
Proceedings Paper:
Thorne, W., James, J., Wang, Y. et al. (2026) Evaluating LLM-based grant proposal review via structured perturbations. In: Brooker, S., Benatti, F., Pisarski, M. and Adamou, A., (eds.) HT '26: Proceedings of the 37th ACM Conference on Hypertext. HT '26: 37th ACM Conference on Hypertext, 14-18 Sep 2026, London, United Kingdom. ACM, New York, pp. 327-341. ISBN: 9798400725647.

https://doi.org/10.1145/3800935.3830838

Evaluating LLM-Based Grant Proposal Review via Structured Perturbations

William Thorne
Computer Science
University of Sheffield
Sheffield, United Kingdom
wthorne1@sheffield.ac.uk

Joseph James
University of Sheffield
Sheffield, United Kingdom
jhfjames1@sheffield.ac.uk

Yang Wang
University of Manchester
Manchester, United Kingdom
yang.wang-2@manchester.ac.uk

Chenghua Lin
University of Manchester
Manchester, United Kingdom
chenghua.lin@manchester.ac.uk

Diana Maynard
University of Sheffield
Sheffield, United Kingdom
d.maynard@sheffield.ac.uk

Abstract

As AI-assisted grant proposals outpace manual review capacity in a kind of "Malthusian trap" for the research ecosystem, this paper investigates the capabilities and limitations of LLM-based grant reviewing for high-stakes evaluation. Using six EPSRC proposals, we develop a perturbation-based framework probing LLM sensitivity across six quality axes: funding, timeline, competency, alignment, clarity, and impact. We compare three review architectures: single-pass review, section-by-section analysis, and a 'Council of Personas' ensemble emulating expert panels. The section-level approach significantly outperforms alternatives in both detection rate and scoring reliability, while the computationally expensive council method performs no better than baseline. Detection varies substantially by perturbation type, with alignment issues readily identified but clarity flaws largely missed by all systems. Human evaluation shows LLM feedback is largely valid but skewed toward compliance checking over holistic assessment. We conclude that current LLMs may provide supplementary value within EPSRC review but exhibit high variability and misaligned review priorities. We release our code and any non-protected data[1].

CCS Concepts

• **Computing methodologies** → **Natural language generation**; • **Social and professional topics**; • **Applied computing** → **Computing in government**; **Decision analysis**;

Keywords

Engineering & Physical Sciences Research Council (EPSRC), Research Grant Reviewing, Automatic Reviewing, Applications of LLMs, Data Augmentation, Perturbation-based Evaluation, Multi-agent Systems, Peer Review Automation, Document Quality Assessment, Human-AI Alignment

[1]We release our code and annotation guidelines here: https://github.com/wrmthorne/grant-perturbation-analysis

HT '26, London, United Kingdom

ACM ISBN 979-8-4007-2564-7/26/09
https://doi.org/10.1145/3800935.3830838

ACM Reference Format:
William Thorne, Joseph James, Yang Wang, Chenghua Lin, and Diana Maynard. 2026. Evaluating LLM-Based Grant Proposal Review via Structured Perturbations. In *37th ACM Conference on Hypertext (HT '26), September 14–18, 2026, London, United Kingdom.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3800935.3830838

1 Introduction

Peer review is the foundational mechanism for ensuring scientific rigour and directing funding towards impactful, feasible research. However, the global research ecosystem is currently caught in a *"Malthusian trap"* [20]: while R&D funding has seen incremental increases, the volume of applications has grown exponentially. In the UK, the number of competitive grant applications assessed by UK Research and Innovation (UKRI) has nearly doubled since 2017, while the overall award rate has plummeted from 36% to 19% [27]. This surge has placed the research ecosystem under unprecedented strain, resulting in systemic reviewer fatigue and a rising administrative burden [33]. Consequently, the peer-review process has seen significantly extended decision cycles, with the end-to-end timeline for major funding schemes now frequently exceeding 18 months [27]. This pressure is exacerbated by a growing *dual-standard* in generative AI (GenAI) policy [26]. While policy is increasingly permissive for applicants, allowing GenAI for brainstorming, structuring, and language editing, it remains strictly prohibited for reviewers. Allowing applicants but not reviewers to use LLMs creates an asymmetry that risks either lower review quality or longer funding timelines.

Grant evaluation also presents challenges that distinguish it from the more widely studied conference setting. Unlike paper reviewing, which is retrospective (evaluating completed work), grant reviewing is prospective and administrative, requiring high-stakes assessments of value for money, multi-year project feasibility, and national impact. It further demands: (1) **Contextual Breadth**, assessing diverse documents from financial spreadsheets to impact statements; (2) **Applicant Visibility**, where non-anonymous metadata increases the risk of prestige or institutional bias; and (3) **Decision Stakes**, where errors carry consequences associated with significant capital and multi-year commitments. While recent research has explored LLM capabilities in academic peer review for conferences [1, 16], evaluating their readiness for grants demands

a fundamentally different methodology. Complete research proposals are tightly guarded assets whose contents comprise novel and highly valuable intellectual property, both financially and in terms of participant careers and reputations. The scarcity of data and the high ethical barriers to access are central to why grant proposals and their reviewing are so understudied, despite their importance.

Recent work using LLMs in the grant domain has been primarily applicant-focused, assisting with proposal drafting, literature discovery and alignment with funding criteria [29, 30]. Okasa et al. [23] developed a supervised pipeline on the content of grant review reports at the Swiss National Science Foundation; however, this only analyses existing human reviews instead of synthesising or evaluating them. To our knowledge, no prior work has systematically evaluated LLM capabilities for the review of grant proposals: assessing whether models can identify substantive weaknesses, produce reliable scores or generate feedback comparable to expert reviewers.

Under this premise, we propose **perturbation-based evaluation** as a principled solution to this issue of data scarcity. Rather than creating supervision through the labelling of many proposals, we construct controlled fault conditions from a limited pool of genuine grant submissions and measure LLM review systems' ability to reliably detect known defects. Using six genuine Engineering and Physical Sciences Research Council (EPSRC) proposals, we define a perturbation taxonomy based on six key axes of quality (funding, timeline, competency, alignment, clarity, and impact), decomposing into 42 total perturbations at the most granular level. Each of these may be applied to every proposal to produce variants that reflect known, targeted weaknesses relative to the original. We summarise our contributions as follows: **(1) a perturbation-based evaluation framework** that enables principled, fine-grained assessment of LLM review systems in data-scarce, high-sensitivity domains, demonstrated here by transforming six proposals into 42 controlled fault conditions across six quality axes; **(2) the development of a Council of Personas architecture** designed to emulate the multi-perspective nature of expert panels; and **(3) a comparative analysis of model-generated feedback** against the nuanced judgements of experienced UKRI reviewers to identify current gaps in automated reasoning and potential sources of bias.

Specifically, we address three exploratory research questions: (**RQ1**) How do different architectural configurations influence the detection of systematic proposal perturbations and the reliability of scoring? (**RQ2**) Which core assessment dimensions are LLM-based systems most sensitive to? (**RQ3**) How does the qualitative feedback from LLMs align with the judgements of experienced UKRI reviewers?

2 Related Work

Research on LLMs in academic assessment spans autonomous review generation and reviewer assistance. While models reliably catch surface-level reproducibility and formatting issues [12, 21], they struggle with nuanced methodological flaws and novelty detection [7, 38]. Evidence on review quality is mixed: large-scale work finds that GPT-4 overlaps with human consensus about as much as humans do with each other, though its qualitative feedback remains weak [16], and systems like *Reviewer3* show only moderate human agreement [28]; other comparisons report low agreement, with models failing to adapt to a venue's specific focus [5]. A 2022 report for the UK HE funding bodies [32] likewise concluded that AI tools could not yet inform scoring decisions in research assessment, while recommending further pilot testing. These conclusions predate current reasoning models, and our work provides an updated empirical assessment of whether their potential has been realised.

Grant assessment poses distinct challenges. Although proposals and publications are both judged on rigour and novelty, the prospective nature of grants shifts the focus toward *justification* and *feasibility* [9], requiring alignment with a specific funding call and a demonstrated gap in the landscape [34]. Prior NLP work here has largely addressed classification of reviewer reports [23] or funding-trend analysis rather than the generative and critical capabilities of LLMs in the review loop. The known limitations of LLMs–summarising content well but missing methodological weaknesses, feasibility, and deep novelty [7, 38]–map directly onto grant criteria such as team competency and resource justification, which demand a holistic synthesis beyond surface-level pattern matching.

2.1 Evaluation via Perturbation

Perturbation-based evaluations have emerged as a standard methodology for probing LLM robustness and the reliability of the "LLM-as-a-judge" paradigm [4, 10]. By systematically altering input text while maintaining core semantic properties, researchers can identify specific model biases and failure modes that remain hidden during standard benchmarking. While foundational frameworks like TextAttack [19] focus primarily on lexical and syntactic transformations, such as word substitutions or character-level noise, recent benchmarks have expanded this to evaluate model consistency across diverse document formats and enterprise-scale contexts [2].

In this work, we extend the scope of perturbation from linguistic variations to domain-specific structural inconsistencies. Unlike general-purpose benchmarks, we target the logical "fatal flaws" unique to the grant domain, such as budget-timeline misalignments or competency gaps, to test whether LLMs can move beyond pattern matching toward the rigorous, high-stakes reasoning required for professional research assessment.

3 Methodology

Research grant proposals are highly sensitive assets, containing both proprietary intellectual property and confidential data of the participants. Due to these privacy constraints and the high ethical burden associated with their processing, the systematic study of grant reviewing remains significantly under-researched. To address these challenges, we evaluate the capability of offline, locally-served LLMs to review proposals submitted to EPSRC. Our methodology maximises the utility of a limited dataset through a perturbation-based approach: we begin with a set of contemporary, human-authored proposals, submitted to EPSRC, and systematically degrade their quality. We deconstruct the notion of proposal quality along six axes: *funding*, *timeline*, *competency*, *alignment*, *clarity* and *quality*, and *impact*, which we derive from the UKRI assessment process itself. Through this approach, we demonstrate that this (albeit limited) dataset can nevertheless be augmented into a robust

benchmark to test the sensitivity, consistency and capabilities of current LLMs for EPSRC proposal reviewing.

3.1 Review Frameworks

3.1.1 Zero-shot Baseline. The baseline system, **GPT-OSS-20B (*high*)** [24], is provided with a zero-shot task description, the official UKRI review guidelines, the specific funding opportunity, and the complete proposal narrative in a single context. The model is tasked with producing an overall score (1-6), following the official EPSRC reviewer scoring scale, and a set of comments justifying this score and providing feedback.

3.1.2 Section-Level Review Framework. Input prompts to the baseline approach often exceed 30,000 tokens. While modern LLMs demonstrate high accuracy in simple information retrieval at these lengths, their ability to perform complex reasoning and synthesis decays significantly as context scales [11, 17]. This performance gap, where models struggle to apply a fact to a critical evaluation, is particularly acute in document-level assessment [35]. To mitigate this, we implement a section-level review process, reducing the cognitive load per inference pass encouraging more specific feedback [36].

We created four logical groups from the sections of the proposal documents, as many sections provide little value to review in isolation (e.g. references) but provide valuable context when accompanying others. The groups are: **Vision-Approach** (vision and approach, references); **Team Capability** (summary, applicant and team capability to deliver, core team, project partners, facilities, references); **Funding Resources** (summary, resources and costs, core team, facilities, references); and **Ethics** (summary, ethics and responsible research and innovation, research involving human participants). We discard purely administrative sections (EPSRC thematic area alignment, letters of support) and sections which were not applicable to any of our reviews (e.g. animal testing, sensitive information).

3.1.3 Council of Personas. The baseline and section level approaches risk propagating single perspective biases and linear errors into the final feedback. To address this, we employ a **Council of Personas**[2], implementing majority-voting for qualitative reasoning via a three-stage process: (1) independent persona-based reviews; (2) blind meta-review and ranking of peer councillor outputs; and (3) final synthesis by a council chair who down-weights anomalous comments based on these rankings.

We use five different personas to encourage feedback diversity: **Cost Analyst, Ethics Assessor, Tech Evangelist, Methodological Sceptic,** and **Impact Champion**. Each persona introduces a deliberate bias: for instance, the Methodological Sceptic prioritises soundness and validity, while the Impact Champion focuses on scalability and industry engagement. Full persona descriptions and prompting templates are provided in **Appendix A.1**. This collective approach ensures that while individual personas may be over-sensitive to specific "fatal flaws," the final output remains holistic and aligned with standard UKRI guidance.

[2]We use Andrej Karpathy's repo as reference https://github.com/karpathy/llm-council.

3.2 Data

Our dataset comprises six full EPSRC funding applications from the School of Computer Science, obtained through collaboration with our institutional research hub. Of these, two proposals were successfully funded, one was unfunded, and the remaining three are awaiting a decision; one funded and one unfunded proposal are accompanied by full expert-review comments and scores. The original submission dates range from May 2023 to August 2025; working with recent proposals mitigates data leakage risks, as it is unlikely this specific content appeared in the pre-training data of current models. Furthermore, as unpublished grant proposals, these documents are not publicly accessible and could not have been included in the training corpora of the models under evaluation.

Assessment Context. Standard UKRI assessment involves evaluation by 3-4 expert reviewers who score applications on a scale of 1-6 across four primary pillars: *Research Excellence*, *National Importance*, *Applicant Track Record*, and *Resources and Management*. Consequently, our analysis focuses on the *Vision and Approach*, *Team Capability to Deliver*, *Justification of Resources*, and *Ethics* sections, as these contain the primary claims where "fatal flaws" in feasibility typically appear.

Perturbation Strategy. To evaluate model sensitivity across the core dimensions of the UKRI assessment process, we systematically degrade proposals along six primary axes. Table 1 summarises the perturbation strategies and the evaluation criteria they target.

Our strategy was informed by an initial round of human evaluation. Four members of the EPSRC review college, each with extensive grant-reviewing experience, were given either the vision or the approach section and asked to score it from one to six following the standard UKRI rubric, justifying their score with highlighted excerpts and discursive positive and negative feedback. All were within computer science but not necessarily domain experts on each proposal.

Three themes emerged that map onto our axes. Evaluators found referencing to be a general strength but often insufficiently developed to make limitations, motivations and novelty explicit, while technical sections could become overly dense and hard to follow (e.g. unexplained acronyms and missing background). Clear organisation, explicit sectioning, named references and worked examples improved accessibility. These findings inform our clarity axes, capturing weaknesses in framing, evidential sufficiency and academic rigour. Second, evaluators noted disconnects within sections: claims of timeliness were not always supported by evidence, the links between timeliness, impact and motivation were sometimes fragmented, and understanding motivation or feasibility often required speculative reasoning (e.g. absent preliminary studies). These inform our axes of timeline and impact, targeting feasibility, justification and contribution. Finally, evaluators stressed alignment between resources, expertise and positioning: proposals appeared weaker when components did not reinforce one another or when justification in one section went unsupported elsewhere, reducing overall credibility even when individual sections were strong. These concerns inform our funding, competency and alignment axes.

Axis	Perturbation Strategy	Targeted Criterion
Funding	Inflating/lowering budgets; removing cost justifications; misaligning resource allocation with project priorities.	Value for money; compliance with UKRI financial policy.
Timeline	Extending periods beyond call limits; unrealistic task compression; misaligning milestones with logical work progression.	Project feasibility; operational justification.
Competency	Removing/replacing key personnel; weakening evidence of technical skills in the *team capability* section.	Demonstration of requisite skills and leadership.
Alignment	Modifying opportunity aims; switching "What we're looking for" sections; introducing cross-disciplinary mandates.	Strategic fit to call; adherence to funding body values.
Clarity	Removing acronym expansions; introducing vagueness in methods; removing novelty markers; deleting factual background.	Accessibility; technical comprehensibility; academic rigour; specificity.
Impact	Replacing key stakeholders with irrelevant parties; modifying outcome scope; removing long-term/short-term outcomes.	Contribution to field; stakeholder engagement.

Table 1: Summary of perturbation axes, strategies, and their mapping to EPSRC assessment criteria. Examples and further details can be found in Appendix A.7 and A.8.

4 Experimental Setup

4.1 Evaluation Tasks

We evaluate our systems using two primary tasks. To simulate secure deployment and satisfy data privacy requirements, all experiments were conducted on isolated single-GPU devices without external access.

Perturbation Identification. We assess the sensitivity of LLM review systems to different aspects of quality by introducing repeatable and independent perturbations into our human-authored grant proposals. Each perturbation was designed to have a solely negative impact on one of the six axes of quality. Sensitivity is measured by observing the proportion of review comments that negatively address the perturbation (e.g. no explanation of abbreviations) or an obvious direct consequence of it (e.g. the section is unclear as none of the methods are explained). Responses are scored as correct only if the perturbation or consequence was both mentioned and with negative sentiment. In cases where the comment partially addresses the perturbation, a consequence could have arisen by multiple means or in any other cases of ambiguity, partial credit is assigned. Contradictory cases of clear identification but positive sentiment are deemed incorrect. This task allows us to explore how architectural choices affect detection performance, which assessment pillars LLMs can and cannot identify, and whether their feedback aligns with expert judgment.

To scale the analysis to all 42 perturbations across all six proposals and in each review setting, we employed a panel of three judge models. The judge first engages in a reasoning phase before responding with one of Correct (C), Partial (P) or Incorrect (I). Providing this list of differences as context mitigates any misinterpretation from our descriptions, while not polluting the context. The prompt template can be found in A.3, further details of the setup in A.4.1, and validation of the judge panel in A.4.2.

Expert–Model Feedback Alignment. Two of our proposals came each with four expert reviews. We decompose these reviews, and those generated by each LLM review system into atomic claims using GPT-OSS-120B (high) that each express a single evaluative aspect. Following a manual assessment, we observe no direct contradictions between human reviewers for each proposal. We define a contradiction to be a factual discrepancies or opposing valence such that two claims cannot both be true at once [14]. LLM claims that do *not* appear in the set of expert claims are subsequently human annotated. Full details of the claim decomposition method are presented in Appendix A.6.

	Baseline			Section-Level			Council		
Perturbation Axis	C	P	I	C	P	I	C	P	I
Funding (n=1688)	159	20	440	220	40	358	115	51	285
Competency (n=889)	28	12	353	32	10	240	15	7	192
Alignment (n=953)	122	2	222	183	1	163	84	0	176
Clarity (n=1859)	6	0	668	99	5	571	12	4	494
Impact (n=639)	10	5	221	71	5	157	7	0	163
Timeline (n=1088)	85	0	314	106	3	289	40	2	249
Overall (n=7116)	410	39	2218	711	64	1778	273	64	1559

Table 2: Perturbation identification rates across review systems. C = Correct (perturbation identified with negative sentiment), P = Partial (ambiguous or indirect identification), I = Incorrect (missed or positive sentiment). Verdicts are determined by majority vote across three independent judge models.

4.2 Metrics

We employ the following metrics to evaluate review system performance:

Perturbation Detection Score. A numerical mapping from the judge's verdict: Correct = 1.0, Partial = 0.5, Incorrect = 0.0. This enables continuous analysis of detection performance.

Score Degradation. The signed difference $\Delta S = S_{\text{original}} - S_{\text{perturbed}}$ between scores assigned to original and perturbed proposals. Positive values indicate appropriate score reduction; negative values flag anomalous cases where perturbation *increased* the score.

Intra-Class Correlation (ICC). We use ICC(2,1) to assess scoring reliability, treating both proposals and evaluation runs as random effects. This two-way random effects model quantifies the proportion of variance attributable to true differences between proposals [13].

5 Results and Discussion

5.1 Perturbation Identification

We evaluated 42 unique perturbations across 6 proposals using 3 review systems (7,347 perturbed observations total). The overall detection rate was 21.2%; nearly four in five perturbations go undetected.

System Comparison. Table 2 presents detection rates, aggregated by perturbation type. The section-level review system achieved the highest detection scores across nearly all categories ($\mu = 0.29$), followed by the baseline ($\mu = 0.17$) and council approach ($\mu = 0.17$). We confirm that the performance difference between systems is significant via a Kruskal-Wallis[3] test ($H = 27.62, p < 0.0001$) [15].

[3]We note that our scores are ordinal and subsequently mapped to floats

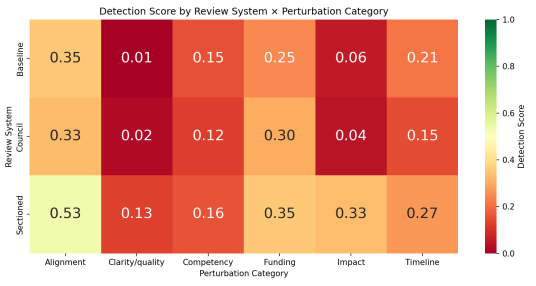

Figure 1: Detection scores across review systems and perturbation categories. Darker cells indicate higher detection rates. The section-level system shows strongest performance on alignment and impact perturbations, while all systems fail on clarity-based changes.

Through a pairwise comparison of the systems, we find the baseline and council methods produce scores that are statistically indistinguishable ($p = 0.83$); this is an especially poor result for the council setup given the token cost. Conversely, we observe that the section-level system consistently prescribes lower scores than the other two with a mean difference of around 1.2 points ($p < 10^{-46}$). This could be seen as miscalibration or an overly harsh reviewer; however, the high detection rate suggests a more critical and accurate reviewer.

Perturbation Sensitivity. Detection rates varied greatly across different perturbation categories. Perturbations to the alignment were most detectable ($\mu = 0.41$), particularly the *cross-cutting theme injections* ($\mu = 0.70$); however, We note that the alignment perturbations were performed on the opportunity documents rather than the proposals themselves. We believe that because many opportunity documents likely appear in models' pre-training data, they have learned the typical structure and conditions of opportunity notices. Deviations from these patterns may appear more salient than comparable changes within proposals which have never been seen. Clarity perturbations, on the other hand, went almost entirely undetected ($\mu = 0.06$); *acronym-related* changes and *connective removal* were never identified. Figure 1 demonstrates the interaction between review system and perturbation type. We attribute the consistently poor performance on clarity in part to the subtlety of these perturbations; however, we believe the LLM review systems rely on contextual inference to resolve ambiguous terminology or acronyms rather than flagging them as missing definitions or quality concerns. While this should be a strength, we observe an over-reliance on this ability, with LLM review systems failing to question random abbreviations in the text.

Reliability. We decompose the total variance in scores to assess the consistency of each review system (Table 3). The section-level approach achieves the highest intra-class correlation (ICC = 0.50), indicating that approximately half of the observed variance reflects true differences between proposal versions rather than noise. In contrast, both the baseline (ICC = 0.14) and council (ICC = 0.11) systems exhibit substantially higher within-sample variance, meaning repeated evaluations of the same proposal yield inconsistent scores. The council's poor reliability is particularly notable given its significantly higher computational cost (see Appendix A.10);

	σ^2_{total}	$\sigma^2_{\text{between}}$	σ^2_{within}	ICC (↑)
Baseline	0.87	0.13	0.80	0.14
Council	0.91	0.11	0.88	0.11
Section-Level	0.88	0.49	0.49	**0.50**

Table 3: Variance decomposition across review systems. ICC (intra-class correlation) measures the proportion of variance attributable to true differences between proposals versus noise from repeated evaluation.

	Human		Baseline		Council		Section-Level	
Category	**Ret**	**Ctr**	**Ret**	**Ctr**	**Ret**	**Ctr**	**Ret**	**Ctr**
Alignment	79.6%	2.8%	68.2%	0.0%	71.8%	1.9%	80.0%	0.0%
Competency	67.7%	2.1%	76.8%	1.8%	66.7%	1.6%	52.2%	0.0%
Ethics	97.7%	0.0%	95.0%	0.0%	94.0%	1.2%	90.0%	0.0%
Funding	81.3%	8.0%	84.0%	0.0%	89.5%	4.8%	100.0%	0.0%
Impact	77.8%	1.2%	85.3%	0.0%	73.8%	1.0%	66.7%	0.0%
Clarity	93.5%	1.6%	98.5%	0.0%	83.9%	1.9%	86.2%	6.9%
Timeline	96.7%	0.0%	100.0%	0.0%	84.6%	2.6%	85.7%	14.3%
TOTAL	**82.2%**	**2.2%**	**86.5%**	**0.4%**	**78.2%**	**2.1%**	**77.4%**	**2.8%**

Table 4: Retained (Ret) and Contradiction (Ctr) Rates of claims. Retained rate is computed from the exclusive set, where consensus claims are removed, isolating claims unique to each reviewer. Higher retained percentages indicate greater reviewer-specific contribution (unique claims), whereas lower values suggest substantial overlap.

the multi-persona architecture does not translate into more stable assessments. These findings suggest that decomposing the review task into focused sections yields more reproducible judgments than either holistic processing or ensemble-based approaches.

5.2 Expert–Model Feedback Alignment

Human annotation. For the validity label, annotators were asked to label each claim, originating from human reviewers or LLM systems, as either valid or invalid. As this label functions as a quality check, the majority of claims are expected to be valid, resulting in a naturally high agreement of 89.5%. In this setting, Fleiss' κ is an unsuitable measure as it penalises agreement that arises from a genuine class imbalance rather than annotator bias [3]. For Agreement and Severity, we see a Fleiss' κ of 0.78 and 0.68, respectively, indicating substantial agreement.

As shown in Figure 2, all systems score above the neutral midpoint for agreement, though human and Council claims show tighter distributions at higher values. For severity, LLM systems tend to prioritise high-severity claims, while human reviewers contribute a broader range, including more low- or "None"-severity observations that provide contextual framing. The Baseline diverges from this pattern, producing a more balanced distribution centred on moderate severity.

Claim analysis. In Table 4, exclusive claim rates are consistently high across all reviewer types, indicating substantial divergence in the issues each raises. Human reviewers show strong exclusivity in Ethics, Clarity, and Timeline, while the baseline exhibits even higher overall exclusivity, particularly in Clarity, Timeline, and Impact, suggesting it frequently introduces points beyond those raised by humans. The council shows slightly lower exclusivity in Competency and Impact, indicating greater overlap, while the

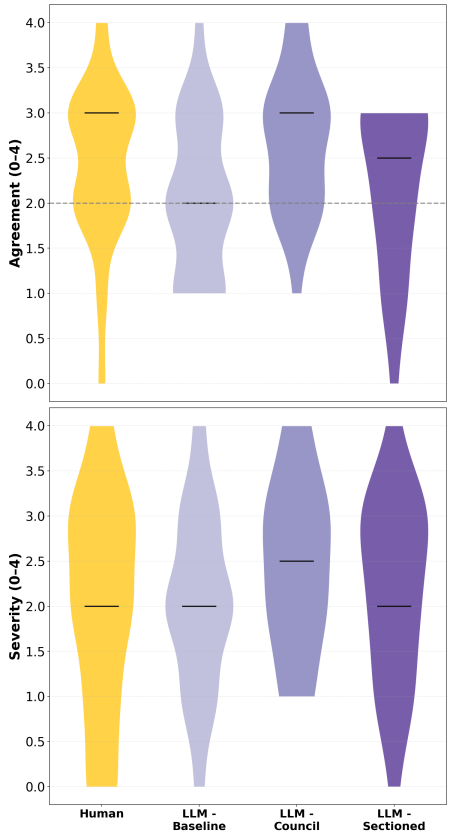

Figure 2: Table 11 shows the scale for severity. The dashed line on agreement indicates the neutral agreement.

	Review	Spearman ρ	Significance
Agreement	Human	0.258	0.014
	Baseline	0.431	< 0.001
	Council	0.732	< 0.001
	Section-Level	0.613	0.006
Severity	Human	0.250	0.018
	Baseline	0.306	0.008
	Council	0.236	0.270
	Section-Level	0.318	0.141

Table 5: Spearman correlations between valence and outcome measures by method.

section-level model displays more variability across dimensions. Critically, claims rarely contradict each other, averaging around 2% of total claims, indicating that unique claims are generally additive rather than conflicting. However, uniqueness alone does not guarantee quality, as a unique claim may still be low severity or disagreeable.

Examining individual claims through Table 7, the most pronounced variance appears in Ethics, where LLM systems raise specific concerns such as data governance, GDPR compliance, and environmental sustainability, whereas human reviewers typically offer broader assessments affirming that RRI considerations are adequate. This suggests LLMs surface granular compliance criteria that human reviewers either overlook or implicitly accept as satisfied earlier in the funding cycle.

	Negative	Neutral	Positive
Human	17.4%	33.1%	49.5%
Baseline	41.7%	20.8%	37.5%
Council	22.2%	25.9%	52.0%
Section-Level	48.1%	12.3%	39.6%

Table 6: Valence distribution (percentage of total claims per source).

Valence Analysis. We examined correlations between valence and human evaluations for both human and LLM claims. Valence was positively correlated with agreement across all systems (Table 5), though the correlation was small for human claims and substantially stronger for LLM systems, particularly the council and section-level variants. These patterns align with the valence distributions in Table 6: human and council claims were predominantly positive, whereas baseline and section-level systems generated more negative claims. Valence-severity correlations were weaker and not statistically significant for the council and section-level systems.

Score distributions (Figure 2) show human and council systems clustering at the upper end of the agreement scale, while baseline and section-level outputs remain more neutral. Human claims exhibit a broader range of severity, reflecting a tendency toward general, affirmatory acknowledgments of proposal content alongside critique rather than focusing exclusively on weaknesses, which accounts for the high variance in severity scores. These findings suggest that while all systems show a positive correlation between valence and agreement, the stronger effect in LLM systems indicates that models are more proficient at generating positive claims than identifying genuine weaknesses, with LLM-generated criticisms consistently less aligned with human judgment.

6 Conclusion

This paper presents an exploratory investigation into the capabilities of LLMs and their potential for use within the EPSRC grant proposal review process. Using a perturbation-based evaluation framework across six axes of grant quality and a human annotation study with members of the EPSRC review college, we show that current LLM systems exhibit highly uneven sensitivity. Alignment perturbations applied to opportunity documents are identified relatively reliably, likely reflecting internalised patterns from pre-training, whereas clarity-based perturbations to proposals are largely missed. These gaps reflect a fundamental difference between grant and paper peer review: where the latter rewards technical depth, grant reviewing demands holistic judgment about whether a proposal merits public investment. Overall, current LLMs show significant limitations for autonomous grant review but may offer value as assistive tools within the process, particularly for structured feedback and alignment checking under human oversight.

Limitations

We evaluate six proposals from a single institution using one model family (GPT-OSS), restricting generalisability across disciplines, funding bodies, and architectures. The scale is sufficient for exploratory analysis but precludes strong statistical claims. The ecological validity of our perturbations also varies: some reflect plausible errors (e.g., budget inflation) while others serve as stress tests (e.g., acronym substitution), so detection rates should be interpreted

as upper bounds on sensitivity. Finally, our human evaluation relies on reviewers from the same institution as the proposal authors, and the per-section annotation design, while ethically necessary, prevents assessment of cross-section coherence.

Ethics Statement

This project has been granted ethical approval by the affiliate institution. All proposals were sourced voluntarily from within the affiliated institution and all handling and processing was conducted on institutional infrastructure. Annotations were conducted also within the institution and in person on an offline device. Annotations were conducted on a per-section level, ensuring any annotator never saw all sections from any single proposal to limit the possibility of plagiarism. All annotators were remunerated at a rate of £25/h.

Acknowledgments

This work was supported by the Arts and Humanities Research Council [grant number AH/X004201/1]. Joseph James was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

[1] AAAI. 2025. AAAI Launches AI-Powered Peer Review Assessment System. https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/ Accessed: 3 December 2025.

[2] Tara Bogavelli, Oluwanifemi Bamgbose, Gabrielle Gauthier Melançon, Fanny Riols, and Roshnee Sharma. 2026. Evaluating Robustness of Large Language Models in Enterprise Applications: Benchmarks for Perturbation Consistency Across Formats and Languages. *arXiv* (2026). doi:10.48550/arXiv.2601.06341

[3] Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of clinical epidemiology* 46, 5 (1993), 423–429.

[4] Manav Chaudhary, Harshit Gupta, Savita Bhat, and Vasudeva Varma. 2024. Towards Understanding the Robustness of LLM-based Evaluations under Perturbations. *arXiv* (2024). doi:10.48550/arXiv.2412.09269

[5] Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. 2023. AI-assisted peer review. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–14.

[6] Deep Search. 2024. *Docling Technical Report.* Technical Report. IBM Research. arXiv:2408.09869 doi:10.48550/arXiv.2408.09869

[7] Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 5081–5099. doi:10.18653/v1/2024.emnlp-main.292

[8] Reinhard Fritsch and Adam Jatowt. 2025. CALLM: A Framework for Systematic Contrastive Analysis of Large Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 6634–6638.

[9] Dan He and DS Parker. 2011. Learning the funding momentum of research projects. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 532–543.

[10] Hanhua Hong, Chenghao Xiao, Yang Wang, Yiqi Liu, Wenge Rong, and Chenghua Lin. 2025. Beyond one-size-fits-all: Inversion learning for highly effective nlg evaluation prompts. *arXiv preprint arXiv:2504.21117* (2025).

[11] Cheng-Ping Hsieh, Simeng Yang, Zeyu Fu, et al. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[12] ICLR. 2025. Assisting ICLR 2025 reviewers with feedback. https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/ Accessed: 3 December 2025.

[13] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (June 2016), 155–163. doi:10.1016/j.jcm.2016.02.012

[14] Venelin Kovatchev, Darina Gold, M. Antonia Marti, Maria Salamo, and Torsten Zesch. 2020. Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 5782–5791. https://aclanthology.org/2020.lrec-1.709

[15] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. http://www.jstor.org/stable/2280779

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

[17] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173. doi:10.1162/tacl_a_00638

[18] Qi Liu, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, and Jiaxin Mao. 2025. E2Rank: Your Text Embedding can Also be an Effective and Efficient Listwise Reranker. arXiv:2510.22733 [cs.CL] https://arxiv.org/abs/2510.22733

[19] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 119–126. doi:10.18653/v1/2020.emnlp-demos.16

[20] Miryam Naddaf. 2025. Is academic research becoming too competitive? Nature examines the data. *Nature* 646, 8087 (2025), 1036–1037.

[21] NeurIPS. 2025. Results of the NeurIPS 2024 Experiment on the Usefulness of LLMs as an Author Checklist Assistant for Scientific Papers. https://blog.neurips.cc/2024/12/10/results-of-the-neurips-2024-experiment-on-the-usefulness-of-llms-as-an-author-checklist-assistant-for-scientific-papers/ Accessed: 3 December 2025.

[22] NVIDIA. 2025. Nemotron 3 Nano: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning. https://arxiv.org/abs/2512.20848 Technical report.

[23] Gabriel Okasa, Alberto de León, Michaela Strinzel, Anne Jorstad, Katrin Milzow, Matthias Egger, and Stefan Müller. 2025. A supervised machine learning approach for assessing grant peer review reports. *Quantitative Science Studies* 6 (2025), 1189–1214.

[24] OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925 [cs.CL] https://arxiv.org/abs/2508.10925

[25] Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5

[26] Daniel D. Reidpath. 2024. UKRI got its A.I. policy half right. Papyrus Walk. https://www.papyruswalk.com/2024/11/ukri-go-its-a-i-policy-half-right/ Accessed December 2025.

[27] Research Professional News. 2025. UKRI grant applications double in seven years as award rate halves. https://www.researchprofessionalnews.com/rr-news-uk-research-councils-2025-8-ukri-grant-applications-double-in-seven-years-as-award-rate-halves/ Accessed: 2026-01-15.

[28] Reviewer3. 2025. Reviewer3. https://reviewer3.com/ Accessed: 3 December 2025.

[29] Krzysztof Rybiński. 2025. Automation of grant application writing with the use of ChatGPT. *Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska* (2025).

[30] Elizabeth Seckel, Brandi Y. Stephens, and Fatima Rodriguez. 2024. Ten Simple Rules to Leverage Large Language Models for Getting Grants. *PLOS Computational Biology* 20, 3 (March 2024), e1011863. doi:10.1371/journal.pcbi.1011863

[31] GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang

Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. 2025. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models. arXiv:2508.06471 [cs.CL] https://arxiv.org/abs/2508.06471

[32] Mike Thelwall, Kayvan Kousha, Mahshid Abdoli, Emma Stuart, Meiko Makita, Paul Wilson, and Jonathan Levitt. 2022. Can REF output quality scores be assigned by AI? Experimental evidence. *arXiv preprint arXiv:2212.08041* (2022).

[33] A Tickell. 2022. Independent review of research bureaucracy. *Recuperado de https://www.gov.uk/government/publications/review-of-research-bureaucracy* (2022).

[34] Anita E Weidmann, Cathal A Cadogan, Daniela Fialová, Ankie Hazen, Martin Henman, Monika Lutters, Betul Okuyan, Vibhu Paudyal, and Francesca Wirth. 2023. How to write a successful grant application: guidance provided by the European Society of Clinical Pharmacy. *International journal of clinical pharmacy* 45, 3 (2023), 781–786.

[35] Kevin Wu, Junxian Huang, Danqi Zhang, et al. 2024. Long-context LLMs Struggle with Long-context Generation. *arXiv preprint arXiv:2402.13718* (2024).

[36] Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is More for Long Document Summary Evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 330–343. doi:10.18653/v1/2024.eacl-short.29

[37] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

[38] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 9340–9351. https://aclanthology.org/2024.lrec-main.816/

A Appendix

A.1 Preprocessing

For each proposal, we obtain relevant contextual data including the target opportunity and, where available, public project data from Gateway to Research (GtR). All textual content is converted to markdown to preserve structural features such as bolding and header nesting. We process PDFs using Docling [6] followed by manual cleaning to reintroduce hyperlink markup.

To handle the critical timeline data often trapped in images, we explored several serialization strategies. While Mermaid and PlantUML render well for humans, their syntax bears little resemblance to the visual timeline. We therefore adopted a markdown table syntax with cells shaded using "####", which proved effective for LLM consumption and facilitated the systematic perturbations described in Section 3.2.

A.2 Multi-Stage Council Prompts

As detailed in Section 3.1.3, the *Council of Personas* follows a three-stage process involving independent review, blind meta-review/ranking, and chair synthesis.

A.2.1 Persona Profiles. The *Council of Personas* uses five distinct roles to simulate a diverse panel of expert reviewers. Each persona is instructed to conduct a holistic review following standard UKRI guidance, but is assigned a deliberate focal bias to ensure comprehensive evaluation across the assessment pillars.

Cost Analyst: Focuses on the financial and administrative feasibility of the project. This persona evaluates budget justifications, spending efficiency, and the proportionality of resource allocation, acting as a critical filter for "value for money".

```
You are financially minded: someone who pays particular
    attention to value for money, resource allocation, and
    cost-effectiveness. While you must still provide a
    comprehensive review covering all aspects, you are
    especially critical of:

- Budget justification and whether costs are reasonable and
    necessary
- Efficient use of resources and personnel
- Whether the proposed outcomes justify the investment
- Risk of cost overruns or inefficient spending
- Whether similar outcomes could be achieved with fewer
    resources
```

Ethics Assessor: Focuses on the societal and ethical implications of the proposed research. Primary concerns include data privacy, environmental sustainability, and responsible research and innovation (RRI).

```
You are ethically minded: someone who emphasizes responsible
     research practices and societal implications. While
    you must still provide a comprehensive review covering
    all aspects, you are especially attentive to:

- Research ethics and responsible innovation
- Data privacy, security, and governance considerations
- Potential societal impacts, both positive and negative
- Inclusivity and equitable access to research benefits
- Environmental sustainability and long-term consequences
```

Tech Evangelist: Focuses on high-risk, high-reward innovation and transformative potential. This persona is incentivized to identify "paradigm shifts" and ambitious technological breakthroughs that may be overlooked by more conservative reviews.

```
You are a tech evangelist: you value innovation, cutting-
    edge approaches, and technological advancement. While
    you must still provide a comprehensive review covering
    all aspects, you are especially excited by:

- Novel technologies and innovative methodologies
- Potential for breakthrough discoveries or transformative
    applications
- Technical sophistication and ambition
- Integration of emerging technologies
- Opportunities to push boundaries and challenge conventions
```

Methodological Sceptic: Focuses on technical soundness, validity, and rigorous experimental design. This persona acts as the primary quality gate, searching for logical inconsistencies or technical "fatal flaws."

```
You are a methodological skeptic who scrutinizes research
     design and scientific rigor. While you must still
     provide a comprehensive review covering all aspects,
     you are especially critical of:

- Methodological soundness and appropriateness
- Validity of proposed approaches and assumptions
- Adequacy of controls, validation strategies, and error
     analysis
- Whether claims are supported by the proposed methods
- Potential confounds, biases, or limitations in the
     research design
```

Impact Champion: Focuses on real-world utility, pathways to impact, and scalability. This persona evaluates how the project engages stakeholders and benefits the broader research and industry landscape.

```
You are an impact champion who focuses on real-world
     applications and broader benefits. While you must still
      provide a comprehensive review covering all aspects,
     you are especially interested in:

- Pathways to impact and how outcomes will be translated
- Engagement with stakeholders, industry, or end-users
- Potential for economic, social, or cultural benefits
- Plans for dissemination and knowledge exchange
- Long-term sustainability and scalability of impacts
```

Chair: Produces the final review based on the strength of arguments made and overall consensus reached during each review and the subsequent meta-reviews by council members.

```
You are a synthesizer who excels at integrating diverse
     expert opinions. You are particularly attuned to:

- When disagreement reflects genuine trade-offs versus
     differences in evidence quality
- The credibility and rigor behind different viewpoints, not
      just their conviction
- Patterns that emerge across independent assessments
- When a minority position raises valid concerns that
     consensus overlooks
- Proportional weighting - giving appropriate influence to
     well-reasoned arguments
- Distinguishing between complementary perspectives and
     genuine contradictions
```

A.2.2 Stage 1: Individual Review. We use the same prompt as the one used for the baseline review system in the initial review stage.

```
For the provided grant proposal, give a score between 1 and
     6 accompanied by a detailed justification for your
     score.

## Score Descriptions

6 - Exceptional: The application is outstanding. It
     addresses all of the assessment criteria and meets them
      to an exceptional level.
5 - Excellent: The application is very high quality. It
     addresses most of the assessment criteria and meets
     them to an excellent level. There are very minor
     weaknesses.
4 - Very good: The application demonstrates considerable
     quality. It meets most of the assessment criteria to a
     high level. There are minor weaknesses.
3 - Good: The application is of good quality. It meets most
     of the assessment criteria to an acceptable level, but
     not across all aspects of the proposed activities.
     There are weaknesses.
2 - Weak: The application is not sufficiently competitive.
     It meets some of the assessment criteria to an adequate
      level. There are, however, significant weaknesses.
1 - Poor: The application is flawed or unsuitable quality
     for funding. It does not meet the assessment criteria
     to an adequate level.

## Review Criteria

We'll only be able to use your review if it meets the
     following criteria:

- you've included enough information to help UKRI staff and
     panellists make an informed judgement on the
     application
- your comments are only based on information that's
     included in the application
- you have not reviewed the application negatively because
     of any equality, diversity and inclusion requirements (
     for example, decisions to work part-time or past
     absences for health reasons)
- your comments are not speculative, inflammatory or
     damaging to applicants
- you have not used journal metrics, conference rankings or
     personal metrics as a substitute measure for assessing
     the applicants' contributions
- you do not have a conflict of interest with the
     application and have not revealed your identity

{proposal_content}

Provide your final assessment as a JSON object with "score"
     (integer 1-6) and "explanation" (string) fields.}


You are evaluating different reviews of the same grant
     proposal.

## Summary

{proposal_summary}

## Reviews

{review_texts}
```

```
Your task:
1. First, evaluate each review individually. For each review
     , explain what it does well and what weaknesses it has
     in its assessment.
2. Then, at the very end of your response, provide a final
     ranking.

IMPORTANT: Your final ranking MUST be formatted EXACTLY as
     follows:
- Start with the line "FINAL RANKING:" (all caps, with colon
     )
- Then list the reviews from best to worst as a numbered
     list
- Each line should be: number, period, space, then ONLY the
     review label (e.g., "1. Review A")
- Do not add any other text or explanations in the ranking
     section

Example format:
Review A provides comprehensive coverage but...
Review B is overly critical on...

FINAL RANKING:
1. Review C
2. Review A
3. Review B

Now provide your evaluation and ranking:



Multiple expert reviewers have provided reviews and then
     ranked each other's assessments. Your task as Chairman
     is to synthesize all of this information into a single
     score (1-6) and explanation. Consider:

- The individual reviews and their insights
- The peer rankings and what they reveal about review
     quality
- Any patterns of agreement or disagreement
- The aggregate rankings showing which perspectives were
     most valued

## Stage 1 - Individual Reviews

{individual_reviews}

## Stage 2 - Peer Rankings

{meta_reviews}

## Aggregate Rankings (Best to Worst)

{aggregation}

Provide your final assessment as a JSON object with "score"
     (integer 1-6) and "explanation" (string) fields.
```

A.3 Perturbation Detection Judge Prompt

```
system: You are an expert evaluator assessing whether an LLM
     -generated grant proposal review correctly identifies a
      known perturbation (intentional flaw) that was
     introduced into the proposal.

You will be given:
1. A description of the perturbation that was applied
2. The exact diff showing what changed between the original
     and perturbed proposal
3. The LLM-generated review of the perturbed proposal

Your task is to determine whether the review identifies the
     perturbation or a direct consequence of it.

user: You are evaluating whether a reviewer identified an
     introduced error in a funding proposal.

## Context

A genuine EPSRC proposal was adversarially modified with the
      following perturbation:
{perturbation_description}

## File Changes

{diff}

## Review Text
{review_text}

## Task
Evaluate whether the review identifies the perturbation.
     Consider:
- Does the review explicitly mention the specific issue
     introduced?
- Does the review identify a direct, obvious consequence of
     the perturbation?
- Vague or generic criticisms that could apply to any
     proposal do NOT count

Award exactly one label:
- C (Correct): Review explicitly identifies and discusses
     the introduced error or direct consequences
- P (Partial): Review makes vague or incomplete reference to
      issues from the perturbation
- I (Incorrect): Review fails to acknowledge the error or
     only mentions it tangentially

Respond with a JSON with two string fields: explanation and
     verdict.
```

A.4 Panel of Judges

A.4.1 Judge Details. Qwen3.5-35B-A3B [25], NVIDIA-Nemotron-3-Nano-30B-A3B [22], and GLM-4.7-Flash [31]. Each judge independently evaluates whether the review identifies the perturbation, and we take the majority verdict. Each judge is given a human-authored description of the perturbation, along with the file diffs between the original and perturbed proposals.

A.4.2 Judge Validation. To validate the judge panel, we assess inter-annotator agreement between the three models. The panel achieves a Krippendorff's α of 0.74, indicating substantial agreement. We also perform a comparison against a human-annotated subset of 50 samples in which we find Qwen3.5 and GLM-4.7 to exhibit perfect agreement with the human annotator ($kappa = 1.0$), while Nemotron shows a fair alignment of $\kappa = 0.57$. The majority voting mechanism effectively compensates for Nemotron's higher rate of disagreement, resulting in reliable verdicts.

A.5 Human Evaluation Breakdown

Category	Human	LLM	Total
ALIGNMENT	13	11	24
COMPETENCY	30	26	56
ETHICS	12	21	33
FUNDING	38	40	78
IMPACT	52	36	88
CLARITY	144	54	198
TIMELINE	13	10	23
TOTAL	**302**	**198**	**500**

Table 7: Claims used for human evaluation.

A.5.1 Expert–Model Feedback Alignment. Review comments provide applicants with critical feedback for resubmission and offer transparency into the decision-making process. We reframe these reviews as sets of atomic claims, each with a valence (positive, neutral, or negative), comparing model-generated claims (M) to expert reviews (E). Their intersection ($M \cap E$) represents consensus, while claims unique to experts ($E \setminus M$) identify areas where the model lacks technical depth or alignment with reviewer priorities. Finally, valid claims exclusive to the model ($M \setminus E$) represent the potential additive value of LLMs, highlighting insights that human reviewers may have overlooked.

We perform a human annotation of LLM claims that do not appear in expert claims. The two proposals we received with review comments each have four original expert reviews. Following a manual assessment, we observe no direct contradictions in the claims made between human reviewers for each proposal. Contradictions are defined as factual discrepancies or opposing valence between claims, such that the claims cannot both be true at the same time [14].

We generate reviews for the two proposals using each of our LLM review systems and extract individual claims from them. To extract atomic claims, we utilise GPT-OSS-120B (high) to break down review sentences into their atomic form. We define a review claim as a claim referring to a specific topic within the grant that expresses a positive, neutral or negative valence, thereby capturing constructive feedback, identified strengths and weaknesses [14]. In many cases, a single claim sentence can contain multiple evaluative aspects, for example assessing feasibility and novelty simultaneously. Decomposing such sentences into atomic units isolates each evaluative component as a separate claim. The full claim decomposition method is shown in Appendix A.6.

We therefore created a taxonomy to split claims into specific categories that overlap with proposal sections (see subsection A.8). Once a claim is labelled with a taxonomy, semantically similar claims are grouped using embedding-based clustering with Qwen3-Embedding-8B [37]. This allows related claims to be grouped before assigning an aspect. For each resulting cluster, the model generates a concise three-word aspect that captures the cluster's topic, reducing variability in aspect labels across similar claims.

Decomposed claims sharing the same source sentence, valence, and aspect are re-merged into their original composite form; otherwise they remain decomposed. After processing, we perform bi-directional relevance matching using E2Rank [18], an embedding-to-rank model combining retrieval and listwise re-ranking. For each claim in the LLM set, we retrieve the most similar claims from the human set above a similarity threshold of 0.5 as potential matches, and vice versa.

Each claim, together with its associated potential matches, is then labelled as EXACT, DIFFERENT or CONTRADICTION, following a similar framework to Fritsch and Jatowt [8]. EXACT indicates that the LLM and human claims address the same topic and express the same valence, while CONTRADICTION indicates alignment in topic but opposing valence. DIFFERENT captures partial overlap in topic without clear agreement in evaluative valence. A further human quality control step was conducted to remove incomplete, overly generic claims introduced during decomposition. The final dataset consists of 500 claims, and the full breakdown is shown in Table 7.

For human evaluation claims are given per section of the proposal due to ethical limitations of sharing personal data and intellectual property outside of the research team. Each annotator reads a section, opportunity and the review guidelines. Each task is a separate claim that they must state 1) the validity of the claim, 2) their agreement with (strong disagree, disagree, neutral, agree, strong agree) and 3) rate the significance of the comment with respect to how much impact that claim would have on the review score, assuming it was correct.

Within human reviewing, an absence of comment can often be viewed as a satisfaction of the requirements. Trivial additional positive comments do not indicate significant value. Similarly human reviewers might list only the most significant issues necessary to kill a proposal. Minor negative comments missing from human reviews indicate potential for value in delivering constructive criticism and actionable feedback to applicants. Major negatives absent from human reviews indicate actual opportunity for value that could be provided by an LLM.

A.6 Claim Decomposition

This section outlines our approach to processing the text into units, detailing how atomic claims are categorised, clustered, and cross-referenced to facilitate a robust comparative analysis.

We created a taxonomy to split claims into specific categories that overlap with proposal sections (see subsection A.8). Once a claim is

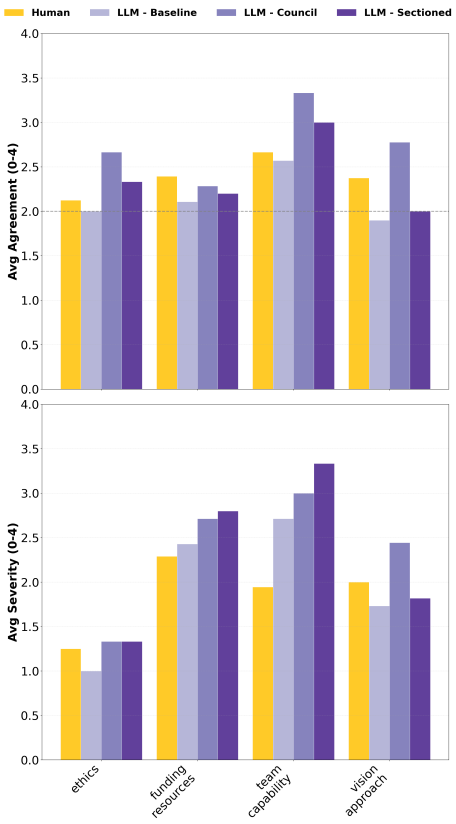

Figure 3: Table 11 shows the scale for severity. The dashed line on agreement indicates the neutral stance (2 on the scale).

labelled with a taxonomy, semantically similar claims are grouped using embedding-based clustering with Qwen3-Embedding-8B [37]. This allows related claims to be grouped before assigning an aspect. For each resulting cluster, the model generates a concise three-word aspect that captures the cluster's topic, reducing variability in aspect labels across similar claims.

Decomposed claims sharing the same source sentence, valence, and aspect are re-merged into their original composite form; otherwise they remain decomposed. After processing, we perform bi-directional relevance matching using E2Rank [18], an embedding-to-rank model combining retrieval and listwise re-ranking. For each claim in the LLM set, we retrieve the most similar claims from the human set above a similarity threshold of 0.5 as potential matches, and vice versa.

Each claim, together with its associated potential matches, is then labelled as EXACT, DIFFERENT or CONTRADICTION, following a similar framework to Fritsch and Jatowt [8]. EXACT indicates that the LLM and human claims address the same topic and express the same valence, while CONTRADICTION indicates alignment in topic but opposing valence. DIFFERENT captures partial overlap in topic without clear agreement in evaluative valence. A further human quality control step was conducted to remove incomplete, overly generic claims introduced during decomposition. The final dataset consists of 500 claims, and the full breakdown is shown in Table 7.

A.7 Perturbation Examples

We provide examples for each perturbation type in Table 8.

A.8 Section Taxonomy

Table 10 provides a breakdown of each of the seven axes and what components, sub-components and aspects were associated with them.

A.9 Claim Severity Definitions

Table 11 provides definitions for each different level of claim severity.

A.10 Review-System Runtimes and Token Usage

Table 12 shows the token usage and wall-clock time for each configuration and reasoning effort.

A.11 Human Evaluation Results

The combined average results from the human evaluation can be seen in Figure 3.

Category / Variant	Example Content
Bracket and Example Removal	
Original	The framework supports multiple modalities (such as text, image, and audio) to ensure versatility in downstream tasks. The system provides several security features (including end-to-end encryption and two-factor authentication) for user protection.
Brackets Removed	The framework supports multiple modalities to ensure versatility in downstream tasks. The system provides several security features for user protection.
Bracket + Example Removed	The framework supports multiple modalities. The system provides several security features.
Numerical De-quantification	
Original	The study surveyed 1,250 participants across 15 different countries to ensure diversity. The model achieved a 98.5% accuracy rate after only 5 epochs of training.
Numerical Removed	The study surveyed many participants across several countries to ensure diversity. The model achieved a high accuracy rate after a few epochs of training.
Framing and Methodological Reduction	
Original	We develop a novel framework to implement real-time anomaly detection using a multi-layered transformer architecture and gradient-based optimisation. The team introduced a new approach for cross-border transactions by utilizing a decentralised ledger with zero-knowledge proofs.
Existing-work Framing	The framework provides real-time anomaly detection using a multi-layered transformer architecture and gradient-based optimisation. The approach facilitates cross-border transactions utilizing a decentralised ledger with zero-knowledge proofs.
Methodological Reduction	We develop a novel framework to implement real-time anomaly detection. The team introduced a new approach for cross-border transactions.
Connective Removal	
Original	The study identified a discrepancy in the results. To bridge the gap, the researchers introduced a secondary validation set. The initial deployment encountered scaling issues. In light of these findings, the architecture was redesigned for distributed systems.
Connectives Removed	The study identified a discrepancy in the results. The researchers introduced a secondary validation set. The initial deployment encountered scaling issues. The architecture was redesigned for distributed systems.
Competency Perturbation	
Original	Name1 has an extensive track record in NLP, with publications at ACL, EMNLP, and NAACL. Their portfolio demonstrates **expertise in efficient transformer architectures and scaling large language models via distributed training**. They have served as a Senior Area Chair and managed multi-institutional grants.
Removal	Name1 has an extensive track record in NLP, with publications at ACL, EMNLP, and NAACL. They have served as a Senior Area Chair and managed several multi-institutional research grants focused on neural machine translation.
Funding	
Original	**Compute resources:** £2,400 for cloud compute over six months. **Travel expenses:** £1,800 for conference attendance.
Addition	Original categories plus **Office supplies:** £850 for ergonomic seating.
Deletion	**Compute resources:** £2,400 for cloud compute over six months.
Funding	
Excessive	Compute: £120,000; Travel: £55,000.
No Values	Budgets requested without numerical specification.
Vague	Funding requested for computing resources and conference attendance.
Impact	
Short-Term	This study will provide a minor refinement in translation efficiency to adjust how different groups communicate.
Original	This study will provide a significant boost in translation efficiency to improve how different groups communicate.
Long-Term	Fundamental shift redefining translation efficiency.This study will provide a fundamental shift in translation efficiency to redefine how different groups communicate.
Timeliness and Stakeholders	
Original	This project introduces more efficient training methods for large models. It directly reduces the environmental impact and carbon footprint of NLP research.
Timeliness Removed	This project introduces more efficient training methods for large models.
Stakeholder Shift	This project introduces more efficient training methods for large models. It directly reduces the environmental impact and carbon footprint of University teaching.

Table 8: Examples of perturbations.

Variant	Summary Totals			Directly Allocated			Staffing	Directly Incurred			
	Full Funding	Org Cont.	Applied	Staff	Estates	Other	%FTE	Staff	Equip.	Travel	Other
Original	£25,000	£5,000	£20,000	£8,000	£2,000	£1,000	40%	£5,000	£2,000	£1,000	£1,000
High Org Cont.	£25,000	**£21,000**	**£4,000**	£1,500	£500	£250	40%	£1,000	£250	£250	£250
Low Eq / High Other	£25,000	£5,000	£20,000	£8,000	£2,000	£1,000	40%	£5,000	**£100**	£1,000	**£2,900**
Low Staff Cost	£25,000	£5,000	£20,000	**£480**	£4,520	£2,000	40%	£8,000	£2,000	£1,500	£1,500
Low Staff FTE	£25,000	£5,000	£20,000	£8,000	£2,000	£1,000	**1%**	£5,000	£2,000	£1,000	£1,000
No Org Cont.	£25,000	**£0**	**£25,000**	£10,000	£3,000	£2,000	40%	£6,000	£2,000	£1,000	£1,000

Table 9: Funding perturbations with constant Full Funding across all variants.

Axis	Component	Sub-component	Aspects
1. Competency	Team Capability	Experience & Track Record	expertise_domain, track_record_outputs, track_record_leadership, career_stage_appropriateness
		Skills & Expertise	skill_coverage, skill_gaps, complementarity
		Leadership & Management	communication_ability, team_development, project_management, cross_sector_influence
2. Funding	Resources & Justification	Resource Specification	staff_justification, travel_justification, compute_resources, resource_completeness
		Appropriateness	staff_time_realistic, resource_alternatives
		Value for Money	outcome_proportionality, impact_optimization
		Infrastructure	facilities_access, institutional_support, collaborative_networks
3. Timeline	Timeline Realism	General Feasibility	duration_appropriateness, milestone_achievability, workpackage_scheduling
4. Alignment	Strategic Alignment	Remit Fit	remit_primary, theme_alignment, critical_tech_relevance
		Strategic Contribution	priority_area_fit, urgency, portfolio_contribution, gap_filling
5. Clarity	Vision Quality	Scientific Excellence	novelty, significance, conceptual_clarity, hypothesis_quality
	Approach Quality	Methodological Rigor	methodology_robustness, methodology_appropriateness, methodology_transparency, validation_strategy
		Risk Management	risk_identification, risk_mitigation, governance_appropriate
		Previous Work	literature_awareness, building_on_previous, preliminary_data
	Writing Quality	Clarity	writing_clarity, structure_logic, technical_precision
		Completeness	information_sufficiency, assumption_explicit, references_quality
6. Impact	Impact Potential	Academic Impact	field_advancement, interdisciplinary_catalyst, capacity_building
		Practical Impact	societal_benefit, economic_value, policy_influence
		Impact Pathway	pathway_credibility, timeline_to_impact, partner_commitment, impact_measurement, dissemination_plans, stakeholder_engagement
		Beneficiaries	beneficiary_identification
7. Ethics	Ethics & RRI	Ethical Considerations	ethics_identification, ethics_management, ethics_acceptability
		Research Integrity	data_management, reproducibility, transparency, conflicts_declared

Table 10: Taxonomy of grant proposals

Rating	Definition	Examples
None	Purely factual or administrative. No bearing on scientific merit or deliverability.	*"The first letter of the sentence was not capitalised."*
Little	Minor issues that are easily correctable or do not affect core assessment criteria.	*"Figure 3 is difficult to read."*
Some	Valid observations affecting secondary criteria. Would influence score by ±0.5 points.	*"The budget justification for travel could be more detailed."*
Substantial	Significant strengths or weaknesses directly affecting Quality or Importance. Would shift score by ±1–2 points.	*"The proposed methodology represents a genuine advance over current techniques."*
Pivotal	Fundamental issues affecting viability or exceptional strengths. Changes fundable/non-fundable status.	*"The underlying theoretical framework contradicts established principles."*

Table 11: Impact rating definitions for different levels of claim severity.

Review	Effort	Wall Clock	Input	Output	Total
Baseline	Low	00:20:37	21,088,675	437,826	21,526,501
	Medium	00:34:04	21,088,675	1,157,919	22,246,594
	High	01:37:06	21,088,675	3,628,537	24,717,212
Section-Level	Low	02:04:33	27,231,048	8,532,425	35,763,473
	Medium	02:40:23	27,828,422	11,221,317	39,049,739
	High	04:22:37	28,094,907	18,489,158	46,584,065
Council	Low	03:56:45	152,950,787	8,073,411	161,024,198
	Medium	06:53:35	156,346,652	18,769,936	175,116,588
	High	–	–	–	–

Table 12: Wall-clock time of GPT-OSS-20B and token breakdown from producing reviews for all original proposals and perturbations at each reasoning level. Each was generated 5 times to assess variability (total: 910). Council on high reasoning was stopped at 6 hours (25%) completion due to excessive runtime cost.