fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e18, 1-25
DOI: 10.22163/fteval.2025.710

INSIDE THE FUNDING PROCESS: USING GENERATIVE AI TO ASSESS REVIEWERS' CRITERIA PRIORITISATION IN MULTI-STAGE APPLICATION ASSESSMENTS

PETER KOLARZ AND DIOGO MACHADO
DOI: 10.22163/FTEVAL.2025.710

ABSTRACT

When evaluating funding schemes with multiple aims (expressed through multiple assessment criteria such as quality, novelty, relevance, collaboration, etc.), there is an inherent challenge in assessing what role different assessment criteria play in the selection and awarding process. As part of a process evaluation of the Austrian FWF's Emerging Fields programme, we used generative AI to analyse peer reviewers' reports on applications submitted to the scheme. The purpose of this analysis was to understand how various assessment criteria were being operationalised in the review process. Specifically, the Emerging Fields scheme has two separate written application assessment stages: a short outline-proposal stage, followed by a full application review stage. Background research on the scheme's design led us to a hypothesis that reviews in the first of these two stages should emphasise and reward innovative potential and novelty of the proposed project ideas, while reviews in the second stage should place a greater emphasis on scientific quality of the research plans. Our analysis of reviews at both assessment stages used generative AI to assess, first, what priority various criteria had in terms of the amounts of text devoted to each criterion. Second, the analysis assessed the extent of positive or negative sentiment with which

reviews covered each criterion (specifically, the criteria covered were scientific quality, novelty, feasibility, and team qualifications). We then compared the results in all combinations between passed applications, failed applications, stage 1 applications and stage 2 applications. Results largely substantiated our hypotheses, most notably with novelty having a significantly higher priority at stage 1 and scientific quality having a significantly higher priority at stage 2. This analysis added substantial value to the Emerging Fields evaluation and has potential to become an invaluable tool to test the effectiveness and appropriateness of research funding processes, particularly in complex multi-stage designs with multiple scheme objectives.

Key words: Research funding, peer-review, generative AI, evaluation

INTRODUCTION

Research funding schemes may have several different aims. These include (but are not limited to), funding basic exploratory research, research that is societally relevant (including to specific societal challenges), research that has particular promise for industrial application, research that is highly innovative and breaks with established paradigms, research that presents new interdisciplinary or transdisciplinary perspectives, or research that seeks to foster collaboration between previously unconnected individuals.

Evaluating funding schemes with diverse objectives requires understanding of how well their processes address complex and varied assessment criteria. Especially in cases of schemes with multiple aims – and consequently, multiple assessment criteria – process evaluations of funding schemes face the challenge of having to assess to what extent application selection processes consider and reward these various aspects of proposed research projects. Put differently: how well do the review processes actually operationalise the aims of the funding instrument?

In recent years, widespread modifications have been made to standard research grant assessment processes, which traditionally involve peer and expert panel reviews. These include short pre-proposals, the inclusion of non-academic reviewers or panellists, and in-person presentations (Kolarz et al., 2023). Funders often introduce such modifications to ensure that the assessment captures a range of different aspects of applications submitted to a scheme. These changes aim to improve the alignment of review processes with the diverse aims of funding schemes, such as fostering interdisciplinarity

or rewarding high-risk, high-reward proposals and address the limitations of conventional peer review, such as biases against highly innovative or interdisciplinary work (OECD, 2021; Criscuolo et al., 2017; Laudel & Gläser 2014; Luukkonen, 2012; Boudreau et al., 2012; Langfeldt, 2006). These evolving practices highlight the growing importance of adapting review processes to better reflect and support the multifaceted goals of research funding schemes.

At the same time, advances in artificial intelligence (AI) present new opportunities to further support and refine the assessment process for funding schemes. Generative AI tools have recently been explored for their potential to enhance peer review processes, from streamlining the review workflow to providing adaptive instructions and automating aspects of assessment. These tools can improve the standardisation and transparency of reviews, particularly when used to identify patterns or summarise reviewer feedback, though they also raise challenges, including confidentiality and ethical considerations (Su et al., 2023; Wong, 2023). Beyond peer review, generative AI has demonstrated value in managing large volumes of data, synthesising evidence, and producing detailed reports, as seen in applications like health technology assessments (Fleurence et al., 2024).

Funders and evaluators must assess how different criteria are featured in application reviews, and whether the process prioritises the funding instrument's aims. For instance, the effectiveness of early-stage assessments in filtering for originality has been well-documented in funding schemes that emphasise novel and unorthodox research approaches (Kolarz et al., 2023; Morgan et al., 2020). Similarly, AI-based tools can play a complementary role in capturing nuanced reviewer priorities, enhancing transparency, and reducing the likelihood of overlooking critical dimensions during assessments. Ultimately, understanding whether the assessment process aligns with the objectives of the funding instrument – and how emerging technologies like AI can support this alignment – is becoming increasingly important in shaping and improving the practices of peer review and evaluation.

CONTEXT: EVALUATION OF THE FWF'S EMERGING FIELDS SCHEME

The Austrian Science Fund's (FWF) Emerging Fields (EF) programme seeks to fund particularly original, innovative, paradigm-shifting research. Launched in 2022 as part of the Excellent=Austria initiative, the first call of the Emerging Fields (EF) Programme attracted 45 applications from a range of

multidisciplinary Austrian research teams and five projects were ultimately funded.

The call for applications for the EF programme was explicitly open to inter- and transdisciplinary proposals. However, these were not strict eligibility requirements because inter- or transdisciplinary ideas are not inherently novel. Proposals may draw on multiple disciplines that reflect established or already integrated approaches without offering a particularly original or emergent combination. Therefore, the programme only encouraged inter- and transdisciplinarity but did not require it, and it asked applicants and reviewers to focus on how each proposal's approach contributed to the novelty of the research.

Following the programme's decision about its primary focus, our analysis did not treat interdisciplinarity as a standalone evaluative dimension. Instead, the elements of interdisciplinary and transdisciplinary indirectly integrate broader categories such as novelty or team qualifications – particularly where reviewers noted the diversity of perspectives, methodological blending, or team composition.

This evaluation accompanied the first call of the EF scheme end-to-end (Excluding the initial application submission window, so the study itself ran from February 2023 to March 2024). It was tasked critically to review the entire process, to identify strengths and weaknesses in its design, and to provide evidence-based recommendations to the FWF and its supervisory bodies on how to improve these processes for the next EF call. The full evaluation report and methodological annex materials are available on the FWF's website.

THE EF SELECTION PROCESS INVOLVES THE FOLLOWING STEPS:

- First, review of 3-page "synopses", i.e. short outlines of the project idea, by an international Jury of 16 experts. Interviews and other scoping work conducted as part of the early stages of the evaluation showed that stage 1 was intended primarily to assess and reward the novelty of the proposed research ideas
- Second, peer review by at least 3 external reviewers. The evaluation's scoping work showed that stage 2 was intended primarily to assess scientific quality in a conventional sense (feasibility, robustness, etc)

- Third, a presentation by shortlisted applicants to the international jury. The evaluation's scoping work showed that stage 3 was intended primarily to assess team composition. We note that the third assessment stage is out of the scope of this paper as it did not involve written review material in the way the other two stages did, and was too small in terms of application numbers to lend itself to meaningful analysis

The intentions behind each stage obtained through the scoping research on the scheme's design led us to a hypothesis that reviews in the first of these two stages should emphasise and reward innovative potential and novelty of the proposed project ideas, while reviews in the second stage should place a greater emphasis on scientific quality of the research plans.

Prominent international scientists reviewed the stage 1 synopses and stage 2 full proposals for the EF programme and produced evaluation documents detailing their judgements. In total, we had access to 140 records: 87 peer-reviews of synopses and 53 reviews of full proposals.

Given the number, heterogeneity and complexity of the review documents, generative AI was particularly useful in facilitating a systematic assessment. The multidisciplinary nature of the EF programme and the fact that EF applications are at the frontier of science make the review documents far from digestible for a general audience. They are heavy on scientific jargon and technical details, which are hard to understand for someone without deep expertise in each topic.

APPROACH/METHOD

GENERATIVE AI CLASSIFICATION AND REVIEW

We used generative AI to analyse peer reviewers' reports on applications submitted to the FWF's Emerging Fields (EF) programme. In total, we processed 140 review reports: 87 from the 3-page synopses in the first assessment stage and 53 from the full proposals in the second stage. We used a rich and comprehensive generative AI model that navigated these technical details to find relevant individual insights about each review document and stylised facts about the selection process in its two stages. The model was OpenAI's GPT4 large language model, accessed programmatically via a dedicated API. This access mode ensures the privacy and confidentiality of the underlying data.

It also enabled us to explore the capabilities of the GPT4 model in large-scale automation (querying all the documents programmatically without manually inputting and querying each review document individually).

Our focus in this analysis was explicitly on the reviewers' perspective. We did not analyse the proposals' text – whether synopses or complete applications – but examined how reviewers articulated their assessments of key criteria. As such, our results capture patterns in how novelty, scientific quality, feasibility, and team qualifications were perceived and expressed during the review process without making claims about the underlying content of the proposals.

To test our hypothesis, we conducted a comprehensive analysis using topic detection, sentiment analysis, and priority detection. These methodologies used the text of peer review documents from both, the first (synopses) and second (full proposals) assessment stages, as input data. Each approach targeted specific aspects of the review process, enabling a detailed exploration of the evaluative dimensions.

Topic detection aimed to identify and categorise distinct sections of the review text, focusing on four core dimensions: each proposal's novelty, risk or feasibility, scientific quality or rigour, and the research team's qualifications or suitability. The topic detection ensured a systematic examination of key aspects addressed by reviewers. This process enabled a systematic examination of the areas, most relevant to the assessment of proposals, providing insights into how reviewers prioritised and discussed these dimensions. For instance, identifying text specific to "novelty" allowed us to assess the emphasis reviewers placed on the innovative potential of the proposal, while sections addressing "risk or feasibility" highlighted their concerns or confidence in the project's practical execution. Similarly, isolating commentary on "scientific quality or rigour" shed light on reviewers' perceptions of the methodological soundness of the work, and text discussing "team qualifications or suitability" provided insights into the perceived capability of the research team to deliver on the proposed objectives.

Sentiment analysis evaluated the tone of reviewers' feedback on each dimension. It determined whether their observations were positive, negative, or neutral, providing an additional layer of understanding regarding the reviewers' perspectives. This analysis was complemented by explanations and direct quotes from the review documents, ensuring transparency and enabling cross-validation with the original text. The sentiment results were particularly valuable for sense-checking the robustness of the analysis. A fundamental assumption was that successful applications should be

associated with more positive reviewer sentiments across these dimensions. If this relationship did not hold – if, for example, successful proposals received consistently negative or neutral sentiment – it would raise concerns about the quality and robustness of the model's outputs. We verified that the AI's classifications aligned with expected patterns by comparing sentiment scores to application outcomes, further validating the approach. This process ensured that the sentiment analysis was descriptive and instrumental in evaluating the trustworthiness and accuracy of the generated insights.

Finally, priority detection quantified the emphasis placed on each dimension by measuring the text length or word count dedicated to it. This served as a proxy for the time and effort reviewers allocated to discussing each aspect, revealing their implicit priorities during the evaluation process. By examining the relative word count devoted to each dimension, priority detection revealed the implicit priorities of reviewers during the evaluation process. For example, a higher word count for "novelty" might indicate that reviewers considered the proposal's innovative potential particularly significant. At the same time, more extensive commentary on "risk or feasibility" could suggest a greater focus on practical concerns. Similarly, longer sections addressing "scientific quality" or "team qualifications" highlighted areas where reviewers invested the most effort to assess the proposal's merit.

The priority detection method provided insights into reviewers' implicit weighting of evaluation criteria and served as a complementary tool to topic detection and sentiment analysis. By triangulating these results, we better understood the review process and how reviewers allocated their attention to various dimensions. Priority detection also enabled comparisons across proposals, offering a standardised way to interpret and analyse the focus of reviewer feedback.

Our prompts were designed to guide the model's responses, ensuring the outputs' precision, clarity, and accountability. For topic sentiment analysis, the model made clear judgements (positive, negative, neutral, or not discussed) for each dimension. We requested concise explanations of up to 100 words to support these judgements, accompanied by direct quotes from the reviewers' text. This approach ensured transparency and allowed for human cross-validation. For priority detection, the prompts instructed the model to rank topics by their importance based on the number of words dedicated to each. The output included a clear ranking and numerical word count estimates for each topic, following a predefined format to ensure consistency and interpretability. The following boxes contain the structured prompts we

deployed to ensure the model's outputs were consistent and interpretable, enabling robust downstream analyses while maintaining accountability.

BOX 1. SYSTEM MESSAGE (UNIVERSAL)

Users will provide text from scientific reviews. These reviews follow a common section template but may include different responses in each section. Answer questions using only information provided in the reviews. Expect users to use the following format:
Review: ***text from the review here***
Question: ***user question here***

Source: Technopolis-Group

BOX 2. TOPIC SENTIMENT PROMPT

Is the reviewer positive, negative, or neutral regarding the proposal's [topic]?

Response format:
Positive/Negative/Neutral/Not discussed. Explanation (maximum 100 words, always include quotes from the reviewer's text).

Source: Technopolis-Group

BOX 3. PRIORITY DETECTION PROMPT

Sort the following topics by importance based on the number of words the reviewer dedicates to each topic. Provide the ranking alongside an estimate of the total number of words per topic. Topics are:
** A. Novelty,
** B. Risk/Feasibility,
** C. Scientific Quality/Rigour,
** D. Qualifications/Suitability of the Team.

##Response format:

[A/C/B/D (104/51/50/20)]
##Note:

** Letters indicate the rank of topics from most to least important.
** Parentheses contain the estimated word count for each topic.
** Provide only the output in brackets; no additional text or commentary.

Source: Technopolis-Group

To address the potential for hallucinations inherent in the black-box nature of AI, we prioritised accountability and transparency in every stage of the analysis. The unpredictable tendency of generative AI models to produce outputs not grounded in the input data required rigorous safeguards to maintain the reliability of results. For each response generated by the model, we explicitly requested reasoning that provided a clear rationale for the decision, accompanied by direct quotes from the original review text, used as the basis for the output. This approach ensured that the model's decisions were grounded in the reviewers' actual statements, avoiding speculative or fabricated content.

We also conducted systematic sense-checking of the outputs by comparing them to real-world examples, such as those presented in Table 1 and Table 2. These examples demonstrate how the model accurately generated explanations for positive and negative assessments of the proposals' novelty, ensuring that its outputs aligned with the source text. Positive examples (e.g. Proposals X, Y, Z) illustrate how the model identified novelty and transformative potential in reviewers' comments. Conversely, negative examples (e.g., Proposals D, E, F) show the model identifying a lack of novelty or incremental contributions in the reviewers' assessments.

The model provided interpretable and verifiable outputs by incorporating direct quotes from the reviewers and framing its responses in a structured format. We manually cross-validated a 20% sample of these outputs, enabling us to agree or disagree with the AI's interpretations and intervene when necessary. In this process, we confirmed that in all sampled cases, the resulting quotes were not hallucinations, and all of the results were reasonable, with just a couple of cases which could be non-obvious/open to interpretation but not patently wrong. The dual-layered process of automated output generation and our human validation ensured robust alignment between the model outputs and the underlying data, significantly mitigating the risk of errors or unsupported conclusions. The tables of examples below further illustrate this process, showcasing how the model's structured approach maintained accuracy and transparency in its outputs.

Table 1. Sense-checking outputs from the LLM (positive examples)

Document	Is the reviewer positive, negative or neutral regarding the proposal's novelty? Answer in the following format: Positive/ Negative/Neutral/Not discussed. Explanation (max. 100 words and always include quotes from the reviewer's text).
Proposal X	Positive. The reviewer states that the proposal is "innovative and extremely difficult but totally right" and that the researchers are 'trying to give birth to a new method of analysis which they call the '[method X]' (…), not by simply carrying out a conventional multidisciplinary research with three different disciplines, but by creating a new research field, a sort of new (and needed) interdisciplinary field". This indicates that the reviewer sees the proposal as bringing novelty to the field.
Proposal Y	Positive. Explanation: The reviewer states that the proposal "has the potential to completely transform [topic 1], [topic 2], and perhaps [topic 3]", indicating a belief in its novelty and transformative potential. They also mention that "the groundwork has already been laid over the past decade", suggesting that the proposal is building on recent advances to potentially create a paradigm shift. The reviewer's enthusiasm is evident in the statement, "This is an impressive proposal. I rank it as the best proposal I have ever had the opportunity to review".
Proposal Z	Positive. Explanation: The reviewer states that the project "has a tremendous potential to generate a quantum leap in [topic X]" and commends the applicants for an idea "that has the potential to fundamentally challenge the field of [topic X] and established notion of research". The reviewer also mentions that "a paradigm shift is likely to emerge", indicating a strong belief in the novelty and transformative potential of the proposal.

Source: Technopolis-Group

Table 2. Sense-checking outputs from the LLM (negative examples)

Document	Is the reviewer positive, negative or neutral regarding the proposal's novelty? Answer in the following format: Positive/ Negative/Neutral/Not discussed. Explanation (max. 100 words and always include quotes from the reviewer's text).
Proposal XX	Negative. Explanation: The reviewer states, "The idea of [research question X] is not a novel idea, it is in fact one of the most sought-after challenges being pursued at an international level". They also mention that "many of the proposed topics come across as incremental to each of the respective PIs" and that "the high-risk high gain for the project as a whole is missing". These statements indicate that the reviewer does not find the proposal's novelty to be particularly strong or transformative.
Proposal YY	Negative. Explanation: The reviewer states that the goal of [work package] "is an ambitious goal, albeit not novel in itself", indicating that the idea is not new. Furthermore, the lack of a clear focus in the proposal as a whole makes it difficult for the reviewer to comment on its novelty, suggesting a negative view on the proposal's potential for novelty.
Proposal ZZ	Negative. Explanation: The reviewer states that "the general idea is not of particular novelty, as [topics X/Y] are now widely applied to [field Z]", and that "It is difficult to imagine how they will lead to a paradigm shift". They also mention that "this proposed project will not be much different" from other projects at large in [field Z] centers, suggesting that the proposal lacks the originality required for a significant impact on the field.

Source: Technopolis-Group

We combined these methodologies to capture qualitative and quantitative insights into the review process. This multifaceted approach enabled us to understand better how reviewers balanced their assessments across different proposal dimensions, such as novelty, risk/feasibility, scientific quality/rigour, and team qualifications/suitability. By systematically integrating the outputs from topic detection, sentiment analysis, and priority detection, we could identify trends and patterns in evaluative criteria and their relative emphasis.

TRANSFORMATION OF QUALITATIVE INSIGHTS INTO QUANTITATIVE SCORES

To facilitate quantitative analysis, we transformed the outputs of the sentiment and priority detection processes into numerical scores. This allowed for a more structured comparison across proposals and reviewers.

In the sentiment indicators, for each dimension d (novelty, risk/feasibility, scientific quality, and team qualifications), the final metric consisted on the following rule:

$$S_{i,d} = \begin{cases} 1, & if\ sentiment\ is\ positive\ for\ dimension\ d\ in\ topic\ i \\ 0, & otherwise \end{cases}$$

Where:

- $S_{i,d}$ *is the sentiment score for dimension d in document i, and*
- $d \in \{novelty, risk, quality, team\}$

This binary scoring approach ensured that only positive sentiments were counted, clearly distinguishing between favourable and neutral/negative evaluations.

Priority detection quantified each dimension's prominence by inverting the topics' ranking based on the word count dedicated to them. The highest-ranked dimension (i.e. the one with the most words) received the maximum value, with subsequent dimensions receiving progressively lower scores. The scoring rule was defined as:

$$P_{i,d} = 4 - R_{i,d}$$

- $P_{i,d}$ *is the priority score for dimension d in document i, and*
- $R_{i,d}$ *is the rank of dimension d based on word count in document i (1 for the most words, 4 for the least words).*

As a result, the dimension with the highest prominence received a score of 3, the second most prominent received 2, the third received a score of 1, and the least prominent scored 0.

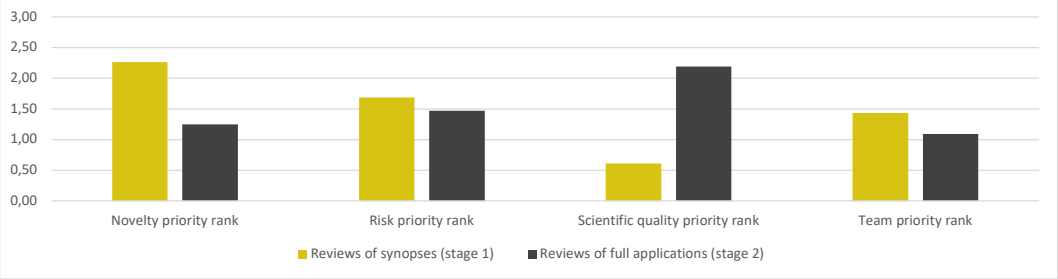
This dual-layered scoring system allowed us to perform statistical analyses and identify patterns, in how reviewers emphasised and evaluated key dimensions of the proposals. Combining sentiment and priority indicators enabled us to develop a detailed framework that facilitated a robust, data-driven understanding of the review process and the evaluation criteria prioritised by reviewers. These scores also enabled comparisons across proposals, further enhancing the interpretability of the results.

RESULTS

Our findings support our main hypotheses, highlighting significant differences in the priorities reviewers assign to proposal dimensions across the two evaluation stages. Most notably, novelty emerges as a significantly higher priority during stage 1 (synopses review), while scientific quality becomes the dominant focus in stage 2 (full proposals).

The analysis of priority rankings reveals a clear shift in focus between the stages. At the synopses stage, reviewers prioritise the novelty of the proposals, reflected by an average ranking score of 2.26 out of a possible 3. In contrast, novelty receives less emphasis during the full proposal stage, where its average ranking score drops to 1.25. Conversely, scientific quality receives minimal attention at stage 1, with an average score of 0.61, but becomes the highest-ranking dimension at stage 2, scoring 2.19. These results align with the expectations that stage 1 primarily evaluates proposals' originality and innovative potential. In contrast, stage 2 involves a more detailed examination of the proposal's scientific robustness and methodological quality (see Figure 1).

Figure 1 Average priority ranking scores for synopses (stage 1) and proposals (stage 2)



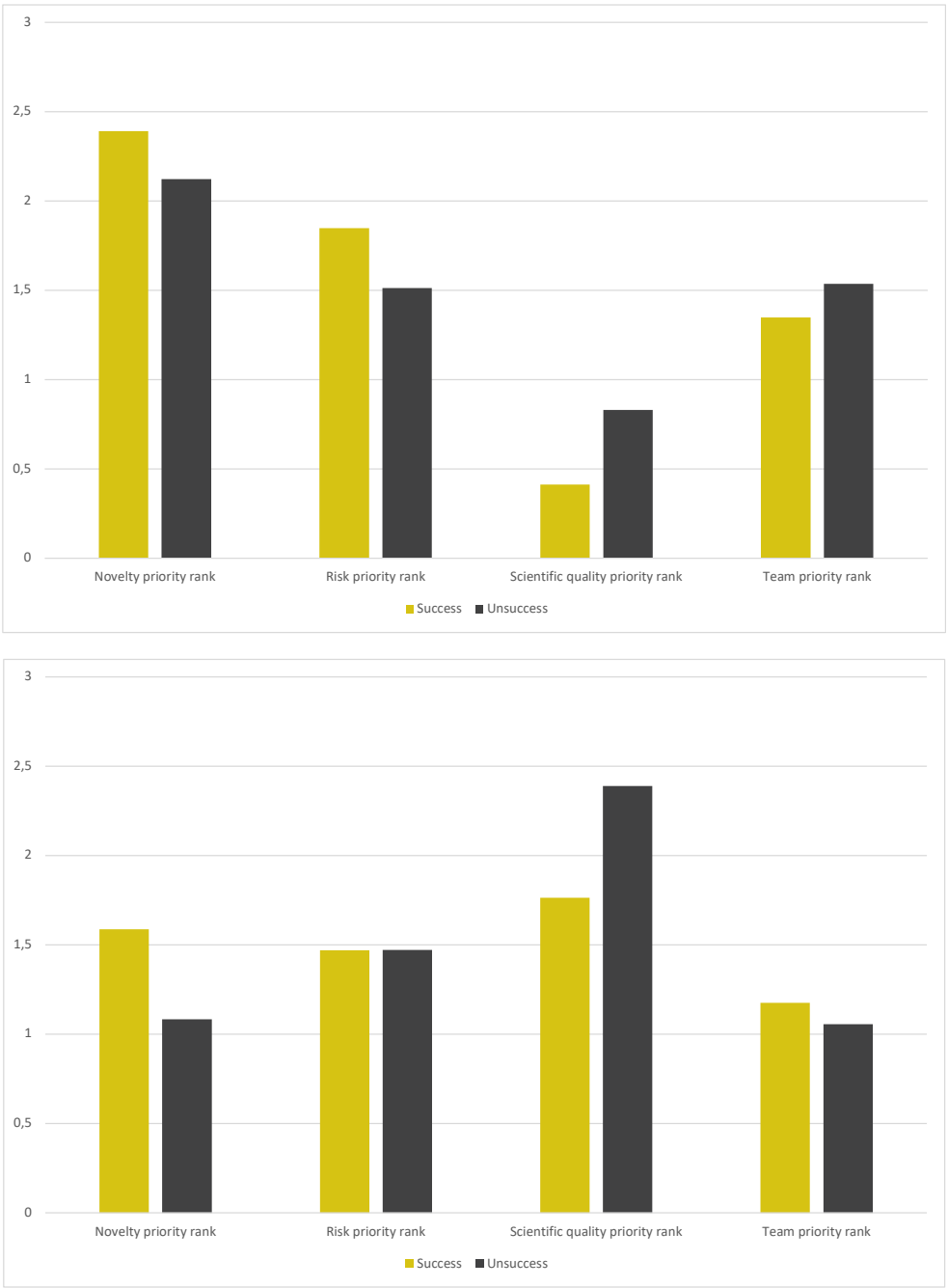
Source: Technopolis Group analysis based on peer-review documents

The dimensions of risk and team qualifications also receive greater attention at stage 1 than at stage 2, albeit with less pronounced differences compared to novelty and scientific quality. This suggests that reviewers dedicate more effort to assessing the feasibility and team capabilities during the initial stage, likely to gauge whether the proposal warrants further consideration.

Additional patterns emerge when we examine the breakdown of priority rankings between successful and unsuccessful applications. Figure 2 illustrates that successful applications consistently score higher in novelty and team qualifications, while unsuccessful applications exhibit slightly higher risk and scientific quality scores. This suggests reviewers may emphasise novelty and team strength when advancing proposals through the selection

process. In contrast, unsuccessful applications are likely to undergo more rigorous scientific quality and feasibility scrutiny, possibly reflecting a focus on identifying methodological shortcomings. These observations further support the idea that the two evaluation stages serve distinct functions, with the synopses stage acting as a filter for innovative and promising proposals. In contrast, the full proposal stage ensures that the selected projects meet high scientific standards.

Figure 2. Average priority ranking scores for successful and unsuccessful synopses and proposals



Source: Technopolis Group analysis based on peer-review documents

The sentiment analysis provides insight into how reviewers assess successful versus unsuccessful proposals. As illustrated in Figure 3, successful proposals consistently exhibit higher positive sentiment across all dimensions, with particularly notable differences for novelty and team qualifications. For novelty sentiment, successful proposals score an average of 0.7, significantly higher than the 0.3 average for unsuccessful applications. Sentiment towards risk also shows a marked difference, with successful proposals scoring 0.5 compared to 0.2 for unsuccessful ones. Similarly, scientific quality sentiment reveals a substantial gap, with successful proposals achieving an average score of 0.67 while unsuccessful applications score only 0.24. Finally, the review of team qualifications shows similar patterns, where successful proposals score an average of 0.82 compared to 0.58 for unsuccessful applications.

These findings suggest that reviewers consistently view successful proposals more favourably across all dimensions, emphasising the importance of novelty, risk management, and team strength in advancing proposals through the selection process. The pronounced differences in sentiment highlight these dimensions' critical role in shaping evaluation outcomes and align with the observed priorities for successful applications.

Figure 3 Average sentiment score for synopses (stage 1)

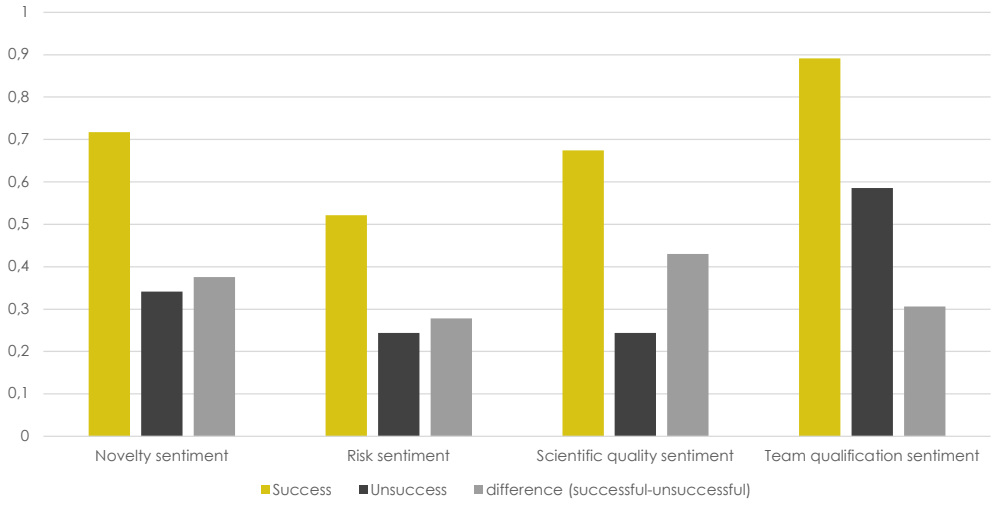
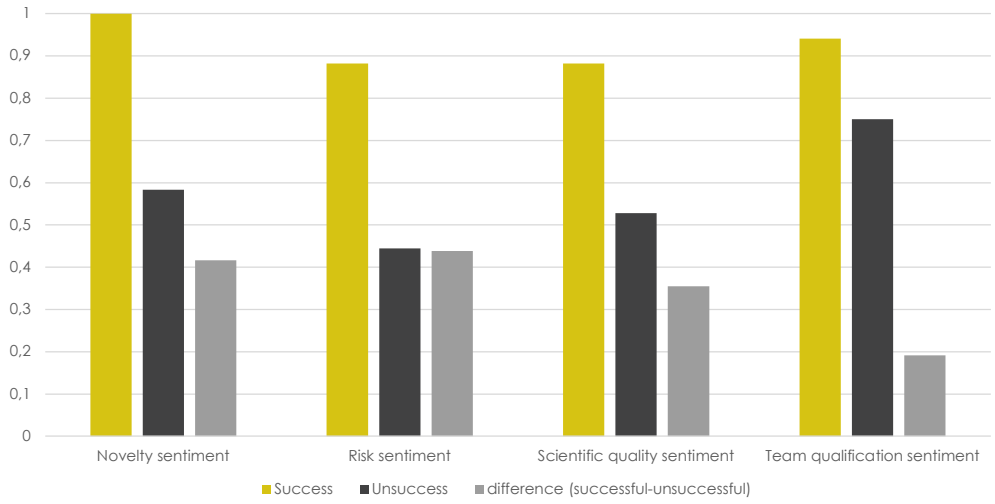


Figure 4 shows how reviewers assess successful versus unsuccessful proposals during the proposal stage (phase 2). Successful proposals consistently generate more positive sentiment across all dimensions, with particularly notable differences in novelty and risk. For novelty sentiment, successful proposals score an average of 1.00, significantly higher than the 0.60 average for unsuccessful applications. Similarly, sentiment towards risk shows a

marked difference, with successful proposals scoring 0.93 compared to 0.48 for unsuccessful ones. Scientific quality sentiment also reveals a substantial gap, with successful proposals achieving an average score of 0.93, while unsuccessful applications score only 0.55. Finally, team qualification sentiment shows the smallest difference but remains significant, with successful proposals scoring an average of 0.96 compared to 0.76 for unsuccessful applications.

In comparing the sentiment ratings between the synopsis and the full proposals, the overall trend is that successful proposals receive higher sentiment scores. However, the magnitude of these differences varies. For instance, the gap in novelty sentiment is slightly larger when evaluating the full proposal (0.40) compared to the synopsis (0.38), and the difference in risk sentiment shows an even more pronounced increase from 0.28 in the synopsis to 0.45 in the full proposal. In contrast, scientific quality sentiment is greater when only the synopsis is assessed (0.43) than when the full proposal is reviewed (0.38). Similarly, team qualification sentiment has a larger gap for the synopsis (0.31) than for the full proposal (0.20). While successful proposals consistently score higher across both stages, the specific extent to which success and failure separate in terms of novelty and risk sentiment becomes more pronounced at the full proposal stage. In contrast, differences in scientific quality and team qualification sentiments are more prominent when evaluators focus on the synopsis alone.

Figure 4. Average sentiment scores for proposals' evaluation (stage 2)



Source: Technopolis Group analysis based on peer-review documents

The comparison of priority rankings and sentiment differences between the synopsis and proposal stages highlights a clear evolution in the evaluation focus.

Novelty is central to reviewers' assessments during the synopses stage, as reflected by its high priority rank (2.26), but it diminishes in relative importance at the proposal stage (1.30). This shift suggests that reviewers emphasise other dimensions more once the initial novelty threshold is met. By contrast, scientific quality experiences a substantial increase in prioritisation, with its rank rising from 0.61 in the synopses stage to 2.17 in the proposal stage. This finding aligns with the observed sentiment patterns and underscores the importance of deeper scrutiny of methodological robustness and scientific rigour as proposals progress through the evaluation process. While important across both stages, risk shows a relatively stable ranking (1.69 at the synopsis stage and 1.46 at the proposal stage) despite an increase in sentiment differences at the proposal stage. This indicates that risk considerations remain a consistent component of the review process, with sentiment differences reflecting the level of engagement with feasibility and risk mitigation as proposals advance. Finally, while receiving high positive sentiment at both stages, team qualifications are comparatively less prioritised, especially during the proposal stage. This suggests that while team strength is acknowledged as an important factor, it plays a more supportive role relative to dimensions such as novelty, risk, and scientific quality.

These findings demonstrate a structured progression in the review process, with novelty playing a crucial filtering role at the synopsis stage and scientific quality emerging as the dominant consideration during the proposal stage. Risk remains a steady focus throughout, while team qualifications, despite positive sentiment, receive comparatively less prioritisation. These shifts reflect the nuanced and evolving priorities of the evaluation process, ensuring that innovative potential and scientific rigour are appropriately assessed at different stages.

TRIANGULATION WITH OTHER EVALUATION APPROACHES

We note that the evaluation of the Emerging Fields scheme also included other method components, some of which we can draw on to further enhance confidence in our results. The full evaluation methodology can be found in the annex sections of the evaluation (Kolarz et al., 2024). However, the following two are relevant to our findings:

The evaluation included in-person observation of the FWF Scientific Board meetings that followed stages 1 and 2. In each case, the Scientific Board discussed the reviews of the synopses/full applications and the evaluation team coded each discussion point thematically and in terms of sentiment. This exercise showed that for stage 1, Board meeting discussions focused strongly on novelty and at stage 2 far more on scientific quality. The Board discussions are of course separate from the written reviews themselves, but the reviews are the main material informing the Board discussions. While there is room here for alternative explanations, it is certainly plausible that the Board discussions were shaped by the reviews. In other words, the strong focus on novelty-aspects in Board discussions on stage 1 synopses was likely at least in part due to an emphasis on novelty in the reviews available to the Board.

Secondly, the evaluation team conducted a survey of individuals who acted as external reviewers for stage 2 full applications. One survey item asked them to self-assess which of a list of criteria they emphasised most in their judgement of the application they reviewed. Reviewers judged themselves to have emphasised criteria around scientific quality over all other aspects. This gives us confidence that the stage 2 reviews of full applications were indeed centrally about scientific quality rather than novelty.

Neither of these two method components allows us to look inside the reviewing process as directly as our analysis presented in this paper, so neither can be a substitute method capable of generating the type of findings presented in this paper. However, we note them here as an additional mechanism of triangulation underscoring the robustness of our findings.

CONCLUSION/DISCUSSION

This study provides evidence for the effectiveness of a two-stage review system in research funding processes, particularly for programmes emphasising scientific novelty. In the case of the Austrian FWF's Emerging Fields (EF) scheme, the two-stage approach – first requiring applicants to submit a short synopsis, followed by a full proposal for shortlisted candidates – has proven beneficial in multiple ways. Moreover, our study demonstrates that generative AI can contribute to a more systematic and transparent review of the evaluation process. This conclusion highlights two major insights: (1) how the initial stage of short synopses helps ensure that novelty is given a strong and meaningful focus, and (2) the potential of using generative AI to enrich and support the peer-review process, capturing nuanced reviewer priorities and

sentiments while reducing the risk of overlooking critical dimensions.

One of the most salient insights from this analysis is that the emphasis on novelty – an essential ingredient of pioneering research – emerges strongly during the first stage of the EF scheme. Indeed, when reviewers assess shorter proposal documents, they often devote significant attention to how innovative or groundbreaking a proposed project might be. A condensed synopsis, typically only a few pages long, appears to be particularly conducive to focusing the reviewer's attention on the "big idea": the conceptual leap or novelty that distinguishes one proposal from another. There are several possible reasons for why this emphasis on originality is pronounced at the first stage:

- Information economy: Short synopses can limit the amount of detail applicants can provide, naturally directing reviewers' attention to the headline objectives, the main research question or hypothesis, and why the proposal is new (Koronis et al., 2021; Nomaguchi et al., 2019). The absence of exhaustive methodological details forces reviewers to focus on conceptual and theoretical novelty. Thus, they can understand whether the proposal pushes the field's boundaries and is worth pursuing further.

- Early filter for originality: The EF scheme aims to seed disruptive or unorthodox projects that might not always fit within traditional funding calls. By spotlighting novelty in the first stage, the scheme ensures that highly innovative yet underdeveloped proposals are not prematurely rejected simply because their details are still taking shape. The short stage 1 application format encourages creative thinkers, whose primary strength might be a bold vision, to apply without being caught up by voluminous instructions or a requirement for fully-formed methodologies.

- Reduced cognitive load for reviewers: Reading a three-page synopsis is far less time-intensive than reading a detailed 30-page proposal. Reviewers can more easily compare multiple synopses, identify the key ideas, and grasp the potential impact of each. This approach can be particularly valuable for busy evaluators, often leading to clearer feedback and more decisive endorsements – or rejections – based on novelty.

Following this first novelty filter, the second stage necessarily foregrounds scientific rigour. Once a proposal has passed an initial threshold for originality, it must be presented in greater detail: applicants provide a full proposal elaborating on data sources, methods, pilot results, team composition, feasibility, risk management, and ethical considerations. Thus, there is a natural shift in reviewer priorities. Our analysis shows a marked increase in attention paid to scientific quality – both in terms of word count devoted to that criterion and the overall sentiment expressed by reviewers. This shift is logical and follows the progressive structure of the most thorough academic and funding processes: first, identify which ideas are truly worth developing; second, ensure that those promising ideas can be executed successfully.

At times, novelty may come at the cost of methodological weaknesses or incomplete feasibility planning. Conversely, high scientific rigour may inadvertently deprioritise groundbreaking concepts, favouring familiar or incremental approaches. The two-stage approach helps strike a balance between these extremes. It provides a structured means to:

- Allow original thinkers to enter the funding pipeline: By focusing on originality first, this process mitigates the risk that truly novel ideas will be lost because their methods are still evolving.
- Evaluate scientific quality at an appropriate time: By deferring an extensive methodology assessment to the second stage, the approach prevents early-stage proposals from being dismissed purely on technical grounds, which might be refined later.
- Promote transparency and accountability: Clear demarcation of stage 1 vs. stage 2 criteria allows applicants to understand precisely what is being evaluated and when. This fosters more targeted writing at each step, improving the coherence and efficiency of the review process.

These strengths underscore that a two-stage system can help funders manage large cohorts of applications while ensuring that novelty receives ample attention. The EF case study illustrates the potential for such systems to be replicated in other funding contexts where multi-dimensional criteria (for instance, interdisciplinarity, societal relevance, or potential for transformative impact) must be weighed alongside more traditional scientific standards.

Our analytical approach focused on the formal dimensions the programme formally prioritised because they had more apparent reviewer consensus and variation to measure with our tools. While recognising that concepts like

inter- and transdisciplinarity were part of the broader evaluative landscape, we did not directly operationalise them in the analysis. Future work could explore more targeted methods to surface how inter- and transdisciplinary framings influence peer assessment and how funders might more clearly guide reviewers in how to weigh such factors.

Relatedly, future work could also explore how the proposals' content and framing shape reviewers' assessments. Analysing both proposals and their corresponding reviews could help disentangle whether an emphasis on novelty or scientific quality reflects the reviewers' preferences or specific signals in the application texts.

A secondary but equally crucial aspect of our work involves integrating generative AI tools into the funding evaluation process. Our study demonstrates that these advanced natural language processing systems can serve as powerful complements to human peer reviewers, with multiple potential benefits:

- Systematic analysis of reviewer reports: AI can parse large volumes of text across multiple documents, detecting recurrent themes, sentiments, and prioritisations. By automating part of this analysis, AI reduces the manual workload, ensuring that commonalities and outliers are consistently surfaced and providing relevant insights at the portfolio level to funders.

- Triangulation with other methods: The generative AI approach can be combined with surveys, interviews, or panel discussions to confirm – and sometimes challenge – emerging patterns. This triangulation is valuable for verifying whether AI-generated insights align with feedback from evaluators or applicants, thereby enhancing the reliability of conclusions.

- Detection of implicit biases: Because AI systematically categorises text, it can more systematically highlight biases or inconsistencies. For instance, an AI might identify that certain criteria – such as team qualifications – are underdiscussed or overshadowed by excessive focus on feasibility. This insight allows funding organisations to refine guidelines or training for reviewers.

- Speed and scalability: As the volume of applications grows, reviewer committees struggle to maintain consistency, timeliness, and thoroughness in their evaluations. Conversely, AI can process

thousands of documents in parallel, providing preliminary analyses that human reviewers can interpret and refine. This scale-up potential means that large or complex calls can be handled more efficiently without sacrificing thoroughness.

While our results highlight the promise of generative AI, there are important cautionary notes and considerations that funding agencies and evaluators must consider. One major concern is hallucinations, as large language models can sometimes produce information that is not grounded in the input data. To address this, we required the AI system to provide direct quotes from reviewers' text wherever possible, explicitly linking claims to textual evidence. However, only requesting quotes is insufficient because, in a worst-case scenario, the requested quotes could be hallucinations. Additional safeguards such as human cross-checking and carefully crafted prompts remain essential to prevent misleading or fabricated outputs. We manually verified a random sample of 20% of the answers to ensure the quotes were not hallucinations and to sense-check the results.

Another critical issue is biases inherent in model training. AI models are trained on vast amounts of publicly available text, which can inadvertently encode societal and cultural biases. While prompt engineering and domain-specific fine-tuning can partially mitigate these effects, users of AI-based tools must remain vigilant about potential skew or unfairness, particularly in sensitive contexts such as equity, diversity, and inclusion.

Overreliance on automation also poses risks, as AI should not entirely replace human judgment in the evaluation process. Peer review is inherently rooted in disciplinary expertise, contextual understanding, and intangible aspects of academic rigour that are challenging to codify in algorithms. The optimal use of AI lies in its role as an assistant, helping to identify patterns or areas that warrant deeper scrutiny while leaving final decisions to human experts. Moreover, data confidentiality is paramount when dealing with sensitive proposals and confidential reviewer statements. Ensuring the security and privacy of this data is critical, and the approach we used – processing reviews through dedicated API connections and securing stored data strictly in EU jurisdictions – should serve as a baseline best practice for funding organisations. Addressing these considerations will allow AI's responsible and effective integration into evaluation processes while maintaining trust, fairness, and accountability.

Taken together, the two primary lessons from this study are deeply intertwined. First, a multi-stage review process helps segment the evaluation of

novelty from subsequent methodological and conceptual depth assessments, thereby allowing truly fresh ideas to surface and preventing premature dismissal. Second, generative AI has clear potential to streamline and enrich the evaluation of those multi-stage processes, furnishing an objective, data-driven lens that can detect patterns in reviewer feedback, highlight areas of discrepancy or consensus, and surface deeper insights into the strengths and weaknesses of each proposal.

Future developments in large language models, including improved capacity to handle domain-specific technical jargon and better safeguards against hallucinations, will likely expand their utility in research funding evaluations. Meanwhile, the ongoing conversation around how best to balance novelty and feasibility—exemplified by a two-stage approach—could continue to evolve. As funding bodies experiment with new ways of fostering ground-breaking science (including open peer review, co-creation processes, and user-driven research agendas), the careful integration of AI tools offers a viable path to ensuring thorough, data-enriched, and balanced assessments.

Ultimately, what this study underscores is the value of intentional design in the funding process. Different evaluation criteria (such as novelty, feasibility, and team capacity) require distinct review frameworks, and by supplementing peer review with structured AI insights, funding agencies can better identify, nurture, and support transformative research proposals that truly push the boundaries of knowledge. The two-stage EF mechanism is one such manifestation of this design philosophy, demonstrating that when novelty is given pride of place early on, many more original and pioneering ideas have the opportunity to flourish, subsequently facing the rigorous scientific scrutiny they require at the full-proposal stage. In tandem, generative AI ensures a level of systematic analysis that is both scalable and transparent, paving the way for an iterative cycle of continuous improvement in research funding evaluation.

BIBLIOGRAPHY

Boudreau, K., Guinan, E., Lakhani, K., & Riedl, C. (2012). The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations. https://doi.org/10.2139/ssrn.2184791.

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2017). Evaluating Novelty: The Role of Panels in the Selection of R&D Projects. IRPN: Other Development of Innovation (Topic). https://doi.org/10.2139/ssrn.3650896.

Fleurence, R., Bian, J., Wang, X., Xu, H., Dawoud, D., Higashi, M., & Chhatwal, J. (2024). Generative AI for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations. ArXiv, abs/2407.11054. https://doi.org/10.48550/arXiv.2407.11054.

Kolarz, P., Vingre, A., Machado, D., Sutinen, L., Dudenbostel, T., & Arnold, E. (2024). Accompanying process evaluation of FWF's Emerging Fields. Zenodo. https://doi.org/10.5281/zenodo.13911479.

Kolarz, P., Vingre, A., Vinnik, A., Neto, A., Vergara, C., Obando Rodrigues, C., Nielsen, K. & Sutinen, L. (2023). Review of peer review. UKRI, UK. https://www.ukri.org/publications/review-of-peer-review/.

Koronis, G., Casakin, H., & Silva, A. (2021). Crafting briefs to stimulate creativity in the design studio. Thinking Skills and Creativity, 40, 100810. https://doi.org/10.1016/J.TSC.2021.100810.

Langfeldt, L. (2006). The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. Research evaluation, 15(1), 31–41.

Laudel, G. & Gläser, J. (2014). 'Beyond breakthrough research: Epistemic properties of research and their consequences for research funding.' Research Policy, 43(7): 1204–16.

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. Research Evaluation, 21(1), 48–60

Morgan, B., Yu, L., Solomon, T., & Ziebland, S. (2020). Assessing health research grant applications: A retrospective comparative review of a one-stage versus a two-stage application assessment process. PLoS ONE, 15. https://doi.org/10.1371/journal.pone.0230118.

Nomaguchi, Y., Kawahara, T., Shoda, K., & Fujita, K. (2019). Assessing Concept Novelty Potential with Lexical and Distributional Word Similarity for Innovative Design. Proceedings of the Design Society: International Conference on Engineering Design. https://doi.org/10.1017/DSI.2019.147.

OECD (2021). „Effective policies to foster high-risk/high-reward research", OECD Science, Technology and Industry Policy Papers, No. 112, OECD Publishing, Paris, https://doi.org/10.1787/06913b3b-en.

Su, X., Wambsganss, T., Rietsche, R., Neshaei, S., & Käser, T. (2023). Reviewriter: AI-Generated Instructions For Peer Review Writing, 57–71. https://doi.org/10.18653/v1/2023.bea-1.5.

Wong, R. (2023). Role of generative artificial intelligence in publishing. What is acceptable, what is not … The Journal of ExtraCorporeal Technology, 55, 103–104. https://doi.org/10.1051/ject/2023033.

AUTHORS

PETER KOLARZ
Research on Research Institute
Email: p.kolarz@researchonresearch.org
ORCID: 0000-0003-4443-580X

DIOGO MACHADO
Technopolis
Email: diogo.machado@technopolis-group.com
ORCID: 0000-0001-6147-876X