliberation, reviewer expertise matching, budget constraints, programmatic priorities, or long-term funding outcomes.

Overall, the benchmark should be interpreted as a resource for studying evaluation methodology, confound control, and cross-source robustness rather than as a definitive measure of scientific merit or funding quality.

9 Ethical considerations

Provenance and redistribution. GrantReview-Bench is constructed from publicly available or permissioned materials, including voluntary sharing platforms, institutional repositories, publicly released funding datasets, and FOIA-accessible records. Benchmark components are distributed using a tiered release structure reflecting source-specific redistribution conditions. Sources permitting redistribution are released directly with attribution, while restricted materials are provided through link-based reconstruction scripts rather than mirrored redistribution.

Personally identifiable information. Some benchmark documents contain personally identifiable information (PII), including investigator names, institutional affiliations, and other identifying details preserved from the original source materials. Because several repositories distribute proposals in near-original form, complete anonymization is not always feasible without substantially altering document structure. To support responsible use, the benchmark preserves provenance metadata and provides an opt-out mechanism for future releases.

Misuse risks. GrantReview-Bench is intended for research on evaluation methodology, robustness, and confound-aware assessment of automated review systems. It is not intended for deployment as an autonomous grant-review or funding-allocation system.

Funding decisions reflect complex institutional processes involving reviewer expertise, panel deliberation, disciplinary priorities, and programmatic

considerations that are not fully captured in benchmark labels. As demonstrated by the inter-source transfer analysis, awarded/declined labels are also highly source-dependent and should not be interpreted as universal measures of proposal quality.

We therefore discourage the use of this benchmark as training data for production funding-allocation systems or as a proxy for scientific merit evaluation.

Broader implications. The benchmark highlights broader methodological concerns in LLM-based evaluation, including source leakage, superficial formatting cues, and non-portable supervision labels. These issues extend beyond grant review and may affect benchmark construction in other LLM-as-judge and institutional decision-support settings.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Stephen A. Gallo, Afton S. Carpenter, David Irwin, Christopher D. McPartland, Joseph Travis, Sandra Reynders, Lisa A. Thompson, and Scott R. Glisson. 2014. The validation of peer review through research funding outcomes. *PLoS ONE*, 9(10):e108184.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*. Updated in CACM 2021, 64(12):86–92.

Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*.

Rachel Heyard, Sonia Pellegrini, Leonhard Held, and Matthias Ott. 2022. Predicting the outcome of grant peer review: A pre-registered evaluation. *eLife*, 11:e75800.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Irena Gao, and 1 others. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pages 5637–5664.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Bing Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, and 1 others. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229.

National Institute of Allergy and Infectious Diseases. 2026. Niaid sample applications and summary statements. https://www.niaid.nih.gov/grants-contracts/sample-applications. Accessed: 2026-05-21.

Open Grants. 2026. Open grants. https://www.ogrants.org/. Accessed: 2026-05-21.

Elizabeth L. Pier, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J. Nathan, Cecilia E. Ford, and Molly Carnes. 2018. Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, 115(12):2952–2957.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Robert Schulz, Adrian Barnett, René Bernard, Nicholas J. L. Brown, Jennifer A. Byrne, Peter Eckmann, Malgorzata A. Gazda, Halil Kilicoglu, Eric M. Prager, Maia Salholz-Hillel, and 1 others. 2022. Is the future of peer review automated? *BMC Research Notes*, 15(1):203.

Science Education Resource Center. 2026. Sample nsf geoscience proposals. https://serc.carleton.edu/NAGTWorkshops/grants/examples.html. Accessed: 2026-05-21.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

A Appendix — Datasheet for Datasets (condensed)

Following Gebru et al. (2018).

Motivation. This dataset was created to enable reproducible evaluation of LLM-based grant proposal review systems. Prior public corpora are dominated by funded applications and lack declined examples. GrantReview-Bench is the largest open corpus we know of with real declined proposals spanning multiple funders. It was created by the (anonymized) authors without external funding.

Composition. The corpus contains 455 grant applications (261 awarded, 194 declined) across 8 sources (Table 1), along with 61 paired NIH peer-reviewer summary statements. Each instance includes metadata such as source dataset, label, funder, funder family, format (PDF/TXT), word count, and extracted text. Proposal length varies substantially: median 4,508 words overall, with awarded proposals much longer than declined ones (median 10,229 vs. 992).

Collection. Data are aggregated from publicly available sources, including voluntary sharing platforms, agency sample-application repositories, and funder-specific releases. Additional declined examples are drawn from publicly shared PDFs and FOIA-accessible materials. Data collection took place between February and May 2026.

Preprocessing and labeling. PDF text is extracted using standard tools; TXT sources are used as provided. Documents with insufficient or corrupted text are excluded. For scanned documents, optical character recognition is applied with quality filtering. Labels for Wellcome entries are derived from the `decision_raw` field using a keyword-based rule, which is validated against both the published manifest and a two-rater human gold standard (50/50 agreement; $\kappa = 1.00$; see Appendix B).

Intended use. The dataset is intended for evaluating LLM-based grant-review systems under the protocol in §4, and for studying alignment between automated and human review using the paired NIH summary statements. It is not intended as ground truth for scientific quality, for deployment in real funding decisions, or for training production grant-scoring models.

Distribution. Distribution follows a tiered, per-source licensing model (§F). Open Grants and Wellcome ORF are redistributed directly, while other sources are provided via link-based access with a build script. A persistent release (HuggingFace + Croissant 1.0 metadata + Zenodo DOI) will be provided upon publication.

Maintenance. The dataset is maintained by the (anonymized) authors. Updates are versioned, with changelogs and issue tracking provided in the repository. The repository will be de-anonymized upon publication. Planned extensions include expanded coverage of declined proposals.

Ethics and legal considerations. All documents were publicly available at the time of collection. Sources include government releases, voluntary sharing platforms, and funder-issued datasets. No additional personal data processing beyond aggregation is performed. An opt-out mechanism allows individuals to request removal of records, providing an ongoing pathway for data-subject rights. The dataset is intended for research use and not for decision-making applications.

B Appendix — Wellcome label validation (full)

The Wellcome 2018/19 ORF release provides a `decision_raw` free-text field. We derive binary labels using a keyword-based rule: `not shortlisted/not funded/unsuccessful/declined` → declined; `funded/awarded` → awarded; ambiguous cases are excluded. Wellcome contributes 138 of 194 declined examples (71%), making label accuracy critical.

(a) Internal consistency (rule vs. manifest). We compare rule-derived labels against the published manifest labels on a uniformly sampled subset (50 entries, fixed random seed):

Metric	Value
Sample size	50 of 155
Agreement	**50/50 (100%)**
Wilson 95% CI	[0.929, 1.000]
Cohen's κ	1.00
Disagreements	0

Table 3: Rule vs. manifest label agreement.

(b) Human validation (rule vs. 2-rater consensus). We further validate the rule against a two-rater human gold standard on the same sampled subset:

Metric	Value
Sample size	50 of 155
Rule-vs-human agreement	**50/50 (100%)**
Wilson 95% CI	[0.929, 1.000]
Cohen's κ	1.00
Inter-rater agreement	50/50 (100%)

Table 4: Rule vs. human gold-standard agreement.

Reproducibility. All audit results are reproducible from the released artifacts using the provided validation scripts and fixed random seeds.

C Appendix — Inter-source transfer matrix (full)

Full 5×5 matrix with bootstrap 95% CIs and one-sided permutation p-values (one-sided p flagged with $*$ if < 0.05, uncorrected) (Table 5):

D Appendix — Reproducibility details

All baseline results in Table 2 are computed from the released corpus, canonical splits, and analysis code provided in the supplementary materials. Classical baselines use scikit-learn default hyperparameters unless otherwise specified; no extensive hyperparameter optimization was performed.

Model configuration. LLM-based systems use pinned model snapshots (e.g., `gpt-4o-2024-08-06` and `gpt-4o-mini-2024-07-18`). All prompts and system configurations used for evaluation are included in the released code.

Randomness control. All experiments use fixed random seeds for reproducibility. Python and NumPy seeds are set at module load. Bootstrap confidence intervals are computed with 2,000 paired iterations (seed 42), and permutation tests use 1,000 iterations (seed 43).

Evaluation pipeline. Baseline predictions, cross-validation metrics, and transfer analyses are produced using dedicated evaluation scripts included in the release. GroupKFold AUC is computed using source-held-out cross-validation. The inter-source transfer matrix (§4.3) is computed using logistic regression on TF-IDF features with consistent cross-validation and resampling settings. Wellcome label validation (§3) is implemented as a separate validation module.

Dependencies and execution. All dependencies are version-pinned, and the full pipeline is designed to be reproducible from a clean environment. Running the provided scripts with the specified seeds reproduces all reported results.

Data processing pipeline. Corpus construction and preprocessing are implemented using a unified build pipeline included in the release. The pipeline is designed to be portable and reproducible across environments.

Dependencies. All dependencies are version-pinned and provided with the released code to ensure consistent execution.

Release. Code and corpus are available in an anonymized repository: https://github.com/grantagent-bench/grantreview-bench. The repository will be de-anonymized upon publication.

E Appendix — Splits and protocol details

Canonical 70/15/15 splits. Data are split into train/dev/test partitions (70/15/15) using stratification by label within each source, ensuring consistent class balance across splits. Splits are generated with a fixed random seed (42) and are fully reproducible from the released code. The NSF/SERC subset (22 awarded, 0 declined) is held out as a single-class, out-of-distribution probe. Because it contains no declined examples, it does not support AUC computation and is instead used for analyzing score distributions and model behavior under distribution shift.

Calibration metrics. Calibration is evaluated using expected calibration error (ECE) with 10 equal-frequency bins, reported both in aggregate and per source (Open Grants and Wellcome). Reliability diagrams are provided with a diagonal reference. Brier score is also computed, though its interpretation is affected by class imbalance.

Statistical reporting. All pairwise comparisons are evaluated using paired bootstrap confidence intervals, with explicit reporting of test direction (one- or two-sided), iteration counts, and random seeds. Multiple-comparison correction (Bonferroni or Benjamini–Hochberg) is applied where appropriate. For small test sets ($n < 20$), permutation-based null tests are used. Random seeds are disclosed for all stochastic procedures.

F Appendix — Distribution and license review

Per-source licensing. Each source is reviewed for redistribution under its respective licensing terms (last verified 2026-05-07). Open Grants (CC BY 4.0) is redistributed directly. Wellcome ORF 2018/19 is included under its public-release terms for non-commercial academic research use with attribution. NIH sample-application repositories (NIAID, NCI, NIDCD, NHGRI) and SERC materials are provided as link-only access due to source-specific usage conditions. Additional declined examples are included from publicly shared

Train\Test	OG-EU/ERC	OG-NIH	OG-NSF	OG-Other	Wellcome
OG-EU/ERC	—	0.66 [0.38, 0.91]	0.35 [0.16, 0.54]	0.53 [0.36, 0.70]	**0.36 [0.19, 0.52]***
OG-NIH	0.54 [0.11, 1.00]	—	0.43 [0.23, 0.64]	0.59 [0.42, 0.75]	0.59 [0.43, 0.74]
OG-NSF/NSERC	**0.08 [0.00, 0.32]***	0.61 [0.35, 0.84]	—	0.50 [0.33, 0.66]	0.59 [0.43, 0.74]
OG-Other	0.71 [0.22, 1.00]	0.80 [0.57, 0.99]	0.47 [0.26, 0.68]	—	0.66 [0.51, 0.80]
Wellcome	0.25 [0.00, 0.70]	0.74 [0.50, 0.95]	0.53 [0.32, 0.73]	0.54 [0.38, 0.70]	—

Table 5: Full inter-source label-transfer matrix. Bootstrap 95% CI from 2,000 paired iterations (seed 42). * = one-sided permutation $p < 0.05$ (uncorrected); 1,000 permutations, seed 43. Sub-corpus sizes: OG-EU/ERC (11), OG-NIH (22), OG-NSF/NSERC (56), OG-Other (59), Wellcome (155). The two flagged cells: NSF→EU/ERC (AUC 0.08, p=0.024), EU/ERC→Wellcome (AUC 0.36, p=0.034). Bonferroni-corrected threshold for 20 off-diagonal cells: $\alpha/20$=0.0025. Neither flagged cell survives multiple-comparison correction.

documents and FOIA-accessible materials, with per-record provenance maintained in the release.

Opt-out and removal mechanism. Any individual identified in a record may request removal by contacting the (anonymous-during-review) maintainer. Requests are processed within 4 weeks for the next versioned release. Removed records are replaced with a stub entry ("removed at request") to preserve cross-version reproducibility. All removals are recorded in the dataset changelog.

Versioning. The dataset is released in versioned form. v1.0 (this release): 455 applications + 61 paired summary statements. v1.1 (target Q4 2026): expanded declined coverage and optional PII-redacted variant. v2.0 (target 2027): further expansion with FOIA-collected declined proposals.

Persistent identifiers and metadata. The dataset is distributed with Croissant 1.0 metadata. A HuggingFace dataset identifier and Zenodo DOI will be provided upon publication.