On the Rigour of Scientific Writing: Criteria, Analysis, and Insights

Joseph James[1*], **Chenghao Xiao**[2*], **Yucheng Li**[3], **Chenghua Lin**[1, 4†]
[1]Department of Computer Science, The University of Sheffield, UK
[2]Department of Computer Science, Durham University, UK
[3]Department of Computer Science, University of Surrey, UK
[4]Department of Computer Science, The University of Manchester, UK
jhfjames1@sheffield.ac.uk chenghao.xiao@durham.ac.uk
yucheng.li@surrey.ac.uk chenghua.lin@manchester.ac.uk

Abstract

Rigour is crucial for scientific research as it ensures the reproducibility and validity of results and findings. Despite its importance, little work exists on modelling rigour computationally, and there is a lack of analysis on whether these criteria can effectively signal or measure the rigour of scientific papers in practice. In this paper, we introduce a bottom-up, data-driven framework to automatically identify and define rigour criteria and assess their relevance in scientific writing. Our framework includes rigour keyword extraction, detailed rigour definition generation, and salient criteria identification. Furthermore, our framework is domain-agnostic and can be tailored to the evaluation of scientific rigour for different areas, accommodating the distinct salient criteria across fields. We conducted comprehensive experiments based on datasets collected from two high impact venues for Machine Learning and NLP (i.e., ICLR and ACL) to demonstrate the effectiveness of our framework in modelling rigour. In addition, we analyse linguistic patterns of rigour, revealing that framing certainty is crucial for enhancing the perception of scientific rigour, while suggestion certainty and probability uncertainty diminish it.

1 Introduction

Rigour is one of the cornerstones of scientific research. Despite its profound importance and the widespread use of the term in both scientific and lay parlance, the scientific literature adds surprisingly little to our understanding of rigour, with the term almost always used without a definition, as if its meaning is self-evident. There are few definitions for scientific rigour available. For instance, the National Institutes of Health (NIH) has defined scientific rigour as "*the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results. This includes full transparency in reporting experimental details so that others may reproduce and extend the findings.*" (NIH, 2015). Whilst this definition may seem useful, it has been criticised for being both overly verbose and disconcertingly vague (Casadevall and Fang, 2016).

Inherently, scientific rigour is multi-faceted—no single criterion can define it fully. Indeed, there are some obvious dimensions of rigour, such as reproducibility, as rigorous scientific practice can enhance the likelihood that the results generated will be reproducible. In Computer Science (CS), recent years have witnessed an exponential increase in the number of publications (e.g., for AI, Machine Learning, and NLP), culminating in nearly half a million publications worldwide in 2021 alone (Maslej et al., 2023). This surge has resulted in what is termed *scientific debt*, where researchers prioritise 'novel' methods without sufficiently grounding their work in theory, conducting extensive ablation studies, or performing comprehensive evaluations (Nityasya et al., 2023). Furthermore, this trend has exacerbated the *reproducibility crisis* in science (Baker, 2016), a widespread problem where many scientific studies are difficult or impossible to reproduce by other researchers.

The community has begun to address some of the issues surrounding scientific rigour. For instance, in response to calls for more transparent and robust research, ACL introduced separate scores for soundness and excitement in 2023[1], and ICLR now includes a breakdown for correctness, technical and empirical novelty, and significance in its review process[2]. While these initiatives are important, there is a significant gap: *different research domains may*

*Equal contribution.
†Corresponding author.

[1]https://2023.aclweb.org/blog/overall-recommendation/
[2]https://iclr.cc/Conferences/2022/ReviewerGuide

Findings of the Association for Computational Linguistics: EACL 2024, pages 6523–6538

have varying preferences and traditions for defining rigour criteria, and there is a lack of analysis on whether those criteria can effectively signal or measure the rigour of scientific papers in practice. Efforts have been made to model different aspects of scientific writing. Notably, there is a rich literature on modelling scientific discourse, which aims not merely to present information and ideas but also to ensure their effective communication, allowing readers to accurately perceive what authors intend (Goldsack et al., 2023). Additionally, there are studies focusing on detecting scientific novelty from text (Savov, 2021; Luo et al., 2022), as novelty is regarded as an important aspect of judging scientific merit. However, compared to novelty, rigour is more challenging to define and analyse, and little work exists on modelling and analysing rigour computationally, particularly in the domain of CS. The limited existing work often focuses on non-CS domains (e.g., biomedical research), employs a top-down approach to defining criteria for scientific rigour (e.g., based on researchers' own experiences), or lacks empirical analysis on the effectiveness of the defined rigour dimensions (Prager et al., 2019).

To address the aforementioned gap in modelling scientific rigour, we have developed a *bottom-up*, *data-driven* framework that can automatically elicit candidate rigour criteria and construct their detailed definitions. Furthermore, our framework is domain-agnostic and can support the analysis of the salience of criteria for signalling the rigour of scientific papers–an important analysis which, to the best of our knowledge, has not been attempted previously. Specifically, we (i) construct a high-quality corpus of publications that exemplify high rigour, where the corpus is used to train a binary rigour classifier by fine-tuning SciBERT (Beltagy et al., 2019). We then extract candidate rigour keywords from two much larger datasets, namely ICLR[3] and ACL Anthology[4], based on the predictions of the rigour classifier using feature selection. Next, (ii) we generate detailed definitions for the candidate rigour keywords by prompting GPT-4, and (iii) analyse the salience of criteria for signalling the rigour of scientific papers by proposing an LLM-based embedding approach.

We conducted comprehensive experiments based on datasets collected from two high impact conferences for ML and NLP (i.e. ICLR and ACL) to demonstrate the effectiveness of our framework in modelling rigour, involving analysing the performance of rigour classifier, and assessing the most salient rigour criteria for each experimented dataset quantitatively. We hypothesise that linguistic pattern differences exist between high rigour and less rigour papers, impacting readers' perception of scientific rigour. Therefore, we further conducted a sentence-level analysis based on the aspect-level uncertainty theory of (Pei and Jurgens, 2021). Experimental results reveal that framing certainty is crucial for enhancing the perception of scientific rigour, while suggestion certainty and probability uncertainty diminish it. To summarise, our contributions are three-fold:

- We propose a bottom-up, domain-agnostic computational framework that automatically identifies candidate rigour criteria and generates their corresponding definitions.
- We create a set of Rigour Criteria and propose an LLM-embedding based method that can effectively measure the silence of specific rigour criteria for a given research domain.
- Our comprehensive analysis provides valuable insights into the linguistic features that signify perceived rigour in scientific writing, promoting transparency and robustness in scientific research.

2 Related Work

2.1 Criterion of Scientific Rigour

Despite its fundamental importance, existing guidelines or definitions for rigour are often vague and general, such as the NIH's suggestion to justify the methodology, identify potential weaknesses, and address limitations (Johnson et al., 2020; Wilson and Botham, 2021). Sansbury et al. (2022) highlight the importance of rigour in study design and conduct, statistical procedures, data preparation, and availability. In addition, there exist many domain-specific requirements for rigour proposed by researchers. For example, Lithgow et al. (2017) believe stricter variability control is necessary for animal research, following strict handling protocols, and adhering to precise methodological guidelines. Hamberg et al. (1994) discuss the multifaceted nature of rigour and suggest that different criteria should be used to assess truthful findings under different circumstances. In family medicine research, Hamberg et al. (1994) argue

[3] https://github.com/berenslab/iclr-dataset

[4] https://github.com/shauryr/ACL-anthology-corpus

that researchers should focus on credibility, dependability, confirmability, and transferability instead of traditional rigour criteria. In the computer science community, leading conferences like NeurIPS and ACL have employed a checklist approach for authors to self-review their submissions and address issues of research ethics and reproducibility (NeurIPS, 2021; ARR, 2023). This encourages authors to clearly describe their research questions, explicitly explain the limitations of their work, and report experiments in as much detail as possible.

These aforementioned criteria are predominantly developed in a top-down fashion, relying heavily on domain experts' experience. Such an approach has several limitations in real practice. Firstly, due to the multi-faceted nature of rigour, it may be challenging to directly apply existing rigour criteria or checklists from other domains. This makes it difficult to scale and adapt rigour assessment practices across various disciplines. Secondly, the criterion/checklist approach has a limited impact on the actual reviewing process. For instance, authors are only required to complete a checklist which can challenge reviewers trying to assess the level of rigour as additional details are often not required to support the claims.

Moreover, Randall (2023) suggests that even professional reviewers might excessively focus on novelty while neglecting the importance of rigour. These highlight the potential and need for bottom-up computational modelling of rigour, where such models can assist authors in improving their narrative and writing, and help editors assess the rigour and truthfulness of paper submissions.

2.2 Computational Modelling of Scientific Rigour

Several attempts have been made to computationally analyse the rigour of scientific papers. For example, Soliman and Siponen (2022) investigated how researchers use the word "rigour" in information system literature but discovered that the exact meaning was ambiguous in current research. Additionally, various automated tools have been proposed to assess the rigour of academic papers. Phillips (2017) developed an online software that spots genetic errors in cancer papers, while Sun et al. (2022) used knowledge graphs to assess the credibility of papers based on meta-data such as publication venue, affiliation, and citations. However, these methods are neither domain-specific nor do they provide sufficient guidance for authors to improve their narrative and writing. In contrast, SciScore (SciScore, 2024) is an online system that uses language models to produce rigour reports for paper drafts, helping authors identify weaknesses in their presentation. However, they rely on existing rigour checklists suggested by NIH and MDAR (Chambers et al., 2019), which are not easily scalable or transferable to other domains.

2.3 Research Excellence Framework (REF)

REF is the UK's national system for assessing the quality of research across UK universities. The assessment is carried out once every 7 years, and the assessment outcome informs the distribution of research funding nationwide, which constitutes a significant portion of each university's research income. Due to its importance, all universities identify the strongest outputs for REF submission (e.g., if a researcher has multiple NeurIPS/ACL publications, only the strongest among them will be submitted, as there is a stringent and limited cap on the number of submissions allowed per individual). The submitted output will then undergo reviews by panel members consisting of senior academics, who will rate the paper using a five-point scale from 4*-*Quality that is world-leading in terms of originality, significance and rigour*, to 0*-*quality that falls below the standard of nationally recognised work*. In addition, 4* papers are given much heavier weights than papers from other categories, e.g., one 4* paper will be allocated *four times* the funding of one 3* paper. See Appendix A.1 for a full description of the categories.

3 Methodology

In this section, we describe our data-driven, bottom-up framework for eliciting the criterion for defining the rigour of scientific writing, which consists of three main components. An illustration of the framework is shown in Fig 1.

3.1 Rigour Keyword Extraction

Extracting candidate keywords that are highly relevant to the rigour of scientific papers requires a corpus of papers that exemplify high rigour. While it may seem reasonable to use the review scores of papers as an indicator of rigour (e.g., papers from ICLR with higher correctness scores are regarded as more rigorous compared to those with lower scores), we argue that this might not be a robust proxy. Prior research has highlighted inconsistencies in the review process, noting that the

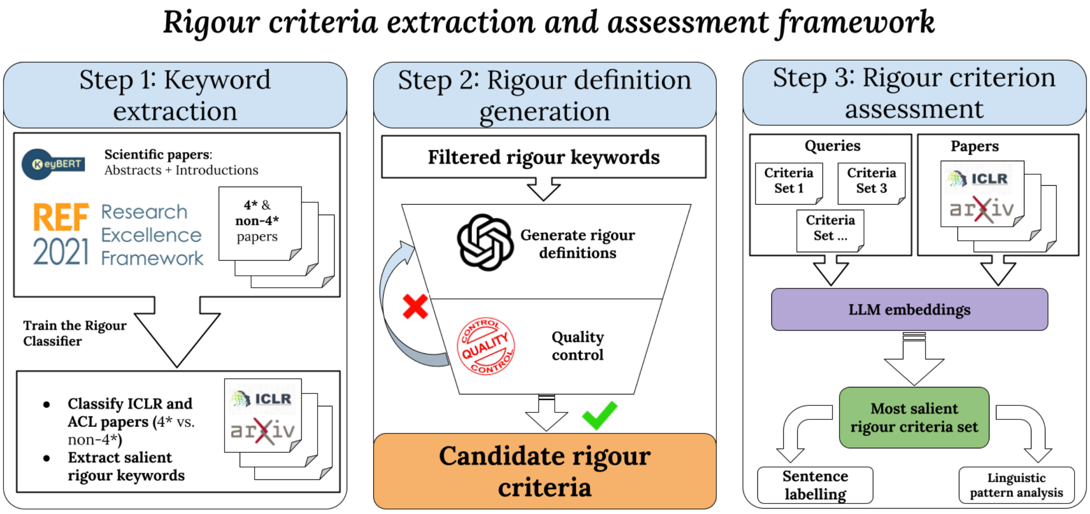

Figure 1: Illustration of the rigour criteria extraction and assessment framework.

acceptance decisions for approximately half of the papers would change if the process were repeated (Cortes and Lawrence, 2021; Beygelzimer et al., 2023).

High rigour paper source. To address this issue, we opted for a more reliable source—papers classified under the Research Excellence Framework (REF). The rationale, as discussed in §2, is that the stringent nomination and review processes (i.e., by expert panel members) provide very high confidence that the 4* publications classified in the REF will exhibit high rigour, which is one of the main assessment criteria. Since REF only publishes scores of publication outputs at the institutional level rather than for each individual publication, we collected publication outputs from institutions predominantly rated as 4* (e.g., Imperial College London and Oxford), as well as from institutions whose outputs are predominantly rated as non-4*. With this, we constructed a REF dataset consisting of papers categorised into two groups: 4* and non-4* (please refer to Table 1 for statistics).

Bias Mitigation. We performed preprocessing to minimise various types of biases that could affect the classification of rigour. First, we manually removed mandated keywords, footers, headers, and section titles from the papers and retained only the abstracts and introduction sections, instead of the entire articles. While other sections are also likely to reflect the quality or rigour of scientific writing, we focus on abstract and introduction due to several factors: (i) a prior study by Afzal et al. (2020) performed rigour classification based on the title and abstract in the Biomedical domain, achieving satisfactory performance; (ii) the REF dataset contains papers from various disciplines within Computer Science, not just NLP and machine learning. As a result, we observe a lack of standardisation across journals and conferences in other sections, e.g., large variations in paper length and the number of sections, as well as the limitations section not always being present. Therefore, we decided to focus on the abstract and introduction, which provide a comprehensive overview of the paper (e.g., contextual foundation, motivation, research problem/objectives, methods and results/findings) and could mitigate potential biases (e.g., paper length and number of sections) in training due to the relatively small dataset size.

Additionally, we removed information related to the authors, publication venues, and affiliated institutions to mitigate author and institution biases (Thelwall, 2022). To mitigate topical biases (e.g., theoretical machine learning, HCI, medical, etc.), we utilise the method proposed by Golchin et al. (2023) and extracted domain-specific keywords using KeyBERT (Grootendorst, 2020) then replaced the keywords with [MASK]. Through this meticulous process, we curated a high-quality dataset containing the abstracts and introductions of 988 papers, which were used to train and test our rigour classifier.

Keyword Extraction. We train a SciBERT (Beltagy et al., 2019) binary classifier for rigour on the constructed REF dataset. As our framework is designed to be domain-agnostic, we extract rigour keywords from two much larger datasets, namely ICLR and ACL (see Table 1 for statistics), which

	4*	non-4*	Train	Val	Test	Total
REF	292	696	790	99	99	988
ICLR	-	-	-	-	-	5,493
ACL	-	-	-	-	-	32,651

Table 1: Dataset statistics for all three datasets

represent two subfields of computer science; machine learning and NLP. This is achieved by first predicting the labels (4* or non-4*) of the papers from the two datasets, followed by feature extraction using mutual information (Kraskov et al., 2004) to separate out the important features for 4* and non-4*. Positive coefficients indicate a higher association with 4* papers, while negative coefficients indicate a higher association with non-4* papers. The keywords prominent in 4* papers are shown in Fig 2 and for non-4* papers in Appendix A.3. The extracted candidate rigour keywords then underwent manual filtering for quality control. We provide the full list of the top 100 keywords in Appendix A.4.

3.2 Rigour Definition Generation

Following the extraction of rigour keywords in §3.1, we generated definitions using GPT-4[5] (Achiam et al., 2023), using the following prompt: "*Give the definition of "[keyword]" in the context of Computer science and Machine learning. In the format: [keyword]: Refers to [definition]*".

We validate our generated criteria manually to determine whether our outputs are adequate at explaining each rigour keyword (more details given in §5.2). For example, the definition of **reproducibility** by Raghupathi et al. (2022) is: "*the ability of an independent research team to produce the same results using the same research method based on the documentation made by the original research team*", in comparison to our GPT-4[6] generated definition: "*Refers to the ability to reliably recreate the same results or outputs from a given model or experiment, given the same input data and configuration settings, by providing the complete source code and using openly available tools and datasets*". Overall, our manual examinations verify that the GPT-4 generated definitions are of good quality. A full list of generated definitions are shown in Appendix A.5.

[5]Mixtral-8x7B, Gemini-1.5, and Claude-3 Opus were also tested and resulted in similar outputs.

[6]gpt-4-turbo-2024-04-09

3.3 Salient Rigour Criterion Assessment

To the best of our knowledge, no prior work has attempted to analyse the salience of criteria for signalling rigour. We approach the problem by assessing the prominence of a rigour criterion or a set of rigour criteria (e.g., **{Settings, Baselines}**, **{Examples, Benchmarks, Justifications}**) based on their semantic similarity to 4* and non-4* papers. By doing so, we measure whether papers containing a specific set of rigour criteria associate more with 4* papers than non-4* papers.

LLM Embeddings. We leverage LLM-based embedding models to calculate the semantic similarity between each rigour criterion definition and papers from the REF, ICLR and ACL datasets. The rationale behind opting for off-the-shelf LLM-based embedding models are three-fold: (i) the lack of fine-grained ground-truth labels for each criterion, thus unable to train a criterion-specific classifiers; (ii) given the limited size of the REF dataset, training a domain-specific embedding model with contrastive learning essentially enforces a uniform embedding space for this specific task, resulting in subpar performance; and (iii) recent advancements in LLM-based embedding models show exceptional generalisation, alignment on reasoning-level language, and instruction-following abilities (Wang et al., 2023; Muennighoff et al., 2024; Xiao et al., 2024). Therefore, we argue that it is reasonable to judge rigour by encoding rigour criteria and scientific papers into the same semantic space, and conducting similarity matching to identify the prevalence of the rigour criteria for each paper. In our experiment, we specifically use the representations of GritLM (Muennighoff et al., 2024), a 7B model that unifies generative and representational abilities in one model and achieves state-of-the-art results on MTEB (Massive Text Embedding Benchmark (Muennighoff et al., 2022)) and RAR-b (Reasoning as Retrieval Benchmark (Xiao et al., 2024)).

Computing semantic similarity. We prepend the query (combinations of the criteria with its associated definition) with the following instruction: "*Given the following definitions, retrieve the appropriate document that contains the following criteria:*". The concatenated instruction and query are passed to the model, and we apply mean pooling to the query tokens, giving us $E(q|i)$ (i.e. query embeddings conditioned on the instruction). Aligning with common practices in instruction-aware embedding systems (Asai et al., 2023; Su et al., 2023), we encode the documents without instructions, giving the document embedding $E(d)$. The cosine similarity between $E(q|i)$ and $E(d)$ is taken as the indicator of a document reaching the corresponding criterion q, allowing our analysis on the corpus-level distributional difference. Formally, we have

$$\mathbf{E}(q|i) = \frac{1}{|q|}\sum_{t\in q}\mathbf{E}([i;q])_t; \mathbf{E}(d) = \frac{1}{|d|}\sum_{t\in d}\mathbf{E}(d)_t \tag{1}$$

where i denotes the instruction built upon the rigour criteria, $\mathbf{E}([i;q])_t$ denotes the embedding of the t-th token in the concatenated sequence $[i;q]$, and $|q|$ is the number of tokens in the query q; $\mathbf{E}(d)_t$ denotes the embedding of the t-th token in the document d, and $|d|$ is the number of tokens in document d without prepending instructions. And $\text{cos\_sim}(\mathbf{E}(q|i), \mathbf{E}(d)) = \frac{\mathbf{E}(q|i)\cdot\mathbf{E}(d)}{\|\mathbf{E}(q|i)\|\|\mathbf{E}(d)\|}$ is taken as the indicator of document d meeting rigour criteria i.

Finally, we identify the most salient set of rigour criteria by analysing whether a significant distributional difference exists by comparing the semantic similarities between the queried set of criteria and the 4* and non-4* papers. Notably, we discovered that appending sets of criteria to a single query is more effective at separating the semantic similarities in comparison to individual criterion queried then summed, as shown in Appendix A.8.

4 Experimental Setup

Datasets. The REF dataset was constructed based on REF 2021[7], which covers papers published between 2014 and 2021. The submissions from UOA (Unit of Assessment) 11 [8], which covers all areas of computer science and information, was used to create the dataset. As described in §3.1, we collected publications from institutions whose outputs are predominantly rated as 4* (e.g., Imperial College London and Oxford), as well as from institutions whose outputs are predominantly rated as non-4*, to form a binary labelled dataset for rigour.

For both the ICLR and ACL papers, we extracted abstracts and introductions from existing datasets or from arXiv[9]. For ICLR, we considered all papers from 2022 and 2023 (González-Márquez and Kobak, 2024), to show the effectiveness of our rigour classifier on papers submitted after 2021. The ACL dataset was developed using the ACL Anthology Corpus (Rohatgi, 2022). Given the high formatting consistency within their respective domains, we automatically extracted abstracts and introductions from both datasets using pattern matching with section titles. The statistics of the three datasets are presented in Table 1.

Rigour classifier and feature extraction. We trained the rigour classifier by fine-tuning a Longformer version of SciBERT[10], a pre-trained language model trained on scientific publications that has proven effective at capturing language patterns and domain-specific knowledge within scientific texts (Beltagy et al., 2019). The full settings, including the model hyperparameters, are given in Appendix A.2. For feature extraction, we implemented mutual information with logistic regression. The default parameters provided by the scikit-learn library[11] were used.

GritLM-7B. We used the embedding-only variant of the GritLM-7B model (Muennighoff et al., 2024), and the representations without the LM head. Mean pooling is taken over token-level embeddings to attain sentence-level embeddings. For query embeddings, the final attention is only given to actual query tokens conditioned by the prepended instructions (Muennighoff et al., 2024), and for document embeddings, we simply take the mean pooling of all tokens.

5 Experimental Results

5.1 Rigour classification and feature extraction

Experimental results of rigour classification based on the REF dataset demonstrated that our classifier gives strong performance, signalling that the abstract and introduction of a paper can provide rich information for predicting the rigour of papers. More specifically, our classifier trained on our processed data achieved an accuracy of 0.94 and F1 score of 0.90, while the unmasked data (topic words included) resulted in an accuracy of 0.93 and F1 score of 0.88. The results highlight the robustness of the classifier in distinguishing 4* and non-4* papers.

We then predicted the rigour labels for the ICLR and ACL dataset with our classifier, and performed

[7] https://2021.ref.ac.uk/

[8] https://results2021.ref.ac.uk/profiles/units-of-assessment/11

[9] https://www.kaggle.com/datasets/Cornell-University/arxiv/data

[10] https://huggingface.co/yorko/scibert_scivocab_uncased_long_4096

[11] https://scikit-learn.org/stable/

feature selection based on mutual information to identify the most salient keywords. The top 30 salient keywords for 4* papers are shown in Fig 2, with keywords related to rigour highlighted. For examples *"Setting"*, *"Generalise"* and *"Robust"*.

5.2 Assessing the Salience of Rigour criteria

From the salient keywords, we manually selected 13 potential rigour criteria. An additional three criteria were chosen based off of prior work, these include *"Reproducibility"*, *"limitations"* and *"justifications"*, derived from literature discussed in §2. Following this, we generated 65,535 unique combinations of the criteria to be used to query REF, ICLR and ACL papers, based on the unique permutations of criteria C and subset of elements in the criteria set r, given by $N = \sum_{r \subseteq C} \frac{C!}{r!(C-r)!}$.

The most salient rigour criteria set, i.e., criteria with the highest correlation with 4* papers, are presented in Table 2, with the corresponding distributions of the cosine similarity between the rigour criteria and each document split by 4* and non-4*, shown in Fig 3. The cosine similarity score here is calculated via the documents and their corresponding salient rigour criteria set. The cosine similarity values indicate the prominence of the rigour criteria set which we utilise to signify what makes 4* papers different to non-4* papers. The correlation between rigour labels and the cosine similarity score are 0.307 for REF, 0.227 for ICLR and 0.240 for ACL ($p < 0.0001$).

A moderate correlation was observed for REF, while a weaker correlation was observed for ICLR and ACL. Nonetheless, we can observe a clear difference in distribution from Fig 3, indicating 4* papers exhibit higher similarity scores with our salient rigour criteria set. By using Kendall's correlations, we justify the statistical significance of our correlations (Gilpin, 1993).

The results in Table 2 suggest each dataset has a slightly different rigour criteria set preference, indicating inherent domain-specific characteristics. We observe that criteria *"Baselines"*, *"Benchmarks"*, *"Assumptions"* and *"Reproducibility"* are prominent across all three datasets. This falls inline with recent work emphasising the importance of reproducibility (Semmelrock et al., 2023). On the other hand, *"Challenges"* and *"Contributions"* show little difference between the 4* and non-4* papers. This suggests that all papers contain these criteria to a similar extent, indicating a consensus across all domains. In summary, we reveal that different sub-fields exhibit some degree of rigour criteria set preference, which can be captured by our framework in a data-driven way.

Criteria	REF	ICLR	ACL
Biases	–	✓	✓
Settings	–	✓	✓
Constraints	✓	✓	–
Limitations	–	–	✓
Baselines	✓	✓	✓
Benchmarks	✓	✓	✓
Empirical Findings	✓	✓	–
Examples	✓	–	–
Motivations	–	–	✓
Generalisation	✓	–	✓
Robustness	–	✓	✓
Assumptions	✓	✓	✓
Justifications	✓	✓	–
Challenges	–	–	–
Contributions	–	–	–
Reproducibility	✓	✓	✓

Table 2: The most salient rigour criteria sets for the REF, ICLR, and ACL datasets, where ✓ indicates that a particular criterion is included in the criteria set for a specific dataset, and – indicates its absence. Essentially, each column corresponds to a criteria set.

5.3 Linguistic Patterns of Scientific Rigour

We first separate the papers into sentences for both 4* and non-4* papers. Then each sentence is labelled via GritLM with a rigour criterion. We utilise a threshold of 0.5 for the cosine similarity between rigour criterion and sentences to remove sentences that are not similar enough to any of the criterion. This is to exclude sentences such as prior work, which are prevalent in the papers but irrelevant to our rigour criteria. In total, we obtain 400k sentence labels which associated with a rigour criteria, a full breakdown of labels are shown in the Appendix A.6. We utilise these sentences to extract linguistic patterns that highlight the differences between 4* and non-4* sentences.

Certainty metric. Certainty is a crucial concept in scientific writing when communicating knowledge to the reader, especially for conveying rigour (National Academies, 2017). For example *framing certainty* indicates how *certain* or *confident* scientific findings are "framed" and "interpreted". Here we evaluate certainty of different aspect for the rigour criteria we identified previously. Specifically, we use the certainty classifier introduced in (Pei and Jurgens, 2021) to label sentences with a certainty aspect. We do this for both 4* and non-4* papers, which allows us to find the most relevant aspect of certainty for each rigour criterion in 4* or non-4* sentences.

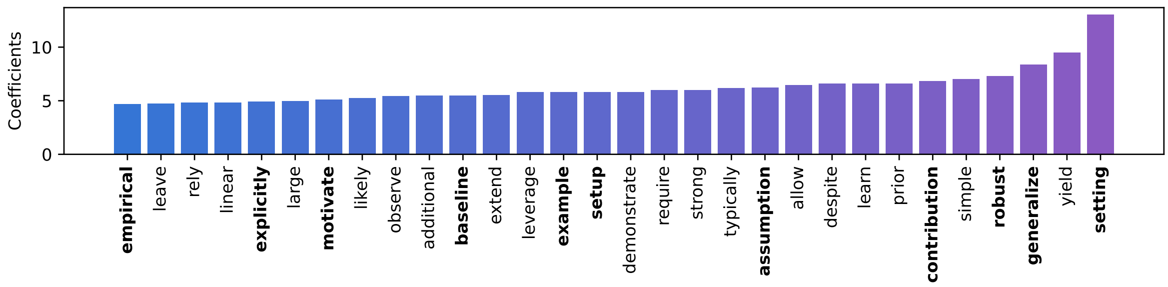

Figure 2: Top 30 salient keywords for 4* predicted papers from ICLR and ACL using Mutual Information with candidate rigour keywords highlighted in **bold**. List of keywords in Table A.4.

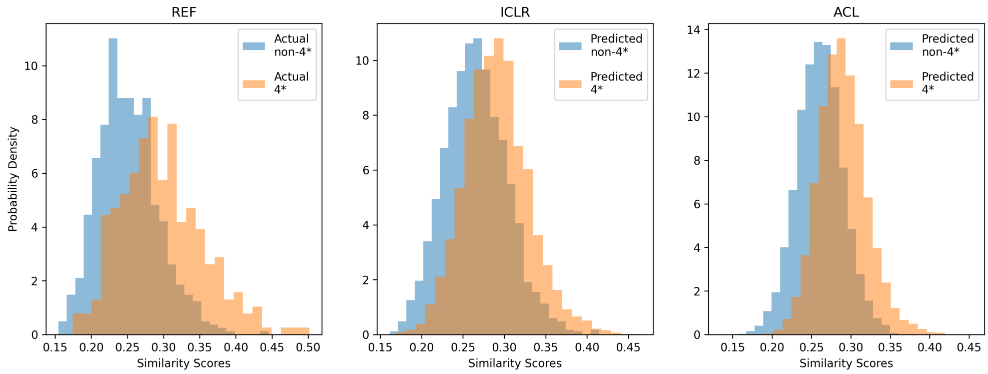

Figure 3: Distribution of similarity score for best rigour criteria set for each of REF, ICLR and ACL datasets

Sentences with rigour criteria contained in both ACL and ICLR (Biases, settings, Baselines, Benchmarks, Robustness, Assumptions and Reproducibility), 346k in total, were used in our evaluation and the most relevant aspects of certainty were identified by their differences in probabilities between 4* and non-4* sentences, as shown in Fig 4. It can be observed that 4* sentences are more likely to contain *framing certainty*. But for non-4* sentences, *suggestion certainty* and *probability uncertainty* are the more common phenomenons, where *suggestion certainty* refers to how certain the findings propose suggestions or future actions, and *probability uncertainty* indicates the usage of uncertain wording such as "possibly". These findings on linguistic patterns around certainty highlight the importance of scientific writing, as it has direct impact on readers' perception of the rigour of a paper. We provide examples in Appendix A.7 and further analysis on all sentences in Appendix A.9.

Human evaluation. We further conducted human evaluation to determine if the perceived level of rigour was different in 4* and non-4* sentences. We recruited two CS postgraduate students to evaluate their preference from pairs of sentences from rigour criterion contained in both ICLR and ACL.

Certainty	Example
Framing certainty	The most widespread family of techniques **are** diagnostic models, which use the internal activations of neural networks trained on a particular task as input to another predictive model.
Suggestion certainty	Such systems **need** to be assessed by system developers for any possible technological improvements and novel research ideas and by potential users for quality comparison purposes.
Probability uncertainty	Various features can *potentially* be used, based on the source and target context as well as syntactic and semantic analysis.

Table 3: A sample of sentences from the three prevelant certainty classes. Words highlight in **bold** indiciate the certainty and words in *italics* indicates the uncertainty.

We selected 35 examples which was calculated using power analysis, this allowed the estimation of the smallest samples required for human evaluation to be statistically significant (for Cohen's d of 0.57, statistical power of 0.80 and significance level of 0.05 (Schuff et al., 2023)). The evaluators were given guidelines to act as a reviewer assessing senrately.

As shown in Table 4, we observe that the inter-annotator agreement for 4* sentences is fair while for non-4* sentences there was no agreement, demonstrating the consensus on sentences from 4* papers and signifying a higher level of perceived rigour. This is further backed up by our certainty results in Fig 4, as specific certainty aspects are favoured in 4* papers such as *framing certainty*. The relatively low agreement scores is not surprising due to rigour being a highly complex and abstractive concept. However, the key take-away of the results is the higher agreement of more rigorous sentences (from 4* papers) compared to less rigorous sentences (from non-4* papers). This shows more rigour in a sentence leads to less ambiguity and thus more likely to be agreed upon by the evaluators. We provide examples of 4* and non-4* sentences with respect to the rigour criteria in Appendix A.7.

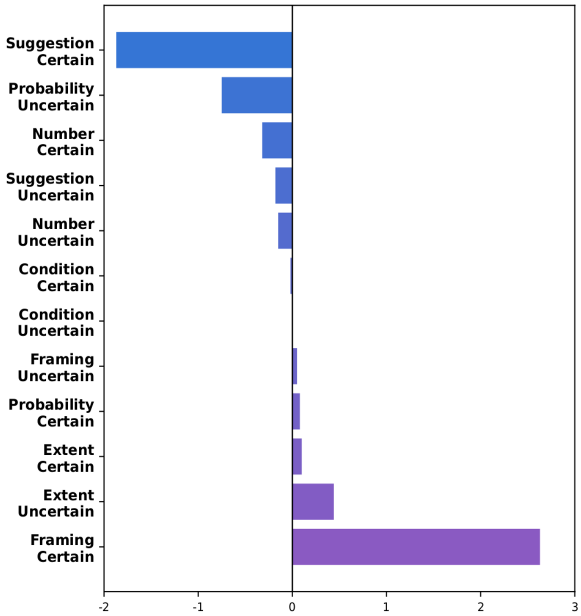

Figure 4: Uncertainty scores for rigour criteria present in both ICLR and ACL. Positive values indicate 4* preference while negative values indicate non-4* preference.

Label	Rigour rating	Cohen's kappa
4*	3.4143	**0.349**
non-4*	3.3714	0.030

Table 4: Human evaluation on 35 pairs of sentences from rigour criteria contained in both ICLR and ACL.

tence pairs in isolation along with the associated rigour criterion. The guideline for rigour includes: (i) is the presented sentence written confidently about the criterion (i.e., confidence and certainty); (ii) does the sentence contain enough information to interpret the criterion (i.e., level of detail and relevance); and (iii) is the sentence straight to the point without excessive filler words (i.e., conciseness).

To mitigate the potential confounding effect of sentence-level topic differences within each rigour criterion, we ensured that sentence pairs we annotated were semantically similar, verified with SciBERT using cosine similarity. This allowed us to assess the nuanced differences in the sentences, eliminating the potential impact of different topics and findings. Sentences were assessed in pairs to give evaluators a frame of reference when scoring the level of perceived rigour and presented in a random anonymised order. Evaluators were given a 5-point Likert scale to rate both sentences sepa-

6 Conclusion

In this paper, we introduce a bottom-up, data-driven framework to automatically identify and define rigour criteria and assess their relevance in scientific writing. Our framework is domain-agnostic and can be tailored to the evaluation of scientific rigour for different areas. Comprehensive experiments based on datasets demonstrate the effectiveness of our framework in modelling rigour. In addition, we analyse linguistic patterns of rigour, revealing that framing certainty is crucial for enhancing the perception of scientific rigour, while suggestion certainty and probability uncertainty diminish it.

Limitation

LLMs. The use of Large language models as a definition generator, semantic measure and sentence label annotator has its limitations, this is due to the stochastic nature of such models that may not capture all the nuances in the text compared to expert annotators. We utilised an embedding-based variant to partially address this limitation in the semantic measure, though it is not without its own drawbacks.

Domain specific. Our approach looks at Computer Science papers and more specifically Machine Learning. Extending to other domains would lead to a more generalised rigour criteria, however this may not be desirable due to differences across domains which would reduce the descriptiveness of the criteria.

Full paper classification. Given the variability in formatting and structure across publications, we focused the analysis on the abstracts and introductions of the papers. This approach allowed more consistent evaluation across topics. However, to develop a comprehensive understanding of the textual factors that contribute to rigour, it is crucial that future investigation should analyse other sections.

Acknowledgements

Joseph James is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Afzal, Beom Joo Park, Maqbool Hussain, and Sungyoung Lee. 2020. Deep learning based biomedical literature classification using criteria of scientific rigor. *Electronics*, 9(8):1253.

ARR. 2023. ACL Rolling Review - Responsible NLP Research Checklist. https://aclrollingreview.org/responsibleNLPresearch/. Accessed: 12 June 2024.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604).

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2023. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*.

Arturo Casadevall and Ferric C Fang. 2016. Rigorous science: a how-to guide.

Karen Chambers, Andy Collings, Chris Graf, Veronique Kiermer, David Thomas Mellor, Malcolm Robert Macleod, Sowmya Swaminathan, Deborah Sweet, et al. 2019. Towards minimum reporting standards for life scientists.

Corinna Cortes and Neil D Lawrence. 2021. Inconsistency in conference peer review: revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*.

Research England, the Scottish Funding Council, the Higher Education Funding Council for Wales, and Northern Ireland the Department for the Economy. 2021. Guidance on ref 2021 results - ref 2021. https://2021.ref.ac.uk/guidance-on-results/guidance-on-ref-2021-results/index.html#outputs. Accessed: 12 June 2024.

Andrew R Gilpin. 1993. Table for conversion of kendall's tau to spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement*, 53(1):87–92.

Shahriar Golchin, Mihai Surdeanu, Nazgol Tavabi, and Ata Kiapour. 2023. Do not mask randomly: Effective domain-adaptive pre-training by masking in-domain keywords. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 13–21, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2023. Domain-driven and discourse-guided scientific summarisation. In *European Conference on Information Retrieval*, pages 361–376. Springer.

Rita González-Márquez and Dmitry Kobak. 2024. Learning representations of learning representations. In *Data-centric Machine Learning Research (DMLR) workshop at ICLR 2024*.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Katarina Hamberg, Eva Johansson, Gerd Lindgren, and Göran Westman. 1994. Scientific rigour in qualitative research—examples from a study of women's health in family practice. *Family practice*, 11(2):176–181.

Jessica L Johnson, Donna Adkins, and Sheila Chauvin. 2020. A review of the quality indicators of rigor in qualitative research. *American journal of pharmaceutical education*, 84(1):7120.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E*, 69(6):066138.

Gordon J Lithgow, Monica Driscoll, and Patrick Phillips. 2017. A long journey to reproducible results. *Nature*, 548(7668):387–388.

Zhuoran Luo, Wei Lu, Jiangen He, and Yuqi Wang. 2022. Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics*, 16(2):101282.

Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, et al. 2023. Artificial intelligence index report 2023. *arXiv preprint arXiv:2310.03715*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

National Academies. 2017. *Communicating science effectively: A research agenda*. National Academies Press.

NeurIPS. 2021. NeurIPS 2021 Paper Checklist Guidelines. https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist. Accessed: 12 June 2024.

NIH. 2015. Enhancing reproducibility through rigor and transparency. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html. Accessed: 12 June 2024.

Made Nindyatama Nityasya, Haryo Wibowo, Alham Fikri Aji, Genta Winata, Radityo Eko Prasojo, Phil Blunsom, and Adhiguna Kuncoro. 2023. On "scientific debt" in NLP: A case for more rigour in language model pre-training research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8554–8572, Toronto, Canada. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicky Phillips. 2017. Online software spots genetic errors in cancer papers. *Nature*, 551(7681).

Eric M Prager, Karen E Chambers, Joshua L Plotkin, David L McArthur, Anita E Bandrowski, Nidhi Bansal, Maryann E Martone, Hadley C Bergstrom, Anton Bespalov, and Chris Graf. 2019. Improving transparency and scientific rigor in academic publishing.

Wullianallur Raghupathi, Viju Raghupathi, and Jie Ren. 2022. Reproducibility in computing research: An empirical study. *IEEE Access*, 10:29207–29223.

Ian Randall. 2023. Theorists are good peer reviewers - but tend to prefer significance over rigour, study finds. https://physicsworld.com/a/theorists-are-good-peer-reviewers-but-tend-to-prefer-significance-over-rigour-study-finds/. Accessed: 12 June 2024.

Shaurya Rohatgi. 2022. Acl anthology corpus with full text. Github.

Brian E Sansbury, Matthew A Nystoriak, Shizuka Uchida, Marcin Wysoczynski, and Joseph B Moore IV. 2022. Rigor me this: What are the basic criteria for a rigorous, transparent, and reproducible scientific study? *Frontiers in Cardiovascular Medicine*, 9:913612.

Pavel Savov. 2021. Measuring the novelty of scientific papers.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

SciScore. 2024. The best methods review tool for scientific research. https://sciscore.com/. Accessed: 12 June 2024.

Harald Semmelrock, Simone Kopeinik, Dieter Theiler, Tony Ross-Hellauer, and Dominik Kowald. 2023. Reproducibility in machine learning-driven research. *arXiv preprint arXiv:2307.10320*.

Wael Soliman and Mikko Siponen. 2022. What do we really mean by rigor in information systems research?

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Kexuan Sun, Zhiqiang Qiu, Abel Salinas, Yuzhong Huang, Dong-Ho Lee, Daniel Benjamin, Fred Morstatter, Xiang Ren, Kristina Lerman, and Jay Pujara. 2022. Assessing scientific research papers with knowledge graphs. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2467–2472.

Mike Thelwall. 2022. Can the quality of published academic journal articles be assessed with machine learning? *Quantitative Science Studies*, 3(1):208–226.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Jennifer L Wilson and Crystal M Botham. 2021. Three questions to address rigour and reproducibility concerns in your grant proposal. *Nature*, 596(7873):609–610.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*.

A Appendix

A.1 REF star rating description

Quality level	Description
Four star	Quality that is world-leading in terms of originality, significance and rigour.
Three star	Quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence.
Two star	Quality that is recognised internationally in terms of originality, significance and rigour
One star	Quality that is recognised nationally in terms of originality, significance and rigour.
Unclassified	Quality that falls below the standard of nationally recognised work. Or work which does not meet the published definition of research for the purposes of this assessment.

Table 5: REF star rating description (England et al., 2021)

A.2 Model training parameters

Model trained on two RTX4090's with hyperparameters shown in Table 6.

A.3 Key features for negative coefficients

Top 30 Salient keywords for non-4* papers are shown in Fig 5.

A.4 Top 100 Salient keywords

The top 100 salient keywords for 4* papers are shown in Table 7.

Setting	Value
num_train_epochs	5
per_device_train_batch_size	1
per_device_eval_batch_size	1
warmup_steps	100
weight_decay	0.01
learning_rate	5e-5
logging_steps	50
evaluation_strategy	'steps'
eval_steps	50
load_best_model_at_end	True
metric_for_best_model	'f1'
gradient_accumulation_steps	8
seed	41

Table 6: Model training hyperparameters

A.5 Rigour criteria defintitions

GPT-4 generated rigour criteria shown in Table 8.

A.6 Sentence labels

Criterion	Frequency
Settings	251,755
Benchmarks	40,402
Baselines	30,100
Generalisation	20,527
Reproducibility	13,185
Motivations	11,820
Biases	10,549
Assumptions	5,726
Robustness	5,490
Examples	3,513
Empirical Findings	2,994
Limitations	2,048
Justification	1,053
Constraints	990
Total	400,152

Table 9: Breakdown of sentence labels for REF, ACL and ICLR papers.

A.7 4* vs non-4* rigour examples

Examples comparing 4* and non-4* sentences, shown in Table 10.

A.8 Appending vs individual

Appending criteria together and querying documents resulted in better separation, as shown in Table 6.

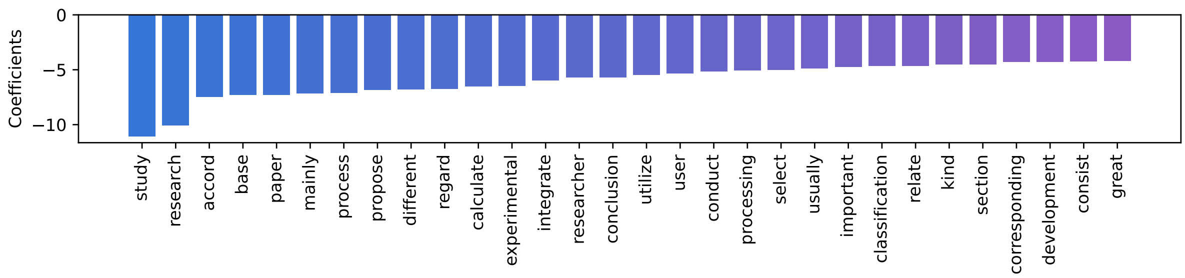

Figure 5: Top 30 salient keywords for non-4* predicted papers from ICLR and ACL using Mutual Information.

Chi-square features				
setting	yield	**generalize**	**robust**	simple
contribution	prior	learn	despite	allow
assumption	typically	strong	require	demonstrate
setup	**example**	leverage	extend	**baseline**
additional	observe	likely	**motivate**	large
explicitly	linear	rely	leave	**empirical**
hold	estimate	**assume**	outperform	remove
instead	encode	replace	train	inference
tune	derive	produce	recent	gradient
subset	particular	contrast	remain	fail
directly	binary	unlike	property	consistent
natural	true	recently	distribution	fix
want	note	count	**constraint**	variable
denote	low	objective	let	log
align	embedding	downstream	state	right
parse	neural	optimization	non	**bias**
choose	simply	noise	small	average
differ	**benchmark**	deep	condition	unsupervised
draw	suggest	single	effect	like
fully	representation	perform	standard	high

Table 7: 4* predicted rigour paper keywords extracted using Logistic regression and percentile Mutual information of 10%, where bold text implies use as rigour criteria.

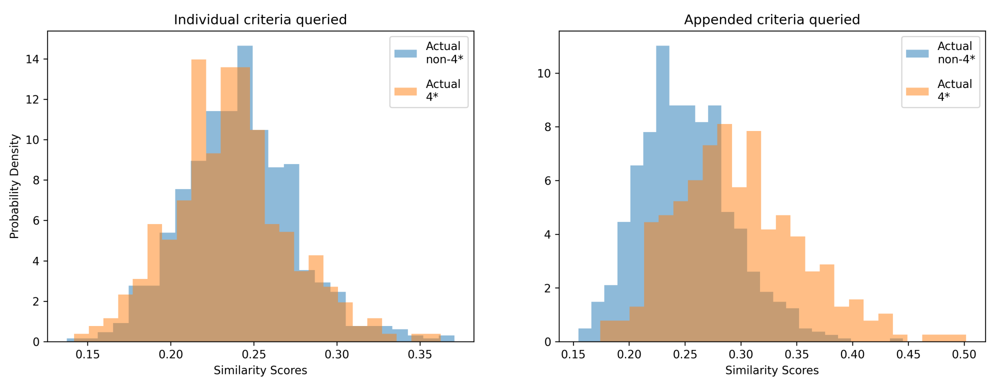

Figure 6: Comparing the difference when the criteria are appended together (Best combination on the right) vs individual criterion queries and then adding the similarity score (Best criteria added together on the left)

Criterion	Definition
Biases	Refers to systematic errors or distortions in the collection, analysis, or interpretation of data that can lead to skewed or inaccurate results. Uncertainties refer to the lack of precision or confidence in measurements, predictions, or conclusions due to limitations in data or knowledge.
Settings	Refers to adjustable parameters or configurations that define the behavior or performance of a system or model. They are typically set before the learning or optimization process and impact the final outcome.
Contributions	Refers to the original ideas, innovations, or advancements made by individuals or groups in the field. Contributions can take various forms and impact different aspects of computer science, including research, technology development, software engineering, algorithms, systems design, or theoretical advancements.
Constraints	Refers to restrictions or bottlenecks imposed on a system, software application, algorithm, or problem-solving process. Constraints define the boundaries within which a solution or system must operate, and they help guide the design, implementation, and behaviour of computer systems.
Limitations	Refers to the inherent shortcomings that exist within a system, technology, algorithm, or problem-solving approach. Limitations define the boundaries of what a system or solution can achieve or the constraints that restrict its performance, functionality, or applicability.
Generalisation	Refers to the process of extracting common patterns, concepts, or properties from specific instances or examples and formulating more abstract or generalised representations or models. Generalisations help capture the essential characteristics or behaviours shared by a set of objects, data, or systems, enabling more efficient and flexible problem-solving, analysis, or design.
Robustness	Refers to the ability of a system, software application, algorithm, or network to effectively handle and recover from abnormal or unexpected conditions, inputs, or events. A robust system is designed to withstand errors, exceptions, invalid inputs, or challenging operating conditions and continue functioning correctly or gracefully degrade without catastrophic failures.
Benchmarks	Refers to standardised tests, metrics, or reference points used to measure and evaluate the performance, efficiency, or capability of computer systems, software applications, algorithms, or hardware components. Benchmarks provide a basis for comparing different systems or solutions and assessing their relative strengths and weaknesses.
Baselines	Refers to reference points or initial measurements that serve as a starting point for comparison or evaluation. Baselines provide a foundation for assessing the performance, effectiveness, or efficiency of systems, algorithms, models, or solutions.
Assumptions	Refers to the statements or conditions that are considered to be true or valid for the purpose of designing, developing, or analysing systems, algorithms, models, or solutions. Assumptions simplify problem-solving processes by providing a set of predefined conditions or constraints under which a particular approach or solution is expected to work correctly.
Examples	Refers to specific instances or data points that are used to illustrate or demonstrate a concept, principle, or the behavior of an algorithm or model. These examples serve to showcase the application of a technique or highlight particular characteristics, allowing for a clearer understanding and communication of the ideas involved.
Empirical Findings	Refers to observations, data, or evidence obtained through direct observations, experiments, or measurements in the real world. They are based on empirical evidence rather than theoretical or speculative reasoning.
Justifications	Refers to present reasons, evidence, or logical justifications in support of a particular claim, position, or viewpoint. It involves making a persuasive case or engaging in a reasoned debate.
Challenges	Refers to difficulties, obstacles, or problems that need to be addressed or overcome in a particular context or task. They may arise due to technical, theoretical, practical, or ethical factors.
Reproducibility	Refers to the ability to reliably recreate the same results or outputs from a given model or experiment, given the same input data and configuration settings, by providing the complete source code and using openly available tools and datasets.
Motivations	Refers to the reasons, goals, or driving factors behind a particular study, project, or research endeavor. A gap refers to a missing or unaddressed aspect or area within existing knowledge or literature, which motivates further investigation or research.

Table 8: GPT-4 generated definitions for rigour criteria.

Criterion	4*	non-4*
Reproducibility	We validate our findings through numerical experiments where our theory accurately predicts empirical findings and remains consistent with observations in deep neural networks.	Our framework provides a reproducible and easy-to-use entry point for the development and evaluation of future bias mitigation algorithms in deep learning.
Benchmark	We design experiments to thoroughly test and objectively score metrics on their ability to measure the diversity and fidelity of generated graphs, as well as their sample and computational efficiency.	Ten years later, the ImageNet dataset is still one of the main benchmarks for state-of-the-art computer vision models (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; He et al., 2016; Liu et al., 2018; Howard et al., 2019; Touvron et al., 2021; Radford et al., 2021).
Assumption	A prerequisite to comparing a machine's performance to human intelligence is, hence, the verification that machines can exhibit a sensitivity to context that would allow them to perform as well on cases that require reasoning about exceptions as on cases that require recalling generic associations.	Such a preference-based decision procedure would then allow stronger valued evidence to override weaker one.

Table 10: Comparison of 4* and non-4* sentences from different rigour criteria.

A.9 Certainty-aspect breakdown

Full breakdown of certainity-aspect for each rigour criterion. As shown in Table 11 and Table 12.

Criterion	Framing	Suggestion	Extent	Condition	Probability	Number
Settings	1.14	-0.37	0.07	0.47	-1.83	-0.55
Benchmarks	2.09	-3.71	0.87	-0.44	1.14	0.53
Biases	3.60	-1.11	-0.61	0.05	0.76	-0.21
Robustness	4.82	-3.09	-0.17	-0.21	0.93	-0.86
Reproducibility	1.50	-1.08	0.32	0.03	-0.61	-0.49
Constraints	2.12	-2.56	0.72	-2.61	5.38	0.70
Baselines	0.93	-0.70	0.08	0.02	1.69	-0.45
Limitations	2.75	0.11	-0.46	-0.37	-2.80	0.05
Generalisation	1.39	-0.36	-0.12	0.17	0.13	-0.09
Motivations	-1.84	-2.27	-0.10	0.62	-0.38	-0.32
Empirical Findings	-0.93	-0.08	-0.03	0.02	1.74	0.02
Assumptions	1.06	-1.65	0.01	1.04	-0.24	-0.34
Examples	0.74	0.13	0.15	-1.44	2.68	2.99
Justification	0.53	-1.42	-0.28	-0.13	-1.56	2.53

Table 11: Certainty Breakdown

Criterion	Framing	Suggestion	Extent	Condition	Probability	Number
Settings	0.24	0.37	0.31	0.03	0.31	-0.20
Benchmarks	0.04	-0.69	1.48	-0.06	-1.21	-0.03
Biases	-0.26	-0.21	-0.23	0.03	-1.72	-0.09
Robustness	0.05	-0.09	0.25	-0.03	-1.15	-0.44
Reproducibility	0.16	-0.27	0.38	0.02	0.03	0.02
Constraints	-0.99	-0.84	1.50	0.02	-3.26	-0.18
Baselines	-0.17	-0.40	0.35	0.03	-1.21	-0.15
Limitations	1.59	0.17	-1.43	0.00	0.65	-0.27
Generalisation	0.65	0.08	-0.55	0.00	-1.22	-0.07
Motivations	1.78	0.05	0.23	0.07	1.82	0.33
Empirical Findings	0.26	-0.02	-1.45	0.00	0.38	0.09
Assumptions	0.17	-0.10	0.47	-0.06	-0.24	-0.12
Examples	0.36	-0.03	-2.81	0.00	-2.38	-0.37
Justification	0.51	0.65	-2.72	0.00	2.19	-0.29

Table 12: Uncertainty Breakdown