Benchmarking Expertise Retrieval for Peer Review: Evaluating Statistical and Neural Representations for the Reviewer Assignment Problem

Anonymous authors
Paper under double-blind review

Abstract

The exponential growth of scientific submissions has strained the peer review system. Despite the rapidly expanding global pool of researchers, this scale has rendered the previous approach of manual expert identification impractical. Therefore, institutions have naturally turned to Large Language Models (LLMs) to automate intricate processes like expertise identification. However, the reliability of these new models in accurately identifying domain experts lacks rigorous evaluation. We conduct an empirical evaluation of statistical and AI-driven expertise representation methods to benchmark their reliability and limitations. Framing expert identification as an information retrieval problem, we utilize the Distributed Peer Review system of the European Southern Observatory (ESO) Period 110 call, where proposal authorship serves as our proxy ground truth for domain expertise. Evaluating six retrieval methodologies utilized across observatories and computer science conferences, we find that traditional statistical representations outperform generative AI. Specifically, Term Frequency-Inverse Document Frequency successfully identified a labeled expert within the top 25 recommendations 79.5% of the time, compared to 51.5% for GPT-4o mini. Our results highlight that distinguishing subfield expertise benefits from fine-grained vocabulary, which is obscured by the semantic smoothing in generative methods. Using operational data from ESO's Distributed Peer Review system, our findings suggest that transparent and reproducible statistical representations are a strong baseline for expertise identification in peer review, outperforming computationally expensive LLMs in this specialized setting. The code, anonymized aggregate scores, and synthetic benchmark functionality are released to enable full reproduction of our evaluation pipeline and to support future work on expertise retrieval benchmarking at (link provided upon acceptance).

1 Introduction

The exponential growth of scientific outputs is placing unprecedented strain on the global peer review system, challenging the sustainability of research publishing, grant funding, and resource allocation (Kovanis et al., 2016; Hanson et al., 2023). As submissions surge across all disciplines, the reliance on manual workflows to identify and assign qualified experts for peer review has become a bottleneck (Zhao et al., 2021; Saveski et al., 2023; Xu et al., 2026). In response, institutions, such as astronomical observatories and computer science conferences, are rapidly adopting automated systems that leverage Large Language Models (LLMs) or statistical methods to estimate scientific expertise and match reviewers to submissions (Charlin et al., 2012; Kerzendorf et al., 2020; Lee et al., 2025; Stelmakh et al., 2025; Teixeira, 2025). However, deploying automated systems in peer review requires systematic evaluation against the scientific standards of transparency and reproducibility, not just administrative performance.

The Reviewer Assignment Problem, the task of identifying and matching qualified experts to peer review opportunities, is typically framed as a two-stage process: first, generating similarity scores to estimate expertise, and second, computationally optimizing the final assignments subject to constraints (e.g., conflicts of interest, reviewer workload) (Stelmakh et al., 2019; Leyton-Brown et al., 2022). In this work, we focus

specifically on the first stage, evaluating how well different representation methods retrieve qualified experts. Identifying qualified experts typically involves mapping reviewer bibliographies and research output texts (e.g., publications or proposals) into shared, high-dimensional vector spaces to compute pairwise similarity (Mimno & McCallum, 2007; Charlin et al., 2012; Zhang et al., 2019). Without an objective definition of a scientific 'expert,' current methods rely on topical similarity as the primary signal of expertise, delegating qualifications (e.g., seniority) to the constraint-based optimization (Kerzendorf et al., 2020; Carpenter et al., 2025).

To handle these scaling challenges, disciplines from astronomy to computer science have adopted various machine learning methods for expertise identification and assignment (Kerzendorf et al., 2020; Stelmakh et al., 2025). For example, the Space Telescope Science Institute and the Toronto Paper Matching System both utilize Term Frequency-Inverse Document Frequency (TF-IDF) to align reviewers to proposals or papers (Strolger et al., 2017; 2023; Charlin et al., 2012). Similarly, the Atacama Large Millimeter Array and several computer science pipelines infer expertise using topic modeling via Latent Dirichlet Allocation (LDA) (Meyer et al., 2022; Carpenter et al., 2025; Rosen-Zvi et al., 2012; Mimno & McCallum, 2007; Anjum et al., 2019). More recently, conference platforms like OpenReview have deployed Transformer-based semantic embeddings (OpenReview, 2025), while funding bodies like the European Research Executive Agency are exploring specialized LLMs to match experts to fellowship proposals (Álvarez-García et al., 2026). These semantic approaches contrast with the explicit keyword-based matching currently utilized by the European Southern Observatory (Jerabkova et al., 2023). Ultimately, the fragmented landscape of algorithms points to a need for systematic evaluation of how these representations retrieve experts in operational environments (Stelmakh et al., 2025).

Despite the widespread deployment of these methodologies, systematically benchmarking their efficacy is hindered by the absence of ground-truth labels for scientific expertise, forcing prior evaluations to rely on inherently biased proxy labels (Anjum et al., 2019; OpenReview, 2025; Stelmakh et al., 2025). For instance, external human annotation introduces noise as annotators may lack specific reviewer knowledge (Mimno & McCallum, 2007; Zhao et al., 2021). Conversely, self-assessments are susceptible to calibration biases (e.g., self-efficacy bias) (Dumais & Nielsen, 1992; Kruger & Dunning, 1999; Ehrlinger & Dunning, 2003; Rodriguez & Bollen, 2008; Aitymbetov & Zorbas, 2025), and proxying expertise strictly through authorship (See OpenReview Github repository: OpenReview, 2025) may not capture a reviewer's broader domain expertise (Stelmakh et al., 2025). Ultimately, the field lacks systematic evaluations of the underlying representation methodologies.

Major astronomical observatories provide a well-suited environment to systematically benchmark expertise representation methods. The allocation of telescope time is a competitive process of distributing limited resources, demanding precision in expert evaluation (Strolger et al., 2017; Kerzendorf et al., 2020). Crucially, the recent adoption of the Distributed Peer Review (DPR) process at various facilities, where proposal authors simultaneously act as the reviewer pool, creates a closed-loop system that allows for the rigorous evaluation of expertise representation methods (Merrifield & Saari, 2009; Patat et al., 2019). Leveraging the DPR system deployed by the European Southern Observatory (ESO) for its Period 110 (P110) call, we conduct a benchmark study across 379 reviewers and 435 proposals. Treating expert identification as an Information Retrieval (IR) problem, we evaluate six distinct methodologies utilized across astronomy and computer science: keywords, TF-IDF, LDA, two Transformer embedding models, and GPT-4o mini. We introduce a dual validation strategy to assess both retrieval performance and alignment with human judgment: treating the submitting researcher as a 'proxy expert' for their own proposal to evaluate how effectively each method retrieves known experts, and comparing the resulting similarity scores against the expertise categories researchers selected for themselves to assess alignment with self-reported expertise.

2 Data

The data for this study originate from the P110 observing call at ESO. P110 refers to the scheduling cycle for ESO telescope time that began in late 2022 (Jerabkova et al., 2023). During this cycle, ESO first deployed the DPR system for all proposals requesting fewer than 16 hours of observing time. In total, P110 received 435 eligible proposals involving 2014 unique investigators. Under the DPR framework, submitting teams

are required to nominate one investigator to serve as a reviewer. Throughout this paper, we refer to this individual as the "proposal-designated reviewer" to distinguish them from other co-investigators. Because some individuals were nominated to represent multiple proposals, the 435 submissions resulted in a final pool of 379 unique reviewers. Each proposal-designated reviewer was assigned to assess 10 proposals.

The primary data for this study consist of (1) the proposal text and (2) reviewer profiles (e.g., self-reported keywords or publication abstracts). Each proposal includes multiple sections, from scientific justification to feasibility, but our experiments focus on the proposal abstract. This serves as a concise summary of the scientific content, with the longest proposal abstract in P110 containing 176 words. We rely on two data sources for representing reviewer expertise: ESO keyword data and publication abstracts retrieved from the NASA/ADS digital library, which offers an open API. Using the ADS Python package[1], we query abstracts for each proposal-designated reviewer by name. Alongside the keywords and abstracts, we utilize the self-reported expertise labels collected during the P110 DPR at ESO. For each of the 4350 reviewer–proposal assignments (corresponding to the 435 proposals, each assigned to ten reviewers), the assigned reviewer categorized their own expertise into one of three distinct labels: Expert, Intermediate, or Non-expert. Detailed review statistics, NASA/ADS query descriptions, and the baseline ESO keyword distributions are provided in Appendix A.

3 Methods

To evaluate how effectively each approach estimates reviewer expertise, we first formalize the task of identifying an expert reviewer as an information retrieval problem before describing our experimental setup and detailing the individual representation methods in Appendix B.

3.1 Problem Formulation

We frame the task of identifying expert reviewers as an information retrieval problem. In this formulation, each proposal acts as a query seeking to return the most relevant results, which corresponds to a set of reviewers. The goal of each representation method is to rank reviewers from most suitable to least for a given proposal.

In our study p_i denotes the embedding representation of proposal i, and r_j denotes the embedding representation of reviewer j as produced by a given method. A similarity function $d(p_i, r_j) = s_{ij}$, such as cosine similarity, assigns a similarity score between the proposal and reviewer vector representations.

For each method, we first compute vector representations for each unique reviewer's publication history ($N = 379$) and for each proposal abstract ($N = 435$). We normalize all vectors using the L_2 norm to project them onto the unit hypersphere. By normalizing the vectors, we measure only the direction of the expertise, ignoring the length (or word count) of a reviewer's collection of abstracts. We then calculate pairwise cosine similarities between all reviewer and proposal vectors, producing an expertise matrix that represents the similarity between each reviewer's expertise and every proposal. The resulting matrix has dimensions 435×379 , where each entry s_{ij} denotes the similarity between proposal p_i and reviewer r_j. The IR formulation provides the foundation for our evaluation, where we assess how effectively each representation ranks reviewers according to their represented expertise.

3.2 Experimental Setup

To systematically compare the various expertise representation methods, we established a standardized evaluation framework. Although each method maps textual inputs into a continuous vector space, these spaces differ significantly in distribution, rendering raw similarity scores directly incomparable. Therefore, our evaluation relies on rank-based IR metrics to assess the ordering of reviewers based on similarity scores. In the absence of a standardized ground-truth benchmark for expertise modeling (Stelmakh et al., 2025), we adopt a dual validation strategy. We evaluate each methodology against two distinct sources of labels

[1] https://ads.readthedocs.io/en/latest/

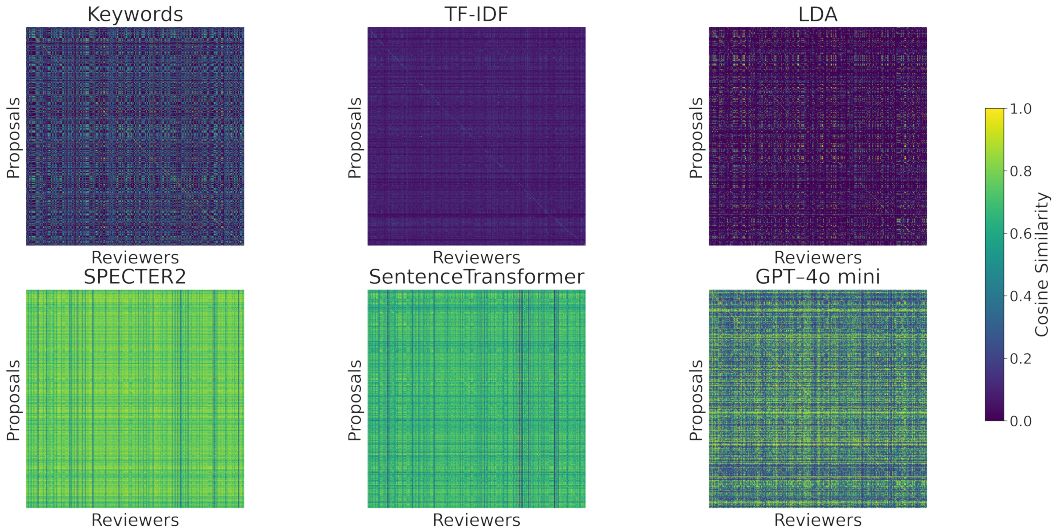

Figure 1: Similarity score matrices across methods for Period P110. Rows and columns are sorted such that the proposal-designated reviewer-proposal pairs align along the diagonal.

derived from the ESO DPR system: (1) proxy gold labels based on proposal authorship and (2) self-reported reviewer expertise.

First, we utilize the unique structure of DPR to assign proxy gold labels. As each proposal designates an investigator to participate as a reviewer in the DPR process, we define each proposal-designated reviewer as a proxy expert on their own proposal. This formulation enables each proposal–reviewer ranking to be evaluated as an IR task. To quantify the quality of the vector representations output by each expertise modeling technique, we compute four evaluation metrics for each proposal-designated reviewer. First, we report the Median Rank of the proposal-designated reviewer within the sorted list of all candidate reviewers. Second, we compute the Mean Reciprocal Rank (MRR), defined as the average of the inverse of the rank ($\frac{1}{\text{rank}}$) across all proposals, which penalizes lower rankings more heavily. Third, we measure Hit@25, the fraction of proposals where the proposal-designated reviewer appears within the top 25 candidates. Finally, to assess how significantly the designated reviewer stands out from the entire sample of reviewers, we calculate a standardized z-score based on the distribution of similarity scores. For a given proposal, this is defined as:

$$z = \frac{S_{\mathrm{I}} - \mu}{\sigma}$$

where S_{I} is the similarity score assigned to the proposal-designated reviewer, and μ and σ are the mean and standard deviation, respectively, of the similarity scores for all potential reviewers for that specific proposal.

Finally, we benchmark the expertise representations against self-reported expertise labels from P110. We utilize the Normalized Discounted Cumulative Gain (NDCG) metric for evaluation to account for the hierarchy of self-reported labels, penalizing models that fail to rank an Expert label above a Non-expert label (see Appendix B.6 for the formal formulation). We utilize the scikit-learn implementation of NDCG (Pedregosa et al., 2018). In our evaluation, a graded relevance score is assigned based on the reviewer's self-reported expertise for the subset of reviewer-proposal pair assignments in P110 totaling 4350 (10 for each proposal). The relevance scores are weighted to reward the correct ranking of 'Expert' reviewers (assigned a score of 10), while only providing a minor gain for 'Intermediate' reviewers (assigned a score of 2), and no gain for 'Non-Expert' candidates. This evaluation metric emphasizes the importance of ranking labeled Experts above Non-experts.

Table 1: Average rank-based retrieval performance across 435 proposals. Values are reported as mean ± margin of error (95% CI). Statistically significant differences from the baseline (Keywords) are marked with * ($p < 0.05$) and †($p < 0.01$). **Bold** indicates the best performance.

Method	Median Rank ↓	MRR ↑	Hit@25 ↑	Z-score↑
Keywords	9.0 ±2.5	0.270 ±0.032	0.703 ±0.044	2.051 ±0.095
LDA (K = 50)	24.0 ±6.0 †	0.148 ±0.024 †	0.503 ±0.047 †	2.114 ±0.208
TF–IDF	**4.0** ±1.0 †	**0.408** ±0.037 †	**0.795** ±0.038 †	**3.969** ±0.340 †
SPECTER2 (Mean pooling)	10.0 ±2.5 †	0.288 ±0.034	0.634 ±0.046 *	0.834 ±0.114 †
SentenceTransformer(Mean pooling)	7.0 ±2.0	0.341 ±0.036 †	0.710 ±0.043	1.252 ±0.123 †
GPT–4o mini	24.0 ±4.0 †	0.166 ±0.027 †	0.515 ±0.047 †	1.684 ±0.079 †

To ensure that our performance metrics are robust and not artifacts of specific sample selection, we calculate 95% confidence intervals for all aggregate results. We utilize bootstrap resampling ($n = 10,000$, percentile method) (Efron, 1979), which allows us to estimate the uncertainty of our metrics (e.g., the error bars on MRR) without assuming a specific underlying distribution. Furthermore, to rigorously compare the proposed methods against the baseline (i.e., keywords), we employ the Wilcoxon signed-rank test (Wilcoxon, 1945). Unlike standard tests that assume Gaussian errors, this non-parametric test is better suited for Information Retrieval metrics like Rank and MRR, which are typically heavily skewed. We report differences as statistically significant when the probability of the result occurring by chance is less than 5% ($p < 0.05$).

4 Results

Comparing raw scores across methods is difficult because different methods produce very different score distributions and sparsity patterns (see Table 3 in Appendix C for a full summary). Keyword-based and topic modeling methods exhibit high sparsity, though for different reasons: keywords produce roughly 52% zero-valued similarities. For topic modeling, while absolute zero similarities are mathematically rare, approximately 68% of the scores fall below 0.01. Because standard implementations (e.g., Gensim) default to truncating probabilities below this 0.01 threshold, these practically operate as zero-valued similarities in the matrix. In contrast, TF–IDF produces less than one percent zero scores, but its maximum score is only 0.50. Conversely, the SPECTER2 (Mean pooling) embeddings yield uniformly high similarities with a minimum score of 0.75. Therefore, simply comparing raw score magnitudes alone does not indicate the superiority of one method over another.

Figure 1 presents the reviewer-proposal similarity matrices where rows correspond to reviewers and columns to proposals. We order matrices so that the proposal-designated reviewer pairs lie approximately on the diagonal. Because a proposal's designated reviewer is expected to be knowledgeable about their own proposal, a visible diagonal provides a baseline check that each method captures reviewer–proposal similarity alignment.

Therefore, we utilize standard retrieval-based metrics to quantitatively compare the performance of each method. We evaluate the similarity scores for each method in two stages: first, by measuring the ability of each method to retrieve the proposal-designated reviewer as a known proxy label of expertise, and second, by assessing how the computed similarity scores align with the reviewers' self-reported expertise for assigned matches. Table 1 details the rank-based retrieval performance across the 435 proposals. We report values with 95% confidence intervals to account for sample variance and determine statistical significance relative to the Keyword baseline. As shown in Figure 2, TF–IDF demonstrates the most robust retrieval performance across all evaluated methods, consistently placing the designated reviewer higher in the ranked list. TF-IDF identified the proposal-designated reviewer with a Median Rank of 4.0 (±1.0) and a Hit@25 of 0.795 (±0.038). The SentenceTransformer (Mean pooling) method follows with a Median Rank of 7.0 (±2.0) and a Hit@25 of 0.710 (±0.043). The Keyword baseline yields a Median Rank of 9.0 (±2.5) and a Hit@25 of 0.703 (±0.044). Notably, while the Keyword method retrieves the designated reviewer at a similar rate to the SentenceTransformer, it achieves a lower MRR of 0.270 compared to 0.341, suggesting the Transformer consistently ranks the designated reviewer higher within that top 25. Conversely, SPECTER2 (Mean pooling)

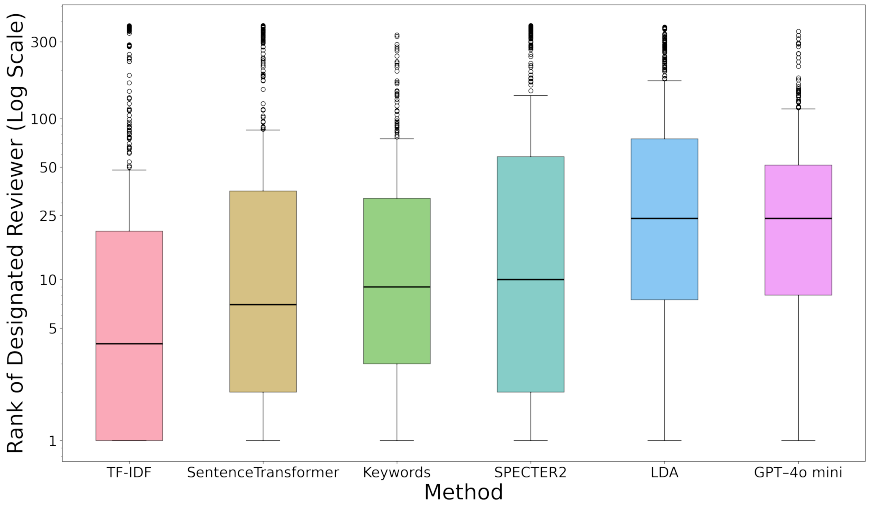

Figure 2: Distribution of the rank assigned to the proposal-designated reviewer across 435 proposals. Boxes span the 25th to 75th percentiles, with whiskers showing the range of non-outlier data. The y-axis is logarithmic to accommodate the wide variance in rankings.

yields a higher Median Rank of 10.0 ($\pm$2.5) and underperforms the baseline in recall, achieving a Hit@25 of only 0.634 ($\pm$0.046). Finally, both LDA and GPT-4o MINI underperform the baseline ($p < 0.01$). GPT-4o MINI yields a Median Rank of 24.0 ($\pm$4.0) and a Hit@25 of 0.515 ($\pm$0.047), performing comparably to LDA, which achieved a Median Rank of 24.0 ($\pm$6.0) and a Hit@25 of 0.503 ($\pm$0.047).

Complementing the proposal-designated reviewer retrieval evaluation, we examine how the computed similarity scores align with reviewers' self-reported expertise labels for their assigned proposals. An effective expertise representation should yield consistently higher similarity scores for self-identified 'Experts' relative to 'Non-Experts.' Table 2 presents the NDCG for each method. Notably, the automated text representations outperform the self-reported Keywords baseline in aligning with human judgment. TF-IDF achieves the highest alignment with an NDCG of 0.832 ($\pm$0.014), closely followed by the SentenceTransformer (Mean pooling) at 0.827 ($\pm$0.014). Both methods demonstrate a statistically significant improvement over the baseline ($p < 0.05$). SPECTER2 (Mean pooling) and GPT-4o mini yield NDCG scores of 0.815 ($\pm$0.014) and 0.811 ($\pm$0.011), respectively, both outperforming the baseline. Conversely, the explicitly selected Keywords baseline exhibits the weakest alignment (0.789 $\pm$0.014), performing comparably only to LDA (0.798 $\pm$0.015), which is the only automated method that does not show a statistically significant improvement over the Keywords baseline.

5 Discussion

Our results show that traditional statistical lexical methods (i.e., TF-IDF) outperformed both semantic neural architectures and generative models in retrieving proposal-designated reviewers. By leveraging exact matches of domain-specific vocabulary (e.g., *Type Ia*, *Brown Dwarf*), TF-IDF encodes precise expertise signals. In contrast, the latent semantic clustering of neural embeddings (SentenceTransformer and SPECTER2) applies a smoothing effect that obscures the fine-grained distinctions required to differentiate highly specialized sub-domain expertise. Although most methods effectively identified relevant experts, GPT-4o MINI and LDA underperformed the Keywords baseline in median rank and recall ($p < 0.01$). Initial exploratory tests with the latest generation of reasoning-based GPT models yielded retrieval performance similar to GPT-4o MINI. The architecture of newer models currently disables the parameter controls (e.g., temperature) (OpenAI et al., 2026) necessary to minimize generation variance and maximize reproducibility for formal benchmarking. To ensure scientific reproducibility in our formal evaluation, we utilized automated

Table 2: Alignment with human self-reported expertise. We report Normalized Discounted Cumulative Gain (NDCG) to measure ranking quality. Values are reported as mean ± margin of error (95% CI). Statistically significant differences from the baseline (Keywords) are marked with * ($p < 0.05$) and †($p < 0.01$). **Bold** indicates the best performance.

Method	NDCG ↑
TF-IDF	**0.832** ±0.014 *
SentenceTransformer (Mean pooling)	0.827 ±0.014 *
SPECTER2 (Mean pooling)	0.815 ±0.014 *
GPT–4o mini	0.811 ±0.011 *
LDA	0.798 ±0.015
Keywords	0.789 ±0.014

context retrieval, structured system prompting, and deterministic guardrails (see Appendix, Section B.4 for details on our harness).

Beyond baseline performance, TF-IDF and SentenceTransformer also proved robust in data-constrained scenarios, maintaining stable performance where LDA and SPECTER2 degraded with limited publication histories (see Appendix, Tables 4–6 for ablations). Furthermore, algorithmic hyperparameters influenced retrieval accuracy; max pooling severely degraded Transformer performance (see Appendix, Table 8), and LDA exhibited sensitivity to the configured topic count (see Appendix, Table 7). Additionally, deploying generative LLMs introduces computational costs and relies on probabilistic generation, which poses challenges for the deterministic reproducibility required for peer review decisions and allocations (see Appendix D). Finally, automated text representations, led by TF-IDF (NDCG = 0.832), better aligned with human self-reported expertise than the self-selected Keywords baseline, highlighting that abstracts capture a reviewer's domain expertise more accurately than coarse manual categorization.

Our findings contrast with recent literature (Stelmakh et al., 2025), which concluded that TF-IDF requires the full text of manuscripts to compete with specialized deep-learning models like SPECTER2. In our evaluation, TF-IDF outperforms semantic neural architectures even when all methods are strictly constrained to publication abstracts. We attribute this divergence to fundamental differences in dataset composition and the framing of the evaluation task. Stelmakh et al. (2025) constructed a dataset of 58 researchers evaluating 5 to 10 papers they had previously read, testing algorithms on their ability to correctly order the small, localized set of papers for a specific researcher. Our evaluation explicitly mirrors the operational reality of proposal/conference peer review administration: globally searching a candidate pool to identify the most relevant experts.

5.1 Limitations

Several limitations should inform how these results are interpreted, spanning the proxy labels used for evaluation, the scale and domain specificity of the dataset, and the construction of reviewer profiles. We acknowledge that this study relies on proxy labels, such as proposal authorship and self-reported labels, as no objective ground truth for scientific expertise exists (Stelmakh et al., 2025). While proposal authorship is a strong proxy for domain knowledge, it may introduce a bias toward a researcher's vocabulary usage rather than capturing broader domain expertise (Stelmakh et al., 2025). Similarly, self-reported expertise labels introduce self-efficacy bias, as they rely on the reviewer's personal perception of their knowledge relative to a proposal rather than a standardized taxonomy (Ehrlinger & Dunning, 2003). Our evaluation is limited in scale, drawing from a single observing cycle with 379 reviewers and 435 proposals, necessitating future work across larger reviewer pools, multi-cycle data, and different domains to explore generalization of results. Additionally, our results are specific to astronomical peer review, where highly specialized and technical vocabulary may amplify the advantage of lexical methods like TF-IDF relative to domains with broader, less field-specific terminology. Reviewer expertise profiles are constructed automatically via NASA/ADS queries, which may lead to incomplete publication histories, particularly for early-career researchers or through author name ambiguity when researchers share similar names (Olivo et al., 2025). Finally, our evaluation is limited to publication abstracts due to the logistical and copyright barriers associated with obtaining full-

text access at scale. A comprehensive discussion of further methodological constraints, including reviewer data completeness, is provided in Appendix D.

6 Conclusion

The sustainability of global peer review relies on the ability to efficiently and accurately match submissions to true domain experts at scale. By formalizing expert identification as a retrieval problem, our evaluation on ESO's Distributed Peer Review data highlights that, in this operational setting, traditional statistical representations outperform modern neural and generative architectures. These results indicate that semantic smoothing in LLMs may obscure the fine-grained vocabulary distinctions that differentiate expertise in specialized domains. As funding agencies and observatories increasingly automate their peer review systems, our evaluation suggests that transparent, computationally efficient lexical methods offer a strong baseline for expertise retrieval.

Data and Code Availability

The specific reviewer identities, self-reported expertise labels, and proposal abstracts from the European Southern Observatory Period 110 contain sensitive investigator information and cannot be made publicly available. To support reproducibility, we release anonymized, aggregate similarity score matrices for all six methods, together with ground-truth and self-reported expertise labels reindexed by anonymous proposal and reviewer identifiers, allowing the exact reproduction of our reported metrics and figures without exposing any ESO-identifying information. All code for the expertise representation methodologies and the information retrieval evaluation framework is available at (link provided upon acceptance). Because the identifying ESO data is restricted, the repository also includes functionality to generate a synthetic distributed peer review dataset, programmatically retrieving publicly available proposal authors and abstracts from various telescope missions (e.g., the Hubble Space Telescope, James Webb Space Telescope, and Chandra X-ray Observatory) via the NASA/ADS API. This synthetic benchmark enables researchers to run the full pipeline end-to-end on realistic astronomical literature, complementing the anonymized aggregate release by allowing verification of the codebase itself rather than only its reported outputs.

References

Nurmukhammed Aitymbetov and Dimitrios Zorbas. Autonomous machine learning-based peer reviewer selection system. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Brodie Mather, and Mark Dras (eds.), *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pp. 199–207, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.coling-demos.20/`.

Elena Álvarez-García, Daniel García-Costa, Ilse De Waele, Ana Marušić, and Francisco Grimaldo. Expert assignment system based on natural language processing for marie sklodowska-curie actions. *Scientific Reports*, 16, 2026. URL `https://api.semanticscholar.org/CorpusID:285068794`.

Dimitar Angelov. Top2vec: Distributed representations of topics. *ArXiv*, abs/2008.09470, 2020. URL `https://api.semanticscholar.org/CorpusID:221246303`.

Omer Anjum, Hongyu Gong, S. Bhat, Wen mei W. Hwu, and Jinjun Xiong. Pare: A paper-reviewer matching approach using a common topic space. *ArXiv*, abs/1909.11258, 2019. URL `https://api.semanticscholar.org/CorpusID:202750013`.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019. URL `https://arxiv.org/abs/1903.10676`.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 (null):993–1022, March 2003. ISSN 1532-4435.

John M. Carpenter, Andrea Corvillón, and Nihar B. Shah. Enhancing Peer Review in Astronomy: A Machine Learning and Optimization Approach to Reviewer Assignments for ALMA. *Publications of the Astronomical Society of the Pacific*, 137(3):034501, March 2025. ISSN 1538-3873. doi: 10.1088/1538-3873/adb5c1. URL https://doi.org/10.1088/1538-3873/adb5c1. Publisher: The Astronomical Society of the Pacific.

Laurent Charlin, Richard S. Zemel, and Craig Boutilier. A Framework for Optimizing Paper Matching, February 2012. URL http://arxiv.org/abs/1202.3706. arXiv:1202.3706 [cs].

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.207. URL https://aclanthology.org/2020.acl-main.207/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Susan T. Dumais and Jakob Nielsen. Automating the assignment of submitted manuscripts to reviewers. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992. URL https://api.semanticscholar.org/CorpusID:15038631.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979. URL https://api.semanticscholar.org/CorpusID:227312712.

Joyce Ehrlinger and David Dunning. How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of personality and social psychology*, 84 1:5–17, 2003. URL https://api.semanticscholar.org/CorpusID:4143192.

Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. The strain on scientific publishing. *Quantitative Science Studies*, 5:823–843, 2023. URL https://api.semanticscholar.org/CorpusID:263136473.

T. Jerabkova, F. Patat, D. Dorigo, F. Sogni, F. Primas, A. De Cia, and E. R. Hoppe. Distributed Peer Review at ESO: Demonstrating Success and Evolving Through Period 115. *The Messenger*, 194:33–36, March 2025. doi: 10.18727/0722-6691/5383.

Tereza Jerabkova, Ferdinando Patat, Francesca Primas, Dario Dorigo, Fabio Sogni, Lucas Astolfi, Thomas Bierwirth, and Michael Prümm. The First Results of Distributed Peer Review at ESO Show Promising Outcomes, 2023. URL https://doi.eso.org/10.18727/0722-6691/5316. ISSN: 0722-6691 Publisher: European Southern Observatory (ESO).

Wolfgang E. Kerzendorf, Ferdinando Patat, Dominic Bordelon, Glenn van de Ven, and Tyler A. Pritchard. Distributed peer review enhanced with natural language processing and machine learning. *Nature Astronomy*, 4(7):711–717, July 2020. ISSN 2397-3366. doi: 10.1038/s41550-020-1038-y. URL https://www.nature.com/articles/s41550-020-1038-y. Publisher: Nature Publishing Group.

Michail Kovanis, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS ONE*, 11, 2016. URL https://api.semanticscholar.org/CorpusID:9484241.

Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77 6:1121–34, 1999. URL https://api.semanticscholar.org/CorpusID:2109278.

Jisoo Lee, Jieun Lee, and Jeong-Ju Yoo. The role of large language models in the peer-review process: opportunities and challenges for medical journal reviewers and editors. *Journal of Educational Evaluation for Health Professions*, 22, 2025. URL https://api.semanticscholar.org/CorpusID:275705044.

Kevin Leyton-Brown, Mausam, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, and Dinesh Raghu. Matching Papers and Reviewers at Large Conferences, August 2022. URL http://arxiv.org/abs/2202.12273. arXiv:2202.12273 [cs].

H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4):309–317, October 1957. ISSN 0018-8646. doi: 10.1147/rd.14.0309. URL https://ieeexplore.ieee.org/document/5392697.

Michael R. Merrifield and Donald G. Saari. Telescope Time Without Tears: A Distributed Approach to Peer Review. *Astronomy & Geophysics*, 50(4):4.16–4.20, August 2009. ISSN 13668781, 14684004. doi: 10.1111/j.1468-4004.2009.50416.x. URL http://arxiv.org/abs/0906.1943. arXiv:0906.1943 [astro-ph].

Jennifer Donovan Meyer, Andrea Corvillón, John M. Carpenter, Adele L. Plunkett, Robert Kurowski, Alex Chalevin, Jakob Bruenker, D.-C. Kim, and Enrique Macías. Analysis of the ALMA Cycle 8 Distributed Peer Review Process. *Bulletin of the AAS*, 54(1), May 2022. doi: 10.3847/25c2cfeb.4ece85d4. URL http://arxiv.org/abs/2204.05390. arXiv:2204.05390 [astro-ph].

David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pp. 500–509, New York, NY, USA, August 2007. Association for Computing Machinery. ISBN 978-1-59593-609-7. doi: 10.1145/1281192.1281247. URL https://doi.org/10.1145/1281192.1281247.

Vicente Amado Olivo, Wolfgang Kerzendorf, Bangjing Lu, Joshua V. Shields, Andreas Flörs, and Nutan Chen. Practical author name disambiguation under metadata constraints: A contrastive learning approach for astronomy literature, 2025. URL https://arxiv.org/abs/2511.10722.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost

Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Bohan Zhang, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg,

Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wenting Zhan, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2026. URL https://arxiv.org/abs/2412.16720.

OpenReview. openreview/openreview-expertise, November 2025. URL https://github.com/openreview/openreview-expertise. original-date: 2018-08-02T18:12:33Z.

F. Patat, W. Kerzendorf, D. Bordelon, G. Van de Ven, and T. Pritchard. The Distributed Peer Review Experiment. *The Messenger*, 177:3–13, September 2019. doi: 10.18727/0722-6691/5147.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine Learning in Python, June 2018. URL http://arxiv.org/abs/1201.0490. arXiv:1201.0490 [cs].

Francesca Primas, Olivier Hainaut, Thomas Bierwirth, Ferdinando Patat, Dario Dorigo, Elisabeth Hoppe, Uwe Lange, Moreno Pasquato, and Fabio Sogni. The New ESO Phase 1 System for Proposal Submission, 2019. URL https://doi.eso.org/10.18727/0722-6691/5141. ISSN: 0722-6691 Publisher: European Southern Observatory (ESO).

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Marko A. Rodriguez and Johan Bollen. An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM08, pp. 319–328. ACM, October 2008. doi: 10.1145/1458082.1458127. URL http://dx.doi.org/10.1145/1458082.1458127.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents, 2012. URL https://arxiv.org/abs/1207.4169.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Martin Saveski, Steven Jecmen, Nihar B. Shah, and Johan Ugander. Counterfactual evaluation of peer-review assignment policies. *ArXiv*, abs/2305.17339, 2023. URL https://api.semanticscholar.org/CorpusID:258959302.

Karen Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval*, pp. 132–142. Taylor Graham Publishing, GBR, 1988. ISBN 0947568212.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review, November 2019. URL http://arxiv.org/abs/1806.06237. arXiv:1806.06237 [stat].

Ivan Stelmakh, John Wieting, Sarina Xi, Graham Neubig, and Nihar B. Shah. A Gold Standard Dataset for the Reviewer Assignment Problem, May 2025. URL http://arxiv.org/abs/2303.16750. arXiv:2303.16750 [cs].

Louis-Gregory Strolger, Sophia Porter, Jill Lagerstrom, Sarah Weissman, I. Neill Reid, and Michael Garcia. The Proposal Auto-Categorizer and Manager for Time Allocation Review at the Space Telescope Science Institute. *The Astronomical Journal*, 153(4):181, March 2017. ISSN 1538-3881. doi: 10.3847/1538-3881/aa6112. URL https://doi.org/10.3847/1538-3881/aa6112. Publisher: The American Astronomical Society.

Louis-Gregory Strolger, Jamila Pegues, Tegan King, Nathan Miles, Michelle Ramsahoye, Keith Ceruti II, Brett Blacker, and I. Neill Reid. PACMan2: Next Steps in Proposal Review Management. *The Astronomical Journal*, 165(5):215, April 2023. ISSN 1538-3881. doi: 10.3847/1538-3881/acc2c4. URL https://doi.org/10.3847/1538-3881/acc2c4. Publisher: The American Astronomical Society.

Antonio L. Teixeira. Ai in peer review: can artificial intelligence be an ally in reducing gender and geographical gaps in peer review? a randomized trial. *Research Integrity and Peer Review*, 10, 2025. URL https://api.semanticscholar.org/CorpusID:282386844.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Frank. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:196–202, 1945. URL https://api.semanticscholar.org/CorpusID:53662922.

Xuanang Xu, Joshua Yan, Gloria Nwachukwu, Hongming Shan, Uwe Kruger, and Ge Wang. Artificial intelligence-aided assignment of journal submissions to associate editors—a feasibility study on ieee transactions on medical imaging. *Visual Computing for Industry, Biomedicine, and Art*, 9, 2026. URL https://api.semanticscholar.org/CorpusID:284600986.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Dong Zhang, Shu Zhao, Zhen Duan, Jie Chen, Yangping Zhang, and Jie Tang. A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation, December 2019. URL http://arxiv.org/abs/1912.08976. arXiv:1912.08976 [cs].

Yue Zhao, Ajay Anand, and Gaurav Sharma. Reviewer recommendations using document vector embeddings and a publisher database: Implementation and evaluation. *IEEE Access*, 10:21798–21811, 2021. URL https://api.semanticscholar.org/CorpusID:237867076.

Appendices

A Data Details

A.1 Reviewer Statistics and Assignment

As outlined in the main text, the Distributed Peer Review process for the European Southern Observatory Period 110 (P110) utilized a final pool of 379 unique proposal-designated reviewers. Over 90% of these

reviewers were the proposal PI, while the remainder were either Co-PIs or designated proxies (e.g., doctoral students with supervision from supervisor). Regarding career stage, roughly 15% of the reviewer pool consisted of students, with the remainder being postdoctoral researchers or professional astronomers.

A.2 Baseline Keyword Matching Pipeline

In P110, the reviewer–proposal matching relied strictly on keyword-based similarity (Jerabkova et al., 2025). Reviewers provided keywords when creating their ESO User Portal profiles, and principal investigators assigned keywords to their submitted proposals. Across all proposals and reviewers in P110, there were 138 unique keywords chosen from ESO's standardized corpus[2]. ESO developed an internal algorithm (detailed in Appendix B) that transforms these selected keywords into vector representations to calculate similarity scores for all reviewer-proposal pairs. Based on these similarity scores, the system assigns ten experts to each proposal using the `peerreview4all`[3] optimization algorithm (Stelmakh et al., 2019; Jerabkova et al., 2025).

A.3 ESO Keyword Categories

In addition to specific sub-field keywords, the ESO proposal system organizes proposal and reviewer subjects into the following nine broad scientific categories. These categories serve as the high-level classification for all submissions:

1. Physical data and processes
2. Astrometry and celestial mechanics
3. The Sun
4. Planetary systems
5. Stars
6. Interstellar medium (ISM), nebulae
7. The Galaxy
8. Galaxies
9. Cosmology

A.4 Publication Data Retrieval

To construct the text-based expertise profiles for evaluation, we retrieved reviewer publication histories via the NASA Astrophysics Data System API. When querying for a maximum of 25 publications over the last five years, the mean number returned per reviewer was 22. Notably, over 68% of these queries returned the maximum limit of 25 publications. Because the API supports fine-grained filtering (e.g., by publication year, authorship position, or publication count), we experimented with multiple query strategies to evaluate how different constraints on a reviewer's publication history affect the resulting vector representations and downstream retrieval performance (see Appendix C.2).

B Representation Methods

This section details the mathematical formulations and algorithmic implementations for the expertise representation methodologies evaluated in the main text. We begin by outlining the baseline keyword-matching framework currently deployed by the ESO, followed by the probabilistic, statistical, and neural architectures utilized in our benchmark.

[2]See the full P110 keyword list at: `https://www.eso.org/p1demo/proposals/19859/keywords`
[3]`https://github.com/niharshah/peerreview4all`

B.1 ESO Keywords

The ESO DPR system represents both reviewers and proposals as keyword vectors representing self-reported scientific keywords. Each reviewer and proposal specifies between two and five keywords in decreasing order of relevance from a static vocabulary compiled by ESO [4] (Primas et al., 2019). Reviewer keywords are specified once during the creation of a reviewer's profile in the ESO portal and can be updated individually, however, it is unknown how frequently reviewers revise them. Reviewers are instructed to rank their keywords by relevance:

$$W = n_{\max} - \text{order} + 1,$$

where $n_{\max} = 5$. Thus, the first keyword receives a weight of 5, the second 4, and so on.

These weights populate a vector corresponding to the global keyword corpus, where all positions are zero except for those representing the selected keywords for a given proposal or reviewer. For example, if a reviewer selects "Stars: supernovae", "Stars: binaries", and "Galaxies: abundances" as their top three keywords, only those three positions in the vector receive nonzero weights (5, 4, and 3 respectively). The similarity between a reviewer (p_i) and a proposal (r_j) is then computed as the cosine similarity between the two vectors:

$$s_k = \frac{p_i \cdot r_j}{|p_i|\,|r_j|},$$

which outputs a score between 0 (i.e., no overlapping keywords) and 1 (i.e., identical keyword lists in the same order).

If the reviewer and the proposal have no shared keywords, a secondary category-level similarity (s_c) is computed. Each keyword belongs to one of several broad scientific categories (e.g., Stars, Galaxies, and Cosmology; see Appendix A.3 for the full list). Category vectors are built similarly to keyword vectors, using the position of the first appearance of each category as a weight.

The total matching score is defined as:

$$s_{ij} = s_k + s_c,$$

ranging from 0 (i.e., no keyword or category match) to 2 (i.e., perfect keyword, category, and order similarity). The total score serves as the metric to identify and rank reviewer-proposal pairs to assign ten reviewers to each proposal. In this work we utilize the keyword based similarity scores computed for P110 provided by ESO for evaluation as a baseline.

B.2 Topic Modeling with Latent Dirichlet Allocation

LDA (Blei et al., 2003) is a generative probabilistic model that represents documents as mixtures of latent topics and each topic is described by a distribution of words. Probabilistic models such as LDA extend beyond a static, predefined set of keywords by describing both reviewers and proposals through shared conceptual topics. Formally, LDA assumes that each document d is represented by a multinomial distribution over K topics drawn from a Dirichlet prior with hyperparameter α, where K is an input. Each topic k is itself a multinomial distribution of words, drawn from a Dirichlet prior with hyperparameter β. The vector θ_d represents the topic proportions for document d, serving as its numerical representation.

We determine the topic distributions for both reviewers and proposals using a jointly trained LDA model. In our preprocessing step, we concatenate all abstracts associated with a given reviewer into a single, comprehensive document (i.e., one document per reviewer) prior to model training. These reviewer-level documents are then combined with the individual proposal abstracts to create the final training corpus. By training the LDA model on this combined corpus, we directly infer the reviewer expertise vectors (θ_{r_j}) and the proposal topic vectors (θ_{p_i}) within the same latent topic space. We utilize the GENSIM library (Řehůřek & Sojka, 2010) for implementation. We preprocess the text by removing stop words (e.g., to, and, because, etc), as these high-frequency terms provide minimal semantic information and may skew the resulting topic distributions (Angelov, 2020). The model was trained on the union of all queried reviewer publication abstracts and proposal abstracts to ensure a comprehensive vocabulary. Determining the appropriate number of topics, K,

[4] See the keyword list at `https://www.eso.org/p1demo/proposals/19859/keywords`

is a critical step as too few topics may overly generalize and too many may fragment meaningful concepts. While the ESO keywords are organized into nine broad categories, we experiment with the sensitivity of K in Appendix C.2 to determine the optimal granularity for our topic model. The ALMA DPR system has adopted LDA to model reviewer expertise based on an investigator's previously submitted proposals (Carpenter et al., 2025).

B.3 Term Frequency-Inverse Document Frequency

TF-IDF leverages the statistical word information of a set of documents to construct feature vectors that reflect the importance of terms within a single document relative to the entire set of documents (Luhn, 1957; Sparck Jones, 1988). Formally, the TF–IDF weight for a term t in a document d is defined as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

We utilize the scikit-learn implementation (Pedregosa et al., 2018):

$$\text{TF}(t, d) = f_{t,d}$$

with $f_{t,d}$ denoting the number of times term t appears in document d. The inverse document frequency down-weights terms that appear frequently across the corpus while up-weighting rare terms specific to a particular context. We utilize the smoothed IDF formulation from the `scikit-learn` library, defined as:

$$\text{IDF}(t) = \log\left(\frac{1+N}{1+n_t}\right) + 1$$

where N is the total number of documents in the corpus, and n_t is the number of documents containing term t.

This method is utilized by the STScI Proposal Auto-Categorizer and Manager (PACMan) tool to represent reviewer expertise in its panel peer review process (Strolger et al., 2017). To achieve this, PACMan queries abstracts from NASA/ADS and computes the similarity between TF–IDF vector representations of reviewer publications and proposal content (specifically, the abstract and scientific justification) (Strolger et al., 2023).

In our experimentation the model was trained on the vocabulary constructed from only the proposal abstracts. For each reviewer, the expertise vector is constructed by concatenating all retrieved abstracts into a single document, and proposal vectors are created from their respective abstracts. Both resulting documents are then vectorized using TF-IDF. We experiment with several hyperparameters to optimize the representation quality. Specifically, we vary the n-gram range to capture both unigrams and bigrams, allowing the model to incorporate two-word expressions (e.g., "stellar evolution", "galaxy formation", or "white dwarf") that convey more domain-specific meaning than separate individual words (e.g., "white", "dwarf"). Additionally, standard English words (i.e., stop words) are removed to reduce noise (e.g., articles, prepositions, etc).

B.4 Transformer Representations

Transformer models extend beyond word-level statistics or topic-based approaches by capturing semantic relationships between words and phrases based on the surrounding context, enabling more sophisticated representations (Devlin et al., 2019). Leveraging direct semantic encoding via attention mechanisms (Vaswani et al., 2017), the Transformer architecture has been widely adapted for specific domains and tasks (Beltagy et al., 2019; Cohan et al., 2020).

In this work, we evaluate two distinct Transformer models: the science-tuned SPECTER2 (Cohan et al., 2020) and the general-purpose Sentence Transformer model (all-distilroberta-v1) (an extension of Sanh et al., 2019).

First, the SPECTER2 model is specifically designed for scientific document representation. It is initialized with SciBERT (Beltagy et al., 2019) and fine-tuned using a citation-based contrastive loss, enforcing that

papers citing each other have similar embedding representations. We utilize the Sentence-Transformers framework (Reimers & Gurevych, 2019) because it is specifically fine-tuned using a contrastive objective to produce semantically meaningful, fixed-length embeddings from full sentences, unlike base models optimized for word-level prediction making them useful for similarity comparisons. The specific model we utilize, all-distilroberta-v1, is trained on over one billion sentence pairs, optimized to capture broad semantic similarity across diverse domains.

A primary challenge in utilizing Transformer methods for expertise modeling is the fixed input context window, typically limited to 512 tokens, which is insufficient to embed a reviewer's full publication history simultaneously. To address this, we adopt a common two-stage aggregation approach. We independently encode each abstract in a reviewer's history into a vector $\mathbf{p}_i$, obtaining a set of abstract vectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$. We then apply mean or max pooling across these vectors to produce a single, centroid-based expertise vector $\mathbf{r}_j$:

$$\mathbf{r}_j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{p}_i$$

In addition to encoder models that represent text and output vector embeddings, we evaluate a generative machine learning model using the GPT-4o mini model accessible via the OpenAI API (OpenAI et al., 2024). To implement this, we construct a structured system prompt where the LLM acts as an expert in astronomical time allocation (see Appendix B.5). We retrieve the reviewer publication histories via the NASA/ADS API to construct the context window for each evaluation. Deploying LLMs for pairwise comparisons at the scale of astronomical observatories presents cost challenges. To address this, we leverage prompt caching since a single reviewer must be compared against the 435 proposals, re-sending the reviewer's full publication history for every comparison is redundant. Finally, to mitigate the non-determinism of generative models and ensure scientific reproducibility, we explicitly set the sampling temperature to $T = 0$ and fix the random seed (seed=42). Our framework leverages guard rails, context retrieval, and reproducibility constraints (e.g., harnesses (Yao et al., 2023)). We deliberately select GPT-4o mini over reasoning models (e.g., the GPT-5.5 series), as the latter utilize hidden chain-of-thought processes that currently restrict temperature control and introduce inherent non-determinism. This configuration maximizes the stability of the outputs, ensuring that the relevance scores remain consistent across repeated experimental runs. Initial exploratory tests utilizing the most recent generation of GPT models (i.e., reasoning-based models) yielded retrieval performance similar to GPT-4o mini.

B.5 GPT-4o mini Prompt Design

To estimate expertise using the GPT-4o mini model, we provided the system with the following zero-shot prompt. The model was instructed to output a scalar score between 0 and 100 based on the semantic alignment between the proposal abstract and the reviewer's publication history.

```
You are an expert in assigning reviewers to proposals at astronomical observatories.
You want to make sure that the reviewers can give high quality and relevant reviews to the
    proposal they are assigned.

You are given the following input:
- "REVIEWER PAPERS", which is a selection of the most recent papers by the reviewer,
    containing the title and abstract of each paper.
- "NEW PROPOSAL", which contains the proposal abstract that is under consideration for
    assignment.

Your task is to assign a score (0-100) evaluating how well the NEW PROPOSAL matches the
    REVIEWER'S PAPERS.

Consider the following criteria:
1. The score should be based on the similarity between the proposal and the reviewer papers.
2. The score should be higher if the reviewer has more background knowledge and expertise in
     the proposal.
3. The score should be lower if the reviewer has less background knowledge and expertise in
    the proposal.
```

```
Scoring Scale: Assign any integer score from 0 to 100.
NOTE: Output ONLY the score. Use integer or float. Do not hallucinate.
```

Listing 1: System prompt used for expertise scoring

B.6 NDCG Definition for Self-reported Expertise Evaluation

To evaluate the similarity scores alignment with the reviewer self-reported labels (e.g., Expert, Intermediate, Non-expert), we utilized the Normalized Discounted Cumulative Gain (NDCG) metric. Unlike standard recall metrics, NDCG is rank-aware, explicitly rewarding algorithms that confidently place experts at the top of the assignment list.

NDCG is built upon Discounted Cumulative Gain (DCG), which measures a ranking based on its position and a graded relevance score. For example, for a single proposal, DCG is computed as:

$$\mathrm{DCG}_k = \sum_{j=1}^{k} \frac{\mathrm{rel}_j}{\log_2(j+1)}$$

where rel_j is the graded relevance score (i.e., self-reported expertise labels) of the result at position j, and k is the total number of reviewers (e.g., 10 in ESO DPR). The formal definition for NDCG is:

$$\mathrm{NDCG}_k = \frac{\mathrm{DCG}_k}{\mathrm{IDCG}_k}$$

where IDCG_k is the maximum possible DCG calculated by placing all relevant reviewers in the ideal rank order. We utilize the scikit-learn implementation of NDCG (Pedregosa et al., 2018).

C Expertise Matrix Statistics and Ablation Studies

C.1 Expertise Matrices Statistics

Table 3 provides a statistical summary of the expertise matrices generated by each method. We report the sparsity (percentage of zeros) and the distribution quartiles to illustrate the varying density and scaling of the scores across different algorithms. All methods apart from the Keywords utilized up to 25 abstracts queried from NASA/ADS.

Table 3: Expertise Matrix Distribution Statistics

Method	% zeros	min	25th	median	75th	max
Keywords	51.78%	0.00	0.00	0.00	0.43	1.00
LDA†	67.93%	0.00	0.00	0.00	0.02	1.00
TF–IDF	0.66%	0.00	0.03	0.04	0.06	0.50
SPECTER2 (Mean pooling)	0.00%	0.75	0.91	0.92	0.94	0.98
SentenceTransformer (Mean pooling)	0.00%	-0.10	0.47	0.55	0.62	0.88
GPT–4o mini	10.19%	0.00	0.25	0.45	0.75	0.95

† Values below 0.01 for LDA are treated as zero.

C.2 Ablation Results

This section presents the full results of our ablation studies. Tables 4, 5, and 6 demonstrate the robustness of our retrieval performance under varying query parameters to NASA/ADS (e.g., timeframes and paper counts). Additionally, Table 7 examines the impact of varying the topic count K in the LDA algorithm, while Table 8 evaluates max pooling in the SPECTER2 and SentenceTransformer models.

We evaluate how different queries for publications from the NASA/ADS API affect retrieval performance. The API enables flexible querying along several dimensions, for example: (1) the maximum number of publications returned, (2) the publication time window, and (3) authorship position (e.g., first-author papers).

Table 4: Ablation 1: Average rank-based retrieval performance across 435 proposals when querying for 12 papers in the last 2 years

Method	Median Rank ↓	MRR ↑	Hit@25 ↑	Z-score ↑
Keywords	9.0 ±2.5	0.270 ±0.032	0.703 ±0.044	2.051 ±0.095
LDA (K = 50)	42.0 ±6.0†	0.093 ±0.018†	0.395 ±0.046†	1.672 ±0.279†
TF–IDF	**6.0** ±2.0	**0.340** ±0.035†	**0.720** ±0.043	**3.583** ±0.343†
SPECTER2 (Mean pooling)	15.0 ±5.5†	0.256 ±0.034	0.566 ±0.046†	0.778 ±0.111†
SentenceTransformer (Mean pooling)	9.0 ±2.5*	0.304 ±0.035	0.662 ±0.045	1.109 ±0.122†

Table 5: Ablation 2: Average rank-based retrieval performance across 435 proposals when querying for 50 papers in the last 10 years

Method	Median Rank ↓	MRR ↑	Hit@25 ↑	Z-score ↑
Keywords	9.0 ±1.0	0.275 ±0.033	0.701 ±0.043	2.051 ±0.096
LDA (K = 50)	19.0 ±4.0 †	0.166 ±0.025 †	0.568 ±0.046 †	2.540 ±0.238 †
TF–IDF	**4.0** ±1.0 †	**0.432** ±0.038 †	**0.807** ±0.037 †	**3.962** ±0.298 †
SPECTER2 (Mean pooling)	9.0 ±3.0 *	0.294 ±0.034	0.634 ±0.046 *	0.884 ±0.114 †
SentenceTransformer (Mean pooling)	6.0 ±1.5	0.352 ±0.037 †	0.720 ±0.041	1.377 ±0.122 †

Retrieving the correct publications for a reviewer is important as shorter time windows, smaller publication sets, and stricter authorship filters can reduce the amount and quality of information available for representing reviewer expertise.

Our baseline results use a query that retrieves 25 publications from the past five years, regardless of author position. To examine the sensitivity of the methods to changes in data retrieval, we evaluate three additional configurations: retrieving 50 publications from the last ten years, twelve publications from the last two years, and ten first-author publications from the last 5 years. These ablations allow us to assess how each representation method behaves under sparse versus more extensive publication histories.

Across all ablation settings, we observe a consistent trend: increasing the number of retrieved publications yields only marginal gains for all methods, whereas restricting the number of publications leads to degrading metrics. However, the magnitude of this sensitivity varies by methodology. TF-IDF and SentenceTransformer demonstrate robustness, exhibiting only minor performance declines with fewer publications. In contrast, LDA and SPECTER2 are sensitive with performance decreasing when insufficient publications are queried. The keyword-based method remains static, as it is independent of the retrieval parameters. Note that we omit GPT-4o mini from this ablation due to cost constraints across multiple experiments.

Table 7 presents the results of our ablation of the sensitivity of the LDA model to the hyperparameter K (number of topics). We observe a distinct performance peak at $K = 50$, which yields the best median rank of 24.0 and Hit@25 of 0.503. Lowering the number of topics decreases all metrics with $K = 15$ yielding the lowest scores, while increasing K to 75 reduces retrieval accuracy. Finally, Table 8 shows that the max pooling variants of SentenceTransformer and SPECTER2 exhibit the weakest overall performance. Both models recorded the lowest retrieval scores, yielding Median Ranks of 77.0 and 109.0, respectively.

Table 6: Ablation 3: Average rank-based retrieval performance across 435 proposals when querying for 10 first author papers in the last 5 years

Method	Median Rank ↓	MRR ↑	Hit@25 ↑	Z-score ↑
Keywords	9.0 ±2.5	0.270 ±0.032	**0.703** ±0.044	2.051 ±0.095
LDA (K = 50)	62.0 ±14.5†	0.091 ±0.019†	0.317 ±0.043†	1.338 ±0.220†
TF–IDF	**5.0** ±1.5	**0.424** ±0.041†	0.678 ±0.043	**4.643** ±0.490†
SPECTER2 (Mean pooling)	29.0 ±15.0†	0.297 ±0.038	0.494 ±0.047†	0.644 ±0.123†
SentenceTransformer (Mean pooling)	12.0 ±4.5†	0.356 ±0.040*	0.575 ±0.047†	0.736 ±0.129†

Table 7: Ablation 4: Average rank-based retrieval performance across 435 proposals when querying 25 publications and varying the number of K topics in the LDA algorithm

Method	Median Rank	MRR	Hit@25	Z-score
LDA (K = 15)	30.0 ±5.5	0.106 ±0.020	0.453 ±0.048	1.642 ±0.123
LDA (K = 25)	29.0 ±5.0	0.118 ±0.020	0.467 ±0.046	1.911 ±0.183
LDA (K = 50)	**24.0** ±6.0	**0.148** ±0.024	**0.503** ±0.047	2.114 ±0.208
LDA (K = 75)	28.0 ±6.5	0.143 ±0.025	0.483 ±0.046	**2.266** ±0.363

Table 8: Ablation 5: Average rank-based retrieval performance across 435 proposals when employing max pooling for Transformer models

Method	Median Rank ↓	MRR ↑	Hit@25 ↑	Z-score↑
SPECTER2 (max pooling)	109.0 ±23.5	0.035 ±0.013	0.110 ±0.030	0.256 ±0.101
SentenceTransformer(max pooling)	77.0 ±20.5	0.071 ±0.017	0.290 ±0.044	0.628 ±0.121

D Methodological Limitations

A key limitation of relying on the NASA/ADS API for automatically querying a researcher's publications is the lack of author name disambiguation. As Olivo et al. (2025) note, an estimated 52% of name groupings (i.e., first initial and last name) in NASA/ADS contain ambiguity, which can introduce noise into the expertise representations when researchers share similar names (Strolger et al., 2017). To mitigate this in a production system, observatories could adopt the approach used by platforms like OpenReview, where reviewers manually select their representative publications. This ensures accurate, up-to-date profiles while eliminating the algorithmic risks associated with querying digital library APIs. We verified the data completeness for the 379 researchers in our dataset by querying their last 25 publications over a five-year window. Despite the inclusion of student reviewers in the Distributed Peer Review process, only two individuals returned zero records. We resolved one case by extending the search window to ten years, and noted the other possessed only unrefereed publications. In a live production environment, the expertise of early-career reviewers returning zero publications could be accurately represented by utilizing the text of the proposal they submitted to the current cycle, guaranteeing a baseline expertise vector (Carpenter et al., 2025). Finally, utilizing a generative Large Language Model (e.g., GPT-4o MINI) introduces computational costs (≈$300-500 USD per experimental run in our benchmark). Furthermore, these models rely on probabilistic generation, which inherently limits the strict deterministic reproducibility required for a standardized, fair peer review pipeline.