When Multi-Agent LLM Grant Proposal Review Fails: A Controlled Evaluation Study

Anonymous ACL submission

Abstract

Grant proposal review has emerged a promising application for large language models (LLMs), with recent work hypothesized that multi-agent architectures may improve evaluation quality through decomposition, retrieval, and debate. We test this hypothesis through a controlled study of **GrantAgent**, a multi-agent grant-review system combining criterion-specific agents, retrieval augmentation, adversarial debate, and meta-review aggregation.

Our central finding is negative: **under controlled evaluation, the multi-agent system fails to demonstrate reliable grant-quality discrimination.** Under aggregate evaluation on a heterogeneous corpus of 455 real grant proposals, GrantAgent achieves ROC-AUC 0.55 and appears to outperform several single-call baselines. However, we show that this apparent gain is largely driven by confounds tied to proposal length and source distribution. Once evaluation is restricted to within-source and length-controlled settings, performance collapses toward chance. On the source that contains both funded and declined proposals, all methods perform near chance, and a four-model cross-LLM analysis further shows that this failure persists across model families and scales.

Overall, our results suggest that current prompt-based and multi-agent LLM evaluator architectures do not reliably recover grant-review quality signals from heterogeneous open corpora. These findings highlight the importance of controlled within-source evaluation when assessing LLM-based evaluators in high-stakes domains.

1 Introduction

Grant peer review is one of the most consequential expert-evaluation processes in science. Agencies such as the U.S. National Science Foundation (NSF) and the National Institutes of Health (NIH) collectively process tens of thousands of proposals annually, with reviewers spending multiple hours per application producing criterion-specific critiques and funding recommendations. At the same time, human grant review is expensive, time-constrained, and only moderately reliable: prior work reports inter-reviewer agreement on grant criteria in the range of Spearman $\rho \approx 0.3$–0.5 (Pier et al., 2018; Gallo et al., 2016). These limitations have motivated growing interest in whether LLMs could assist or partially automate aspects of grant evaluation.

Recent progress in multi-agent LLM systems has further strengthened this possibility. Previous work has shown that decomposition into specialized agents, adversarial debate, retrieval augmentation, and meta-review aggregation can improve reasoning and evaluation performance on complex tasks (Wu et al., 2024; Chan et al., 2023; Du et al., 2024; Li et al., 2023). This has led to an emerging intuition that grant review may benefit from similar architectures: instead of asking a single LLM to judge an entire proposal in one pass, a system could decompose review into criterion-specific evaluations aligned with funder rubrics, synthesize multiple perspectives, and produce calibrated funding recommendations.

In this paper, we test this hypothesis directly through a case study of **GrantAgent**, a multi-agent LLM system for grant proposal review. GrantAgent decomposes review into six criterion-specific agents aligned with NIH and NSF evaluation criteria, augments reasoning with retrieval from funded grants, triggers adversarial debate when agents disagree, and synthesizes outputs through a meta-review stage into funding probabilities and interpretable review flags.

Our central finding is negative: **under controlled evaluation, the multi-agent approach fails to demonstrate reliable grant-quality discrimination.** At first glance, aggregate evaluation suggests modest success. On a heterogeneous corpus of 455 real proposals spanning multiple fun-

ders, GrantAgent achieves aggregate ROC-AUC 0.55, outperforming several single-call GPT-4 baselines. However, we show that this apparent improvement is largely explained by two major evaluation confounds. First, proposal length is highly predictive of funding outcome in open grant corpora: declined proposals are systematically shorter than funded proposals, allowing a length-only baseline to achieve surprisingly high aggregate performance. Second, aggregate evaluation across multiple funders introduces a source-distribution confound resembling Simpson's paradox: models can distinguish proposal *formats* and funder-specific structures rather than proposal quality itself. NIH full applications, NSF proposals, and Wellcome short-form submissions differ systematically in length, organization, and label distribution, creating exploitable surface signals unrelated to substantive scientific merit.

Once these confounds are controlled, the apparent gains of the multi-agent system largely disappear. Under within-source evaluation, no tested method achieves statistically reliable discrimination. On Open Grants, the only source with substantial funded and declined coverage, GrantAgent achieves AUC 0.43, statistically indistinguishable from chance. A four-model cross-LLM panel (GPT-4o-mini, GPT-4o, Claude Sonnet, Claude Opus) further shows that no tested model family achieves robust discrimination once cross-source format effects are removed. Increasing model size also fails to recover meaningful within-source performance, suggesting that the limitation is not simply a small-model artifact.

Interestingly, the strongest positive signal emerges in the longest word-count quartile subset of proposals, where proposal length ceases to be predictive. In this regime, single-call GPT-4 baselines collapse below chance, while the multi-agent system achieves substantially higher point estimates. However, these comparisons remain underpowered due to the scarcity of long declined proposals, preventing strong claims of reliable content-quality discrimination. We therefore treat these findings as suggestive directional evidence rather than definitive proof of capability.

Despite failing as a robust classifier, the multi-agent system exhibits two narrower strengths. First, it produces comparatively well-calibrated funding probabilities relative to single-call baselines. Second, its criterion-specific outputs and interpretable flags provide potentially useful signals for human-in-the-loop triage workflows. We therefore argue that multi-agent LLM systems may still have practical value as uncertainty-aware review assistants, even if they do not reliably recover latent grant quality.

The broader implication of our findings is methodological. Aggregate evaluation on heterogeneous grant corpora can substantially overstate LLM review capability by rewarding shortcut signals tied to proposal length and source distribution. We therefore argue that future work on LLM-based grant review should prioritize controlled evaluation protocols, including within-source analysis and length-controlled comparisons, rather than relying on aggregate metrics alone. More fundamentally, our results suggest that current LLMs — including sophisticated multi-agent architectures — do not reliably infer grant-review quality signals from heterogeneous open corpora. The failure persists across prompting strategies, retrieval augmentation, debate mechanisms, model families, and model scales. This reframes the central challenge of automated grant review: the limiting factor may not simply be model architecture, but the structure and consistency of the underlying evaluation signal itself.

Our contributions are therefore primarily methodological and empirical rather than benchmark-oriented:

1. We present a controlled case study of a multi-agent LLM architecture for grant review, including criterion-specific decomposition, adversarial debate, retrieval augmentation, and meta-review aggregation.

2. We show that aggregate grant-review evaluation is strongly confounded by proposal length and source distribution, causing apparent performance gains that do not survive controlled evaluation.

3. We demonstrate that neither single-call nor multi-agent LLM systems achieve robust within-source grant-quality discrimination under the tested settings.

4. We provide evidence that multi-agent systems may still offer narrower utility through calibration and interpretable triage-oriented signals, despite failing as reliable classifiers.

5. We argue for controlled within-source and length-aware evaluation as the appropriate

methodology for future work on LLM-based grant review.

Our goal is not to rule out all possible machine-learning approaches to grant outcome prediction. In particular, we do not evaluate supervised discriminative classifiers trained directly on proposal labels or embeddings. Instead, we focus specifically on prompt-based and multi-agent LLM evaluation architectures, which are increasingly proposed as general-purpose reviewer systems. The paper therefore studies whether these architectures recover robust grant-quality discrimination under controlled evaluation.

2 Related Work

2.1 Reliability of Human Grant Review

Grant peer review itself is an imperfect and only moderately reliable evaluation process. Previous studies of NIH study sections report inter-reviewer agreement on proposal criteria in the range of Spearman $\rho \approx 0.3$–0.5 (Pier et al., 2018), with related work reporting similarly modest agreement under Cohen's κ (Gallo et al., 2016). These findings motivate interest in computational assistance for grant review, but also suggest that funded/declined labels may not represent a fully consistent latent quality signal.

More broadly, grant-review outcomes are shaped not only by proposal quality, but also by funder-specific criteria, panel dynamics, and programmatic fit. As a result, aggregate prediction performance may overestimate true content-level understanding, particularly in heterogeneous multi-funder settings.

2.2 LLM-as-Judge and Evaluation Bias

Recent work has explored using LLMs as evaluators for generation quality, reasoning quality, and scientific review. (Zheng et al., 2023) showed that strong LLMs can approximate human preferences in conversational evaluation, but also identified systematic biases including verbosity bias, positional bias, and self-enhancement effects. Other work has similarly raised concerns about evaluator instability and high sensitivity to prompt structure and presentation order (Wang et al., 2024).

Our work extends this line of inquiry to grant proposal review, a challenging long-form evaluation setting involving heterogeneous document structures and noisy outcome labels. We show that proposal length and source-specific formatting can substantially confound aggregate evaluation metrics, creating the appearance of meaningful evaluator capability even when within-source discrimination remains near chance.

Our work is also related to recent studies of LLM-based scientific review. (Liang et al., 2024b) evaluated GPT-4-generated feedback on scientific papers and found partial overlap with human reviewer judgments. We study a related but distinct setting: grant proposal evaluation under heterogeneous funder distributions and binary funded/declined outcomes.

2.3 Multi-Agent LLM Systems

A growing literature argues that multi-agent architectures can improve LLM reasoning through decomposition, debate, and role specialization. Frameworks such as AutoGen (Wu et al., 2024) and CAMEL (Li et al., 2023) provide general infrastructure for coordinated multi-agent interaction. Debate-based systems such as MAD (Liang et al., 2024a) report gains on reasoning tasks by encouraging adversarial critique and iterative refinement between agents.

Several works have applied these ideas specifically to evaluation. ChatEval (Chan et al., 2023)) uses debating evaluator agents to improve assessment quality, while (Wang et al., 2023) study multi-persona self-collaboration within a single LLM. These approaches motivate the hypothesis that evaluator decomposition may reduce individual-model biases and improve robustness on complex judgment tasks.

Our work focuses specifically on prompt-based evaluator architectures rather than supervised discriminative text classifiers. Unlike prior multi-agent evaluation system, we study a high-stakes long-document evaluation setting where labels are noisy and heterogeneous rather than crowd-aggregated preference judgments. More importantly, our primary contribution is not a new multi-agent architecture, but an empirical evaluation showing that apparent gains from such architectures can disappear under controlled evaluation. In this sense, our work is closer to a negative-result study on evaluator robustness than to a system-design paper.

2.4 Confounded Evaluation and Shortcut Learning

Our findings connect to a broader literature on shortcut learning and benchmark confounds in ma-

chine learning. Models often achieve strong aggregate performance by exploiting spurious correlations that are predictive within a dataset but unrelated to the intended task (Geirhos et al., 2020). Previous work has documented such effects in natural language inference through annotation artifacts (Gururangan et al., 2018), in question answering through adversarial evaluation (Jia and Liang, 2017), and in summarization and evaluation benchmarks through instability and metric mismatch (Fabbri et al., 2021).

We show that heterogeneous grant-review corpora contain strong shortcut signals tied to proposal length and source distribution. Under aggregate evaluation, these effects can create the appearance of meaningful evaluator capability even when within-source discrimination remains near chance. Our work therefore emphasizes the importance of controlled within-source and length-aware evaluation for assessing LLM-based evaluators.

3 Method

Our goal is not to propose a fundamentally novel agent architecture, but to evaluate whether commonly proposed multi-agent evaluation strategies improve grant-review capability under controlled conditions. We therefore construct a representative multi-agent grant-review pipeline combining four widely used design patterns in recent LLM systems: criterion decomposition, retrieval augmentation, adversarial debate, and meta-review aggregation.

3.1 Multi-agent architecture

We implement a representative multi-agent evaluation pipeline, GrantAgent, designed to reflect common architectural motifs in recent LLM-agent systems: role decomposition, retrieval augmentation, adversarial debate, and meta-review aggregation. Each agent receives the full proposal text plus retrieved similar funded grants and produces (a) a 1–5 score, (b) flagged issues from a fixed taxonomy, and (c) a textual rationale. Criterion agents are routed by funder profile: **NSF profile (NSF_CORE):** intellectual_merit, broader_impacts, feasibility_risk, team_capability, budget_plan, citation_integrity; and **NIH profile (NIH_R01):** significance, innovation, approach, investigator, environment, citation_integrity.

Adversarial debate When per-criterion agent scores disagree (range > 1.0 on the 1–5 scale), a debate sub-routine is invoked: each agent produces a defense of its score given the others' rationales, and converges via weighted average. Debate is bounded to 3 rounds. This conditional triggering controls compute cost while addressing cases of genuine inter-agent disagreement.

Retrieval grounding A ChromaDB index of 750 funded NSF/NIH grant abstracts (collected from public NSF Awards API and NIH RePORTER API) provides top-k retrieval (k=4) for in-context similarity grounding. We verified zero exact-title overlap between this index and our test corpus to rule out direct test-set leakage.

Meta-reviewer A meta-reviewer LLM call synthesizes agent outputs into: 1) `overall_weighted_score` (1-5); 2) `verdict` $\in$ Fund-Aligned, Borderline, Misaligned; 3) `p_fund` (0-1) – calibrated funding probability; and 4) `flag_for_human` (boolean) – whether the case should escalate to human review.

3.2 Implementation

GrantAgent is implemented in Python 3.10 (Python Software Foundation, 2021) with FastAPI orchestration. Agents call OpenAI gpt-4o-mini (OpenAI, 2024a) for cost-effective inference. Each review consumes approximately 30K input tokens $\times$ 18 LLM calls per proposal, costing \$0.50–\$1.00 and taking 80–160 seconds wall-clock.

4 Experimental Setup

Our goal is to evaluate whether multi-agent LLM architectures recover meaningful grant-review signals under controlled conditions. Because heterogeneous grant corpora contain substantial distributional variation across funders and proposal formats, we focus on evaluation protocols designed to separate genuine content-quality discrimination from exploitable surface correlations. This section describes the evaluation corpus, baselines, and controlled evaluation procedures used throughout the paper.

4.1 Evaluation Corpus

We evaluate on a heterogeneous corpus of 455 real grant proposals collected from publicly available or voluntarily shared sources spanning NSF, NIH, Wellcome, EU/ERC, and related agencies (Table 1). The corpus contains 261 funded and 194 declined proposals. All documents are in English and preserve their original, heterogeneous submission formats.

Source	Total	Awd.	Dec.
Open Grants	166	119	47
Wellcome Open Research Fund	155	17	138
NIH sample applications	102	102	0
SERC / NSF Geosciences	22	22	0
Other declined sources	10	1	9
Total	455	261	194

Table 1: Evaluation corpus composition aggregated by source family.

The corpus is intentionally heterogeneous: proposals differ substantially across funders in structure, length, formatting conventions, and funding distributions. Funded proposals are systematically longer than declined proposals (median 10,229 vs. 992 words), creating a strong length confound under aggregate evaluation.

4.2 Controlled Evaluation Protocol

Aggregate evaluation on heterogeneous grant corpora can substantially overestimate evaluator capability due to shortcut signals tied to proposal length and source distribution. We therefore evaluate systems under four complementary settings.

Aggregate evaluation. We first report standard aggregate discrimination metrics across the full corpus for comparability with prior work, including ROC-AUC, balanced accuracy, macro-F1, and Matthews correlation coefficient (MCC).

Cross-source evaluation. In addition to standard StratifiedKFold evaluation, we evaluate using GroupKFold split by source. This prevents proposals from the same source distribution from appearing in both train and test folds and tests whether systems generalize across funder distributions rather than exploiting source-specific correlations.

Within-source evaluation. We compute ROC-AUC separately for sources containing both funded and declined proposals. This provides the clearest estimate of whether models discriminate proposal quality within a consistent source distribution.

Length-controlled evaluation. To reduce the influence of proposal length, we partition proposals into word-count quartiles and compute AUC within each quartile. The longest quartile serves as the primary length-controlled setting because proposal length ceases to be predictive in this regime.

Confidence intervals are computed using paired bootstrap resampling with 2,000 iterations. Because several subgroup evaluations contain limited numbers of declined proposals, especially in the longest quartile, we emphasize confidence intervals and effect consistency in addition to nominal significance tests.

4.3 Baselines

We compare GrantAgent against a set of single-call LLM and non-LLM baselines designed to test whether multi-agent decomposition provides meaningful advantages beyond simpler evaluation strategies.

Single-call GPT-4. The simplest baseline uses a single LLM call to evaluate the entire proposal and directly predict funding likelihood. The model receives the full proposal text and produces a scalar funding probability.

Structured-prompt baseline. This baseline augments the single-call setup with rubric-style prompting aligned to common grant-review criteria (e.g., significance, innovation, feasibility, investigator strength). The goal is to test whether explicit prompting structure alone accounts for apparent gains from the multi-agent system.

Chain-of-thought baseline. We additionally evaluate a chain-of-thought (CoT) variant in which the model is instructed to reason step-by-step before producing a final funding assessment. This tests whether explicit deliberative reasoning improves grant-review discrimination without requiring multi-agent decomposition.

Retrieval-augmented baseline. To isolate the effect of retrieval grounding, we evaluate a single-call retrieval-augmented generation (RAG) baseline using the same retrieval corpus as GrantAgent. Retrieved funded-grant abstracts are appended to the prompt before evaluation.

Length-only baseline. Because proposal length is highly correlated with funding outcome in heterogeneous open corpora, we include a logistic regression classifier using proposal word count as its sole feature. This baseline is critical for determining whether apparent model performance reflects genuine content-quality discrimination or shortcut exploitation.

Random baseline. We report a stratified random baseline for reference when interpreting subgroup evaluations with small sample sizes.

Unless otherwise specified, all LLM baselines use GPT-4-family models with identical proposal text inputs and produce scalar funding probabilities in the range $[0, 1]$. This design intentionally minimizes differences in input access so that comparisons primarily reflect architectural choices rather than information asymmetry.

In addition to the main comparisons, we evaluate a four-model cross-family panel consisting of GPT-4o-mini (OpenAI, 2024a), GPT-4o (OpenAI, 2024b), Claude Sonnet, and Claude Opus (Anthropic, 2024) on a controlled subset of non-biomedical proposals. The purpose of this experiment is not leaderboard comparison, but to test whether the observed evaluation failures persist across model families and scales.

4.4 Metrics

We report ROC-AUC as the primary metric, along with balanced accuracy, macro-F1, and Matthews correlation coefficient (MCC). Because several methods output scalar funding probabilities, we additionally evaluate calibration using expected calibration error (ECE) and Brier score.

Threshold-dependent metrics are computed using cross-validation with thresholds selected on training folds and evaluated on held-out folds. Confidence intervals are estimated using paired bootstrap resampling with 2,000 iterations.

5 Results

5.1 Aggregate Evaluation is Misleading

Under aggregate evaluation, GrantAgent appears to modestly outperform the prompting-based baselines, achieving ROC-AUC 0.55 compared to 0.39–0.47 for the GPT-4 variants (Table 2).

However, the aggregate results are dominated by shortcut structure. A trivial length-only baseline achieves AUC 0.87 using proposal word count alone, indicating that proposal length is highly predictive in the heterogeneous corpus. Because funded proposals are often full submissions while many declined proposals are shorter summaries or drafts, aggregate evaluation rewards systems that exploit superficial length correlations rather than substantive proposal quality.

We therefore additionally evaluate using GroupKFold split by source, which prevents proposals from the same source distribution from appearing in both train and test folds.

Under cross-source evaluation, the apparent dominance of the length-only baseline collapses from AUC 0.87 to 0.65, indicating that much of its aggregate performance derives from source-distribution artifacts. GrantAgent remains marginally above the single-call baselines, but the separation between methods narrows substantially.

These results suggest that aggregate evaluation alone substantially overestimates evaluator capability in heterogeneous grant-review settings. The following sections therefore focus on controlled within-source and length-aware evaluation.

5.2 Controlled Evaluation Removes the Apparent Gains

We next evaluate whether the apparent aggregate gains survive under controlled settings designed to reduce shortcut correlations.

Length-controlled evaluation. To reduce the influence of proposal length, we partition proposals into word-count quartiles and compute AUC within each quartile. The longest quartile ($\geq$17,313 words) serves as the primary length-controlled setting because proposal length is no longer predictive in this regime.

All single-call baselines collapse in the longest-quartile setting, with AUC dropping to 0.12–0.33, while GrantAgent reaches AUC 0.69 (Table 4). This directional pattern suggests that the multi-agent pipeline may be somewhat more robust to superficial length cues than single-call prompting approaches.

However, the longest quartile contains only four declined proposals, making the comparison substantially underpowered. We therefore interpret the Q4 result as suggestive directional evidence rather than definitive proof of reliable content-quality discrimination.

Within-source evaluation. We next evaluate systems separately within each source containing both funded and declined proposals (Table 5).

The within-source results substantially weaken the apparent aggregate gains. On Open Grants, the only source containing substantial numbers of both funded and declined proposals, all methods perform near chance, including GrantAgent (AUC 0.43). Performance on Wellcome is somewhat higher for several methods, but the subset remains highly imbalanced and structurally homogeneous.

Taken together, the controlled evaluations suggest that much of the apparent success under ag-

Method	AUC	Bal. Acc.	Macro-F1	MCC
Length-only logistic	**0.87**	**0.80**	**0.80**	**0.61**
GrantAgent	0.55	0.56	0.49	0.16
GPT-4 + RAG	0.47	0.49	0.38	−0.03
GPT-4 + structured	0.45	0.53	0.46	0.08
GPT-4 + CoT	0.40	0.52	0.34	0.09
GPT-4 single-call	0.39	0.51	0.32	0.05

Table 2: Aggregate evaluation on the full corpus.

Method	AUC	Bal. Acc.	Macro-F1	MCC
GrantAgent	**0.65**	0.56	0.38	0.10
Length-only logistic	0.65	0.51	0.47	0.01
GPT-4 + structured	0.62	0.50	0.30	0.01
GPT-4 single-call	0.60	0.51	0.18	0.00
GPT-4 + RAG	0.58	0.53	0.43	0.04
GPT-4 + CoT	0.52	0.52	0.33	0.06

Table 3: Cross-source GroupKFold evaluation.

Method	Q1	Q2	Q3	Q4	Overall
Length-only	0.82	0.54	0.65	0.50	0.87
GPT-4 single call	0.46	0.54	0.49	0.33	0.39
GPT-4 + struct.	0.52	0.48	0.50	0.31	0.45
GPT-4 + RAG	0.51	0.55	0.58	0.20	0.47
GPT-4 + CoT	0.45	0.56	0.58	0.12	0.40
GrantAgent	0.45	0.52	0.42	0.69	0.55

Table 4: Within-quartile ROC-AUC by proposal length.

gregate evaluation derives from exploitable correlations tied to proposal length and source-specific formatting conventions rather than robust proposal-quality discrimination.

5.3 Cross-Model Analysis

To test whether the observed failures are specific to a particular model family, we evaluate the GrantAgent pipeline across four LLMs spanning OpenAI and Anthropic models: GPT-4o-mini, GPT-4o, Claude Sonnet, and Claude Opus.

Because Anthropic models frequently refused biomedical proposal evaluation, we restrict this analysis to a controlled 60-proposal non-biomedical subset containing balanced funded and declined examples.

Across all four models, confidence intervals include chance performance (Table 6). Although larger models produce slightly higher point estimates than GPT-4o-mini, none demonstrates statistically reliable discrimination under controlled evaluation.

We additionally find that replacing GPT-4o-mini with GPT-4o in the full GrantAgent pipeline modestly improves aggregate AUC but does not improve within-source performance. On Open Grants, both models remain near or below chance.

Taken together, the cross-model experiments suggest that the observed failures are not specific to a particular prompting strategy, model scale, or provider family. Across all tested configurations, apparent aggregate gains weaken substantially once source and formatting confounds are controlled.

5.4 Calibration and Triage Utility

Although the multi-agent system fails to demonstrate reliable within-source grant-quality discrimination, we observe somewhat stronger calibration behavior than the single-call baselines (Table 7).

On aggregate evaluation, GrantAgent achieves the lowest expected calibration error (ECE) among the evaluated LLM systems, suggesting that the multi-agent aggregation process produces smoother probability estimates than single-call prompting approaches. However, this advantage weakens under within-source evaluation, where several single-call baselines achieve comparable calibration under overlapping bootstrap confidence intervals.

In addition to scalar probabilities, GrantAgent produces criterion-specific rationales and interpretable review flags. While these signals do not reliably separate funded from declined proposals at the corpus level, they may still provide limited utility in human-in-the-loop triage workflows.

Overall, the results suggest a narrower role for current multi-agent LLM review systems: uncertainty-aware decision support rather than reliable autonomous evaluation.

Source	GrantAgent	Length	GPT-4	Struct	RAG	CoT
Open Grants	0.43	0.52	0.55	0.49	0.54	0.49
Wellcome	0.63	0.60	0.55	0.71	0.61	0.62

Table 5: Within-source ROC-AUC.

Model	AUC	95% CI	Relative Cost
GPT-4o-mini	0.51	[0.35, 0.66]	1×
GPT-4o	0.57	[0.41, 0.72]	∼5×
Claude Sonnet	0.53	[0.38, 0.68]	∼6×
Claude Opus	0.55	[0.40, 0.71]	∼30×

Table 6: Cross-model evaluation on the controlled non-biomedical subset.

Method	ECE	Brier Score
GrantAgent	**0.09**	0.25
Length-only logistic	0.14	**0.17**
GPT-4 + structured	0.21	0.28
GPT-4 + RAG	0.25	0.31
GPT-4 single-call	0.28	0.32
GPT-4 + CoT	0.29	0.34

Table 7: Aggregate calibration metrics. Lower is better.

6 Conclusion

We presented a controlled evaluation study of multi-agent LLM systems for grant proposal review. Using a representative architecture combining criterion decomposition, retrieval augmentation, adversarial debate, and meta-review aggregation, we tested whether commonly proposed multi-agent evaluation strategies improve grant-review capability beyond single-call prompting approaches.

Our central finding is negative. Although aggregate evaluation initially suggests modest gains for the multi-agent system, these apparent improvements weaken substantially once proposal length and source-distribution confounds are controlled. Under within-source evaluation, no tested model or model family demonstrates statistically reliable grant-quality discrimination.

The results suggest that current LLM-based grant-review systems are highly sensitive to shortcut signals tied to formatting, proposal length, and source structure. More broadly, the paper demonstrates that aggregate evaluation on heterogeneous review corpora can substantially overestimate evaluator capability. Controlled within-source and length-aware evaluation are therefore necessary for assessing whether LLM systems recover genuine content-level signals.

At the same time, our experiments indicate that multi-agent systems may still provide narrower utility as assistive tools. Compared to single-call baselines, the multi-agent pipeline produces comparatively stable calibration behavior and interpretable criterion-level outputs that may support uncertainty-aware human triage workflows. The evidence is therefore more consistent with decision-support augmentation than with reliable autonomous review.

More broadly, we view this work as a case study in evaluator robustness rather than a grant-review-specific benchmark. The core lesson extends beyond funding applications: sophisticated multi-agent architectures can appear successful under aggregate evaluation while failing to demonstrate stable within-distribution discrimination once confounds are controlled. As LLM-based evaluators become increasingly common across scientific and administrative domains, evaluation methodology may matter as much as architecture design itself.

Reproducibility. Code, prompts, evaluation scripts, and analysis notebooks will be released upon acceptance. The evaluation corpus is rebuildable from publicly redistributable or voluntarily shared sources using the provided data manifests and reconstruction scripts.

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.

Limitations

Several limitations constrain the interpretation of our results.

First, the evaluation corpus remains heterogeneous and partially imbalanced despite our controlled evaluation protocol. Although this heterogeneity is central to the paper's analysis of confounded evaluation, it also limits statistical power for several within-source comparisons. In particular, the longest-quartile setting contains relatively

few declined proposals, making some subgroup analyses underpowered despite large directional differences.

Second, the corpus combines proposals originating from different funders, review cultures, and submission formats. While we explicitly study this heterogeneity as a source confound, the resulting labels may still fail to represent a single coherent notion of proposal quality. Funding decisions reflect not only scientific merit, but also programmatic fit, reviewer calibration, institutional context, and panel-level dynamics that are not observable from proposal text alone.

Third, our study evaluates only a limited set of architectural strategies. Although we examine multiple prompting styles, retrieval augmentation, adversarial debate, model families, and model scales, these experiments do not exhaust the space of possible multi-agent designs. Future systems incorporating stronger external grounding, reviewer-history modeling, or explicit panel-dynamics simulation may behave differently.

Fourth, several analyses rely on open and voluntarily shared proposals. This introduces potential selection effects, particularly among declined proposals, which may differ systematically from the broader population of unfunded submissions. We therefore caution against interpreting the present corpus as fully representative of real-world funding distributions.

Fifth, we study only prompting-based and multi-agent LLM evaluation systems rather than supervised discriminative approaches trained directly on proposal labels or embeddings. It remains possible that task-specific discriminative models could recover signals not accessible through zero-shot or agentic prompting strategies. Our conclusions should therefore be interpreted narrowly: under the prompting-based evaluation settings studied here, current multi-agent LLM systems do not demonstrate robust within-source grant-quality discrimination.

Finally, our conclusions are limited to the task setting studied here: heterogeneous open grant-review corpora under controlled within-source evaluation. We do not claim that LLMs are universally incapable of scientific evaluation or peer review. Rather, our findings indicate that current multi-agent LLM systems do not reliably recover robust grant-quality discrimination under the tested conditions.

7 Ethical Considerations

This paper studies LLM-based grant-review systems in a high-stakes evaluation setting involving scientific funding decisions. Automated or semi-automated review systems carry substantial ethical risks, including opacity, unfairness, amplification of institutional biases, and over-reliance on unreliable model outputs. These risks are especially serious in grant review, where funding decisions can significantly affect scientific careers and research directions.

Our findings argue against deploying current multi-agent LLM systems as autonomous grant evaluators. Although aggregate metrics may suggest moderate performance, controlled within-source evaluation reveals that current systems do not reliably discriminate proposal quality under the tested conditions. We therefore caution against interpreting aggregate benchmark performance as evidence of deployable review capability.

A central ethical concern raised by this work is the danger of confounded evaluation. Models can appear effective by exploiting superficial correlations tied to proposal length, formatting conventions, or source distributions rather than scientific merit. In high-stakes domains, such shortcut behavior may create misleading impressions of reliability while obscuring systematic failure modes.

At the same time, our results suggest that limited assistive uses may still be possible. Calibrated uncertainty estimates, criterion-level summaries, and interpretable review flags may help support human reviewers in triage-oriented workflows when used cautiously and under human oversight. We emphasize that such systems should function as decision-support tools rather than replacements for expert review.

The evaluation corpus used in this work consists of publicly available or voluntarily shared proposals. **Prior to evaluation, all document data was thoroughly audited to ensure the removal of sensitive personally identifying information (PII),** and we do not release private reviewer identities or confidential panel discussions. Nevertheless, the broader problem of automated grant evaluation raises important questions about consent, reviewer transparency, and the governance of AI-assisted funding systems. We encourage future work to prioritize controlled evaluation, transparency, and human-centered oversight when studying LLM-based evaluators in scientific and administrative

domains.

Data Privacy & API Compliance: All experiments were conducted in strict compliance with the commercial developer terms of use for the types of OpenAI and Anthropic APIs utilized. To protect the intellectual property within the open proposal corpus, all data routing utilized standard enterprise endpoints, ensuring that proposal text remained confidential and was explicitly exempted from provider model-retraining pipelines.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Stephen A Gallo, Joanne H Sullivan, and Scott R Glisson. 2016. The influence of peer reviewer expertise on the evaluation of research funding applications. *PloS one*, 11(10):e0165147.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2021–2031.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024a. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, and 1 others. 2024b. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. Accessed: May 2026.

OpenAI. 2024b. Hello gpt-4o. Accessed: May 2026.

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.

Python Software Foundation. 2021. Python language reference, version 3.10.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing cognitive synergy in large language models: Atask-solving agent through multipersona self-collaboration. *arXiv preprint arXiv*, 2307.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Model Size and Computational Budget

This appendix provides the model versions, API snapshots, and computational budget used throughout the experiments, following the ARR / EMNLP 2026 reproducibility checklist requirements.

A.1 Model Parameters

All models used in this paper are closed-source proprietary LLMs accessed via API; the providers do not publicly disclose parameter counts. We therefore report the exact dated snapshot versions used for reproducibility (Table 8).

Snapshots are pinned at module load in the supplementary implementation files to ensure reproducibility. We use dated snapshots rather than rolling aliases (e.g., `gpt-4o`) because provider-side rolling aliases may silently change over time, potentially altering evaluation outputs.

A.2 Computational Budget

All inference was performed through commercial APIs. We therefore report token-based API expenditure rather than GPU-hours, which are not directly applicable to hosted API inference (Table 9).

Local analysis compute. All local statistical analysis was CPU-based and required no GPU acceleration. Bootstrap confidence intervals (2,000 iterations) and permutation-null experiments (1,000 iterations) were implemented using standard `scikit-learn` and `numpy` pipelines.

On the filtered evaluation subset ($N = 148$ after the $\geq$3-per-class filter), the full transfer-matrix pipeline executes in under 90 seconds on a single CPU core.

B Data Governance

All proposal materials were used exclusively for non-commercial research and evaluation purposes consistent with their original public availability or voluntary sharing conditions. We do not redistribute proposal text directly. Released artifacts are limited to reconstruction scripts, metadata manifests, prompts, and evaluation code intended solely for research reproducibility and methodological analysis.

Role in Paper	Model	Snapshot
GPT-4 single-call, structured, RAG, and CoT baselines	OpenAI GPT-4o	`gpt-4o-2024-08-06`
Multi-agent system (primary configuration)	OpenAI GPT-4o-mini	`gpt-4o-mini-2024-07-18`
Multi-agent system (high-budget configuration)	OpenAI GPT-4o	`gpt-4o-2024-08-06`
Cross-LLM panel — Sonnet	Anthropic Claude Sonnet 4.6	`claude-sonnet-4-6`
Cross-LLM panel — Opus	Anthropic Claude Opus 4.7	`claude-opus-4-7`

Table 8: LLM models and pinned snapshot versions used in the experiments.

System	Per-review Cost	N	Total Cost
Random / length-only baselines	$0.00	455	$0
GPT-4 single-call	~$0.01	455	~$5
GPT-4 + structured prompting	~$0.01	455	~$5
GPT-4 + RAG	~$0.02	455	~$10
GPT-4 + CoT	~$0.02	455	~$10
Multi-agent + GPT-4o-mini	$0.50–1.00	455	~$230–460
Multi-agent + GPT-4o	$2.50–5.00	~120	~$300–600
Cross-LLM panel	$0.50–30 per review	240 API calls	~$500–1,500
Total project API spend			~$1,000–2,500 USD

Table 9: Approximate API inference costs across all experiments.