Identifying Aspects in Peer Reviews

Anonymous ACL submission

Abstract

Peer review is central to academic publishing, but the growing volume of submissions is straining the process. This motivates the development of computational approaches to support peer review. While each review is tailored to a specific paper, reviewers often make assessments according to certain *aspects* such as Novelty, which reflect the values of the research community. This alignment creates opportunities for standardizing the reviewing process, improving quality control, and enabling computational support. While prior work has demonstrated the potential of aspect analysis for peer review assistance, the notion of aspect remains poorly formalized. Existing approaches often derive aspect sets from review forms and guidelines of major NLP venues, yet data-driven methods for aspect identification are largely underexplored. To address this gap, our work takes a bottom-up approach: we propose an operational definition of aspect and develop a data-driven schema for deriving fine-grained aspects from a corpus of peer reviews. We introduce a dataset of peer reviews augmented with aspects and show how it can be used for community-level review analysis. We further show how the choice of aspects can impact downstream applications, such as LLM-generated review detection. Our results lay a foundation for a principled and data-driven investigation of review aspects, and pave the path for new applications of NLP to support peer review.[1]

1 Introduction

Peer review is an essential part of academic publishing. It is a complex, multifaceted task that requires a range of competencies including paper understanding, domain-specific knowledge, and critical thinking (Shah, 2022; Yuan et al., 2022). Ensuring review quality, especially among novice reviewers, is an open challenge (Stelmakh et al., 2021a,b; Sun et al., 2024a). The increasing volume of publications puts further strain on the process, which motivates the development of computational approaches to support different stages of peer review, from reading the paper to the final decision-making by the program committees (Arous et al., 2021; Checco et al., 2021; Stelmakh et al., 2021a; Shah, 2022; Schulz et al., 2022; Yuan et al., 2022; Lin et al., 2023b; Kuznetsov et al., 2024).

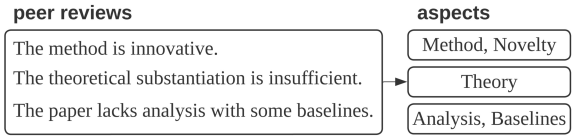


Figure 1: An aspect is a characteristic of a paper that a reviewer makes a judgment on. In this paper, we leverage an LLM to derive aspects from a corpus of peer reviews.

Peer reviews are the central component of the reviewing process. While individual review texts can vary widely, reviewers within the same research community tend to focus on a specific set of general quality categories, or aspects, such as Methodology and Novelty. These aspects can be found in review forms, instructional materials, guidelines, and the resulting review texts. Aspects allow fine-grained comparison of submissions across different dimensions of quality. A comprehensive set of aspects shared among reviewers is critical for ensuring review quality and consistency, and prior work has demonstrated the potential of aspect-based tools to support the review writing process (Sun et al., 2024a,b).

Yet, several open questions remain. First, **what is an aspect?** Most prior work derives aspects from review forms and guidelines. The lack of an operational definition of aspect prevents the comparison of aspect schemata across studies. Second, **what**

[1] Our code and data are available at https://anonymous.4open.science/r/aspects-in-reviews-737E/.

is the optimal level of aspect granularity? While prior work operates with top-down, coarse-grained aspect schemata derived from review forms and guidelines, it is unclear whether these schemata are comprehensive or provide sufficient granularity for NLP applications. Third, **how can fine-grained aspect analysis support peer review?** While aspects have been applied to certain tasks, the tasks that require a higher level of granularity are underexplored.

To address these questions, this work introduces an alternative, data-driven approach to peer review aspect analysis. We propose an operational definition of aspect grounded in its role within the evaluation process. We develop a semi-automatic approach that leverages a state-of-the-art large language model (LLM) to identify aspects in reviews, and apply it to a large collection of peer reviews to extract aspects in a bottom-up fashion (see Figure 1). Building on these results, we develop a multi-level taxonomy of aspects and present a novel dataset of peer reviews augmented with their corresponding aspects. Our dataset facilitates the exploration of two tasks: predicting the review aspects that should be focused on given a paper (paper aspect prediction), and identifying the aspects that a review focuses on (review aspect prediction). Based on these tasks, we conduct a detailed empirical study of aspect at the community level. Furthermore, we demonstrate that a comprehensive, fine-grained set of aspects allows for a new dimension in comparing reviews, offers a nuanced assessment of review quality in terms of specificity, and helps the detection of automatically generated reviews. Our work paves the path for data-driven, bottom-up analysis of aspects in peer reviews and enables new applications of NLP to support review writing. To summarize, this work contributes the following:

- We propose an semi-automatic approach which leverages a state-of-the-art LLM to derive a comprehensive set of aspects from peer reviews in a bottom-up fashion.
- We develop a taxonomy of aspect with different granularity and a new dataset of peer reviews augmented with aspects.
- We evaluate models on two tasks: predicting review aspects that should be focused on given a paper, and identifying the aspects covered in a review.
- We show how aspect analysis provides new insights into the reviewing process and how a finer-grained set of aspects helps in detecting LLM-generated reviews.

2 Related Work

2.1 Peer review in the era of LLMs

NLP for peer review is an emerging research area that aims to support different stages of the peer review process, including improving paper-reviewer matching, increasing reviewing efficiency and reproducibility, tracking dishonest behavior, and more (Shah, 2022; Schulz et al., 2022; Biswas et al., 2023; Lin et al., 2023b; Kuznetsov et al., 2024).

The emergence of LLMs has opened new opportunities for assisting peer review, such as assisting in the verification of checklists and supporting review writing (Liu and Shah, 2023; Gao et al., 2024; Liang et al., 2024b; Jin et al., 2024). In this work, we leverage an LLM to identify aspects in peer reviews, building on evidence that state-of-the-art LLMs are shown to perform well in aspect identification (Lin et al., 2023a), and in summarization, a related task to aspect identification (Pu et al., 2023).

2.2 Aspect in peer review

Beyond their role in review forms and guidelines, aspects are utilized in several other contexts related to peer review. Aspects are used in analyzing the sentiment of reviews and discovering the discourse relations within reviews (Chakraborty et al., 2020; Kennard et al., 2022). In automatic review generation, Wang et al. (2020) generate reviews using aspect-based knowledge graphs, and Gao et al. (2024) prompt LLMs with aspect-based questions to generate reviews. Aspects are included in supporting systems for review writing, which have been shown to improve the comprehensiveness of the reviews written by reviewers with varying levels of experience (Sun et al., 2024a,b). Table 1 summarizes the aspects used in the studies cited in this section. In some other review guidelines, such as the ARR review form and ACL'23 review policies, aspects are presented in an "informal" manner – they are listed as examples rather than parts of a comprehensive checklist. As noted by Kuznetsov et al. (2024), review guidelines for major NLP venues often rely on coarse-grained aspects and they lack comprehensiveness. Our work advances the study of aspects in peer review by exploring an alternative, data-driven approach to deriving finer-grained and more comprehensive aspect sets.

ACL'16, Chakraborty et al. (2020), Wang et al. (2020) Appropriateness, Clarity, Empirical/Theoretical Soundness, Impact of Ideas/Results/Dataset, Meaningful Comparison, Originality, Recommendation, Substance
ACL'18, Kennard et al. (2022), Yuan et al. (2022), Sun et al. (2024b) Clarity, Impact/Motivation, Meaningful Comparison, Originality, Replicability, Soundness/Correctness, Substance, Summary
Sun et al. (2024a) Clarity, Importance, Novelty, Validity
Wang et al. (2024) Clarity, Integrity, Novelty, Significance

Table 1: The aspects used in the studies cited in this section. Gao et al. (2024) use a model to generate aspect-based questions without using a pre-defined aspect set.

2.3 Quality of review writing

Defining the desiderata for review writing is essential to assess its effectiveness (Jefferson et al., 2002). However, the desiderata are often not well defined or operationalizable in terms of automatic measurement (Kuznetsov et al., 2024). While using simple proxies such as "helpfulness" may seem like a straightforward way to assess review quality, research has revealed that such kinds of evaluation can be biased in several ways – for example, an evaluator may be biased towards longer reviews, or those that recommend acceptance of their papers (Wang et al., 2021; Goldberg et al., 2023).

Comprehensiveness is a key desideratum for high quality reviews (Yuan et al., 2022). A frequently reported issue in ACL 2023 is the lack of specificity in reviews (Rogers et al., 2023). In this context, our work shows how fine-grained aspects can be used to compare reviews and evaluate their specificity, which contributes to a nuanced and practical assessment of review quality.

2.4 LLM-generated review detection

The strong capabilities of state-of-the-art LLMs in text generation have led to the need for detectors to identify LLM-generated contents and to prevent potential misuse (Clark et al., 2021; Gao et al., 2023; Wu et al., 2023). In peer review, LLMs pose a risk that reviewers may exploit LLMs to produce reviews entirely, which raises serious ethical concerns (Yu et al., 2024). Current strategies for detecting LLM-generated reviews align with methods for detecting LLM-generated texts: (a) LLM-as-a-judge, which prompts LLMs to identify LLM-generated contents (Zheng et al., 2023), (b) detection models, which are fine-tuned on both human and LLM-generated texts (Guo et al., 2023), (c) reference-based methods, which compare the similarity between a candidate text and one generated by an LLM (Gehrmann et al., 2019; Ippolito et al., 2020; Liang et al., 2024a; Yu et al., 2024). In our work, we provide an alternative approach – we demonstrate a comprehensive, fine-grained aspect set helps the detection of LLM-generated reviews.

3 Aspect set construction

3.1 Definition of aspect

While aspects are outlined in review guidelines and used in related work, a formal definition of aspect is lacking. To address this, we propose an operational definition of aspect as *a characteristic of a paper that a reviewer makes a judgment on when evaluating the paper, which is later used to compare the manuscripts to each other and to the quality standards provided by the review guidelines*. From this definition, we do not consider terms such as Acceptance Decision and Summary as aspects, since they are not characteristics of a paper. An aspect can be general, such as Soundness, or specific, such as Entity-aware Article Generation. However, aspects that are either too general or too specific are difficult to use for comparing manuscripts to each other or to the publication standards. Therefore, an aspect taxonomy that accommodates different levels of granularity is needed.

We assume that an aspect is expressed in a review sentence, and we allow multiple aspects per sentence. We treat the aspects of a review as a set. Formally, for a paper p, let R_p denote the set of reviews for p. For each review $r \in R_p$, we derive A_r, the set of aspects in r.

3.2 Method

To establish a comprehensive set of aspects, we selected paper reviews from NLP and machine learning (ML) conferences across different time periods. We randomly selected 50 papers from each of the NLPeer (Dycke et al., 2023) and EMNLP23[2] datasets. We used the keywords "natural" and "language" to filter NLP-related papers from ICLR. We randomly selected 50 NLP-related papers from each of the ICLR conferences from 2020 to 2024. We selected 350 papers in total, corresponding to 1094 reviews. We segmented the reviews into

[2]Publicly available through the OpenReview API.

sentences using NLTK Punkt Tokenizer (Bird and Loper, 2004) and performed the identification at the sentence level. See Appendix B for more details regarding the data.

We used OpenAI GPT-4o[3] to identify aspects from the reviews. The prompt we used is shown in Table 7. Since the identification was performed in an unsupervised setting, the identified aspects appear inconsistent, with variations like Result and Results. We post-processed the results to reduce such variations. We first identified the most frequently occurring aspects in the results, which were used as keywords to categorize and match the remaining ones. We grouped related terms together, such as Clarification and Clarity. We removed terms that we considered to be too general, such as Weakness, Strength, Question, and Comment. For the terms that cannot be matched, we omitted those that appear less than 50 times in the annotations.

We notice that our post-processing method may lead to unrelated terms being grouped together. For example, using Improvement as a keyword groups Performance Improvement and Improvement Recommendation (i.e., suggestions to improve paper quality) together. We manually checked the categorization results to verify and correct any inappropriate groupings.

In cases where a single term contains multiple aspects, it is placed into more than one category. For example, Comparison with Related Work is included in both Comparison and Related Work categories.

3.3 Results

GPT-4o identified 14574 unique aspects from the reviews. We excluded 9764 terms that were related to paper decision, too general or specific, or could not be matched by keywords and appeared less than 50 times. These excluded terms occurred 26578 times in the corpus. The most common excluded terms are Weaknesses, Questions, Strengths, Weakness, and Comments. The remaining 4810 aspects appeared 25394 times in the corpus, with the most common ones being Comparability/Comparison, Clarification/Clarity, Performance, Experiments, and Results.

Based on the results, we created a taxonomy that groups the aspects into 16 broad categories (See Table 2). This taxonomy shows the granularity of aspects across 3 levels: (a) the broad category names (**COARSE**, the most coarse), (b) the most frequently occurring aspects (**FINE**, finer), and (c) the raw GPT-4o outputs (**LLM annotation**, the finest). We also distinguish between paper-agnostic and paper-dependent aspects. Paper-agnostic aspects are relevant across all papers, while paper-dependent aspects are specific to individual papers and may not appear universally. Section A shows an example review using the proposed aspect taxonomy.

Some aspects are not present in our taxonomy as review forms for major NLP venues have dedicated fields for their evaluation, making them appear much less frequently in the review text (e.g., Reproducibility). See Table 8 for examples of the dataset we created which pairs aspects with real-life review sentences and Table 9 for the aspect frequencies. In the following sections of the paper, we use both the COARSE and FINE label sets in our experiments and analyses.

3.4 Validity check

Since LLMs have been shown to be sensitive to prompts (Chen et al., 2023; Ajith et al., 2024), we experimented with different prompts and `temperature` to assess the consistency of the model annotations. We used three consistency metrics: exact match, BERTScore similarity (Zhang et al., 2020), and Jaccard similarity. We show in Table 10 that the consistency between annotations generated under different settings is moderate to strong. 45.50% and 67.14% of the annotations obtained using different prompts and `temperature` have BERTScore similarities greater than 0.9. The increase in exact matches between the raw annotations and those mapped to the COARSE label set suggests that differences introduced by varying prompts and `temperature` do not significantly affect the categorization, as most of them still fall within the same label category. The Jaccard similarities of aspects mapped to the COARSE label set also indicate strong consistency. Table 11 and 12 help interpret these scores by showing examples of how pairs of texts and sets correspond to different BERTScore and Jaccard similarities.

We conducted a human evaluation to verify the model annotations, involving three human annotators to evaluate the LLM annotations mapped to the COARSE label set. We asked the annotators to determine whether a review sentence addresses the aspects identified by GPT-4o (see Table 14 for

[3]Version `2024-02-01`. Results were obtained between Oct 31 and Nov 20, 2024.

COARSE	FINE	LLM annotation
Contribution	Contribution	Community Contribution, Methodology Contribution
DDDDEI	Definition, Description, Detail, Discussion, Explanation, Interpretation	Dataset Description, Missing Details, Task Definition
IJMV	Intuition, Justification, Motivation, Validation	Approach Justification, Dataset Validity, Model Intuition
Novelty	Innovation/Novelty/Originality	Algorithmic Innovations, Technical Novelty
Presentation	Clarity, Figure, Grammar, Presentation, Typo	Dataset Clarification, Paper Presentation, Term Clarity
Related Work	Citation/Literature/Related Work	Existing Literature, Missing Citations, Previous Works
Significance	Impact, Importance, Significance	Empirical Importance, Practical Significance

(a) paper-agnostic

COARSE	FINE	LLM annotation
Ablation	Ablation	Ablation Analysis, Ablation Study, Ablation Tests
Analysis	Analysis	Ablation Analysis, Complexity Analysis, Data Analysis
Comparison	Comparison	Comparison Fairness, Comparison to SOTA
Data/Task	Annotation, Benchmark, Data, Task	Annotation Detail, Alternative Tasks, Data Preparation
Evaluation	Evaluation, Metric	Accuracy Metric, Evaluation Scheme, Human Evaluation
Experiment	Experiment	Control Experiment, Experimental Procedure
Methodology	Algorithm, Implementation, Method	Language Model, Methodological Soundness
Theory	Theory	Lack of Theoretical Guarantee, Theoretical Correctness
Result	Findings, Improvement, Performance, Result	BLEU Improvement, Statistical Test

(b) paper-dependent

Table 2: The taxonomy of aspects. See here for the complete taxonomy.

examples of the questionnaire entries). We observe that on average human annotators agree with 91% of the LLM annotations, along with fair inter-annotator agreement as measured by Fleiss' Kappa (Fleiss, 1971) (see Table 4). These results suggest that the model annotations largely align with human judgments, and that human annotators can effectively understand our taxonomy.

See Appendix B.2 for more details regarding the validity check.

4 Aspect prediction

Our fine-grained approach to aspect analysis enables two tasks: predicting the aspects that should be focused on given a paper (paper aspect prediction, **PAP**), and identifying the aspects that are covered in the review (review aspect prediction, **RAP**). These tasks are formalized as follows:

$$f : \begin{cases} \text{PAP}, p \rightarrow \hat{A}_p; \\ \text{RAP}, r \rightarrow \hat{A}_r; \end{cases} \quad (1)$$

where f denotes a model, $\hat{A}_p$ is the predicted aspects for a given paper, and $\hat{A}_r$ is the predicted aspects for a given review.

4.1 Method

For PAP, we only focus on predicting paper-dependent aspects, as our categorization defines paper-agnostic aspects as those that are relevant across all papers. We experimented with different parts of the paper as input, including the full paper, title, keywords, and abstract, which have different implications. Using the title, keywords, or abstract as input is a heuristic method that is grounded in statistics, that reviews for similar types of papers (as determined by the title, keywords, or abstract) tend to emphasize similar aspects. Using the full paper as input may offer broader insights. Beyond leveraging heuristics, a model with access to the full paper may also capture strengths or weaknesses in certain aspects of the paper that are not evident from the title, keywords, or abstract alone.

RAP differs from the task described in Section 3 in that it is implemented using a supervised approach. We utilized our curated data to train models to identify aspects within reviews. We segmented the reviews into sentences using NLTK.

We modeled both tasks as multi-label sequence classification. For both tasks, we tested bag-of-words with random forest (BoW+RF) and RoBERTa (Liu et al., 2019). Both models are strong in supervised settings, and they are lighter-weight alternatives to LLMs. We used focal loss (Lin et al., 2017) to address label imbalance. For PAP, we also experimented with GPT-4o in both zero-shot and few-shot settings. Our evaluation metrics include precision, recall, F1 score, and Jaccard similarity score. See Appendix C for more details regarding the experimental setup.

4.2 Results

Table 3 shows the results of PAP experiments. Overall, the models do not perform well regardless of the type of input used. We observe a small advantage for BoW+RF when using the full paper as input, possibly because the model captures more nuanced information when provided with the entire paper. Using the COARSE label set results in much higher performance than the FINE label set (see Table 18). This is partly due to the prevalence of certain COARSE aspect labels, such as Methodology, which appear in nearly all papers.

In the GPT-4o experiments[4], the evaluation of using the full paper or abstract as input is in the zero-shot setting, and the rest is in the few-shot setting. Though it is in the few-shot setting, GPT-4o outperforms RoBERTa which is trained on the full training data. GPT-4o tends to be better at capturing heuristics.

model	precision	recall	f1	Jaccard
BoW+RF	0.5343	0.6250	0.5552	0.4998
GPT-4o	0.5625	0.5542	0.5352	0.4024

(a) full paper

model	precision	recall	f1	Jaccard
BoW+RF	0.5568	0.6108	0.5511	0.4901
RoBERTa	0.5224	0.6599	0.5781	0.5439
GPT-4o	0.7138	0.7193	0.6756	0.5451

(b) keywords

Table 3: The highest precision, recall, F1 score, and Jaccard similarity; PAP; FINE label set. For GPT-4o, the evaluation of using the full paper as input is in the zero-shot setting, and that of using the keywords is in the few-shot setting. See Table 17 for more results.

Table 4 shows the results of RAP experiments. RoBERTa is the best performing model for this task. Models trained using the COARSE label set achieve higher performance compared to those trained with the FINE label set, and the few-shot setting offers small improvement for GPT-4o over the zero-shot setting (see Table 19 and 20).

model	precision	recall	f1	Jaccard
BoW+RF	0.8058	0.5810	0.6416	0.5070
RoBERTa	0.7720	0.7664	0.7675	0.7089
GPT-4o	0.5696	0.5757	0.5328	0.4573

Table 4: The highest precision, recall, F1 score, and Jaccard similarity; RAP; COARSE label set. For GPT-4o, the evaluation is in the few-shot setting.

In general, models perform better on RAP and the COARSE label set yields higher performance across both tasks.

5 Practical applications

We now demonstrate how a comprehensive, fine-grained aspect set and the proposed tasks enable new types of NLP assistance in the peer review process. We picked the trained RoBERTa model on the COARSE label set with the best F1 score and obtained predicted aspects using it. This model achieves an F1 score above 0.75 for 10 out of 17 labels (see Table 21 for the classification report).

5.1 Aspect analysis

A more comprehensive set of aspects allows for more detailed aspect analysis. Figure 2 shows the most frequent aspects across 4 submission tracks in EMNLP23.[5] While *Machine Translation* and *Multilinguality and Linguistic Diversity* are different tracks, the set of aspects that reviewers emphasize most are very similar. This suggests that there may be overlaps in the papers within these tracks (also as indicated by the track names). We observe a different pattern in the *Resources and Evaluation* track, where reviewers focus more on Data/Task and Evaluation than in other tracks.

Figure 5 in Appendix D.1 shows the similarity of submission tracks based on the Levenshtein similarity of the 10 most frequent aspects within each track in EMNLP23. Some tracks are more similar to each other. For example, *Question Answering* is most similar to tracks such as *Summarization* and *Information Extraction*.

[4]Results were obtained between Dec 28 and Dec 30, 2024.

[5]Frequencies are calculated based on the number of reviews where an aspect appears (hereinafter the same).

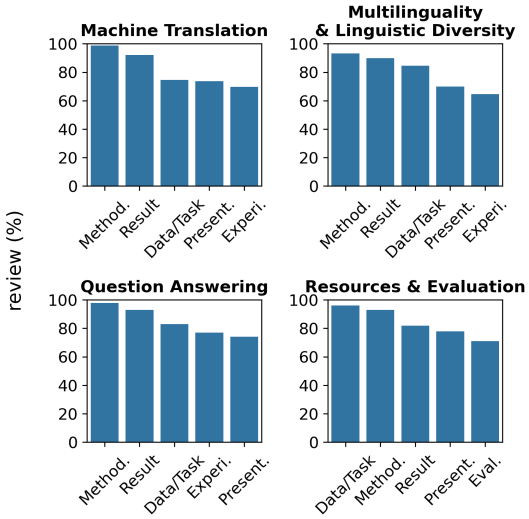


Figure 2: The 5 most frequent aspects in 4 submission tracks in EMNLP23. Figure 6 shows the full results.

Table 5 shows additional examples of how reviewers emphasize certain aspects more in some tracks than in others. For example, the frequency of Analysis is the highest in the *Computational Social Science and Cultural Analytics* track.

track	review (%)
Comp. Social Science & Cultural Analytics	69.23
Ling. Theor., Cogn. Model., & Psycholing.	68.75
Commonsense Reasoning	62.62
Multilinguality & Linguistic Diversity	60.68
Machine Learning for NLP	60.00

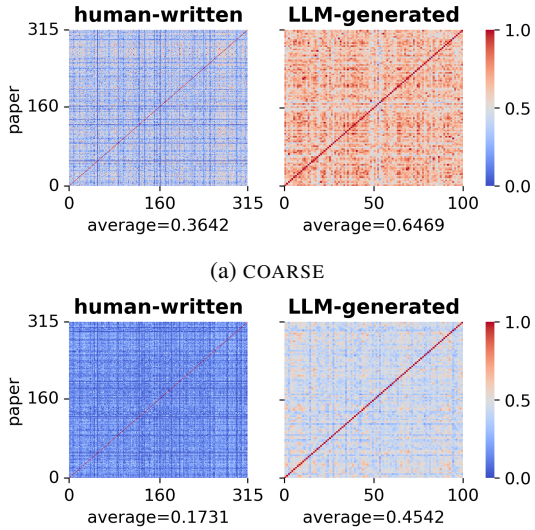
Table 5: The submission tracks with the 5 highest frequencies of Analysis in EMNLP23. Table 22 shows the full results.

This type of analysis helps compare reviewing across different tracks and venues and can inform the development of review forms and guidelines that could prompt reviewers to focus on certain aspects relevant to particular tracks to ensure a more comprehensive review.

5.2 Review comparison

In this section, we demonstrate that a more comprehensive set of aspects introduces a new dimension for comparing reviews. We compare human-written reviews with LLM-generated reviews. We used the LLM-generated reviews used in Du et al. (2024) and generated reviews for 100 randomly sampled papers from each of EMNLP23 and ICLR24. We generated reviews using GPT-4o with the prompt used in Liang et al. (2024b) and a prompt of our own (see Table 23 for the prompts).

To compare the reviews, we predicted the aspects in each review and calculated the Jaccard similarity between sets of aspects. Figure 3 visualizes the similarity between each pair of human-written and LLM-generated reviews generated using Liang et al. (2024b)'s prompt.



(a) COARSE

(b) FINE

Figure 3: The heatmap of the Jaccard similarity between each pair of the human-written reviews and LLM-generated reviews generated using EMNLP23 papers Liang et al. (2024b)'s prompt. Figure 7 in Appendix D.2 shows the rest of the results.

We obverve that LLM-generated reviews show a higher degree of similarity to each other in terms of aspects. This suggests that LLM-generated reviews may be more general than human-written ones, with the model tending to comment on similar sets of aspects across different papers. This could also imply that the quality of LLM-generated reviews is still lacking in terms of specificity.

5.3 LLM-generated review detection

We used the same data as in Section 5.2, where LLM-generated reviews are created using prompts with different levels of prompt engineering: Du et al. (2024) has the heaviest prompt engineering, followed by Liang et al. (2024b), and our own prompt is the simplest and involves the least prompt engineering (see Table 23 for more details). These prompts represent the common types of prompts used to generate reviews end-to-end. There is one LLM-generated review for each paper.

Based on our observations in Section 5.2, we

design a simple strategy to detect LLM-generated reviews. We define $sim(A_i, A_j)$ as the similarity between two sets of aspects. For each review r, we define the intra-similarity S_{intra} as the average similarity between reviews for the same paper, and the inter-similarity S_{inter} as the average similarity of reviews for different papers:

$$S_t = \frac{1}{n} \sum sim(A_r, A_{r_i}),$$
$$t = \begin{cases} \text{intra, if } r, r_i \in R_p,\ p \in P,\ r \neq r_i; \\ \text{inter, if } r \in R_p,\ r_i \in R_q,\ p, q \in P,\ p \neq q; \end{cases} \quad (2)$$

where n is the number of A_r, A_{r_i} pairs. Intuitively, reviews for the same paper are more similar to each other than reviews for different papers, as a review is specific to the paper. Therefore, we expect $|S_{intra} - S_{inter}|$ to be large. Based on this intuition, we calculate $|S_{intra} - S_{inter}|$ for each r in R_p and propose two performance metrics: (a) **@1**, which measures accuracy by determining whether the LLM-generated review is the one with the lowest score, and (b) **@2**, which calculates the percentage of cases where the LLM-generated review is among the 2 reviews with the lowest scores.

Table 6 shows the detection results. We used Jaccard similarity and compared our method with an implementation of Equation 2 using Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), which calculates the similarity between the embeddings of review texts. Experiments were conducted using both the COARSE and FINE label sets, as well as ACL'18 which is a commonly used aspect set in previous studies (see Table 1). We calculated a random baseline that selects a review at random as the LLM-generated review. Note that the FINE aspect set consistently outperform ACL'18 and SBERT, and our method is robust across LLM-generated reviews generated using prompts with different levels of prompt engineering.

Practically, the detection method introduced in this section does not perform as well as more common reference-based methods (e.g., Yu et al. (2024), which reports over 90% accuracy) or zero-shot detection using state-of-the-art LLMs (see GPT-4o results in Table 24), and the setting is specific, with only one LLM-generated review per paper. Nevertheless, our findings validate the aspects and the dataset curated in this work, and more importantly, they provide a key insight into LLM-generated reviews, that current LLMs tend to produce reviews that are general in terms of aspects, covering similar aspects across different papers.

dataset	aspect set	@1	@2
Du et al. (2024) (avg=4.80)	random	0.27	0.55
	SBERT	0.35	**0.80**
	ACL'18	0.40	**0.80**
	COARSE	**0.50**	0.75
	FINE	**0.50**	0.70
Liang et al. (2024b) (avg=4.52)	random	0.25	0.47
	SBERT	0.47	0.66
	ACL'18	0.49	0.72
	COARSE	0.53	0.73
	FINE	**0.66**	**0.89**
ours (avg=4.52)	random	0.25	0.47
	SBERT	0.55	0.71
	ACL'18	0.56	0.82
	COARSE	0.48	0.72
	FINE	**0.65**	**0.87**

Table 6: The **@1** and **@2** accuracy of the detection across different aspect sets and methods. "avg" is the average number of reviews per paper. There is one LLM-generated review for each paper.

6 Conclusion

In this paper, we introduce a data-driven approach to peer review aspect analysis. We provide an operational definition of aspect, and develop an semi-automatic approach using GPT-4o to identify aspects from peer reviews. We propose a taxonomy that involves a comprehensive set of aspects with different granularity, and introduce a new dataset of peer reviews augmented with aspects. We introduce two tasks, paper aspect prediction and review aspect prediction, and show how they contribute to a detailed empirical study of aspects. We further demonstrate that a comprehensive, fine-grained set of aspects opens a new dimension in review comparison and provides insights into the quality of automatically generated reviews.

Limitations

As discussed in Section 3.1, aspects are difficult to define. In this work, we adopted an operational definition as a practical approach. Defining individual aspects is also challenging. For instance, it is difficult to determine the scope of "Methodology." Polysemy further complicates this issue–for example, "Improvement" may refer to either a method's improvement, or places where the paper can be improved. We did a manual inspection of the categorization results to minimize the impact of these issues.

It is important to note that the categorization proposed in this work may not be optimal, and

alternative approaches to categorizing aspects may exist. Our taxonomy is an attempt to streamline the analysis and application of aspects, and it has been shown to be effective for the purposes of this study.

As reported in Section 3.4, GPT-4o shows some sensitivity to prompts and `temperature`, which may indicate uncertainty or inaccuracies in some annotations. Nevertheless, our human evaluation affirms the overall validity of the GPT-4o annotations. The applications of these aspect annotations, especially in the LLM-generated review detection task, provide further support for their reliability. Though we specify the `seed` (See Appendix B and C), the exact reproducibility of the results related to GPT-4o may still be a concern due to its closed source nature. We chose to use GPT-4o because it is among the best performing models across a wide range of tasks.

For human annotation, we recognize that using aspect labels from our taxonomy may introduce potential biases. Asking the annotators to do an open-ended annotation task without pre-defined labels seems more straight forward, but it is very challenging in terms of the cognitive burden on annotators and annotation consistency. Anchoring bias is also a concern (Tversky and Kahneman, 1974). We address this by asking annotators to provide explanations for their annotations.

All papers associated with the ICLR20-24 and EMNLP23 datasets are camera-ready versions. Though initial drafts are more preferable, we did not have access to them via OpenReview. This may affect the validity of the results of PAP experiments. However, we do not consider the use of camera-ready versions problematic in Section 5.2 and 5.3. Even if some issues raised by the reviewers were fixed in the camera-ready versions, the generated reviews are still expected to differ from each other. While the NLPeer dataset includes initial drafts, we did not use them for generating reviews due to the limited number of reviews per paper (around 2 reviews per paper).

Another limitation of our analysis is the potential biases introduced by data imbalance regarding paper acceptance. The EMNLP23 dataset contains only 9 rejected papers, which leads to a very imbalanced distribution. Similarly, in the NLPeer dataset, 69% of the papers are accepted papers (Dycke et al., 2023). These imbalances may influence the outcomes of our analysis. We chose these datasets because they are so far the most comprehensive review datasets in the domain of NLP.

Ethics Statement

This work does not suggest or imply that the proposed task of predicting the aspects to focus on a given paper may replace human involvement. Instead, it is designed to assist reviewers by recommending aspects to consider. Reviewers retain full autonomy over their decisions regarding whether to include the suggested aspects.

Since all the review data used in this study are publicly available and anonymized, processing it with commercial LLMs such as OpenAI GPT-4o does not raise ethical concerns.

Acknowledgements

References

Anirudh Ajith, Chris Pan, Mengzhou Xia, Ameet Deshpande, and Karthik Narasimhan. 2024. Instructeval: Systematic evaluation of instruction selection methods. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4336–4350. Association for Computational Linguistics.

Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2021. Peer grading the peer reviews: A dual-role approach for lightening the scholarly paper review process. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1916–1927. ACM / IW3C2.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Som Biswas, Dushyant Dobaria, and Harris L Cohen. 2023. Focus: Big data: Chatgpt and the future of journal reviews: A feasibility study. *The Yale Journal of Biology and Medicine*, 96(3):415.

Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*, pages 207–216. ACM.

Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. 2021. Ai-assisted peer review. *Humanities and Social Sciences Communications*, 8(1):1–11.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023. On the relation between sensitivity and accuracy in in-context learning. In

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 155–167, Singapore. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7282–7296. Association for Computational Linguistics.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *CoRR*, abs/2402.10886.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: statistical detection and visualization of generated text. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 111–116. Association for Computational Linguistics.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B. Shah. 2023. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *CoRR*, abs/2311.09497.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1808–1822. Association for Computational Linguistics.

Tom Jefferson, Elizabeth Wager, and Frank Davidoff. 2002. Measuring the Quality of Editorial Peer Review. *JAMA*, 287(21):2786–2790.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with LLM agents. *CoRR*, abs/2406.12708.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1234–1249. Association for Computational Linguistics.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. 2024. What can natural language processing do for peer review? *CoRR*, abs/2405.06563.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on AI conference peer reviews. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023a. The unlocking spell on base llms: Rethinking alignment via in-context learning. *CoRR*, abs/2312.01552.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023b. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *CoRR*, abs/2306.00622.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *CoRR*, abs/2309.09558.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program chairs' report on peer review at acl 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv, Toronto, Canada. Association for Computational Linguistics.

Robert Schulz, Adrian Barnett, René Bernard, Nicholas JL Brown, Jennifer A Byrne, Peter Eckmann, Małgorzata A Gazda, Halil Kilicoglu, Eric M Prager, Maia Salholz-Hillel, et al. 2022. Is the future of peer review automated? *BMC research notes*, 15(1):203.

Nihar B. Shah. 2022. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87.

Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. 2021a. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4785–4793. AAAI Press.

Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. 2021b. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):75:1–75:17.

Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024a. Reviewflow: Intelligent scaffolding to support academic peer reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI 2024, Greenville, SC, USA, March 18-21, 2024*, pages 120–137. ACM.

Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024b. Metawriter: Exploring the potential and perils of AI writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar B. Shah. 2021. Debiasing evaluations that are biased by evaluations. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10120–10128. AAAI Press.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 384–397. Association for Computational Linguistics.

Zhongyi Wang, Haoxuan Zhang, Haihua Chen, Yunhe Feng, and Junhua Ding. 2024. Content-based quality evaluation of scientific papers using coarse feature and knowledge entity network. *Journal of King Saud University - Computer and Information Sciences*, 36(6):102119.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2023. A survey on llm-generated text detection: Necessity, methods, and future directions. *CoRR*, abs/2310.14724.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. 2024. Is your paper being reviewed by an llm? investigating AI text detectability in peer review. *CoRR*, abs/2410.03019.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Using the proposed aspect set to evaluate this paper

We conducted a self-review of this paper using the aspect set proposed in this work. We maintained neutral, and the points listed below serve as examples and they are not exhaustive. This section is not generated by an LLM. For paper-agnostic aspects:

- **Contribution:** this paper derives a comprehensive set of aspects from NLP paper reviews (Section 3); this paper creates a new dataset of NLP paper reviews augmented with aspects (Section 3); this paper evaluates models using this dataset on two tasks (Section 4); this paper shows practical applications of a comprehensive set of aspects (Section 5).
- **DDDDEI:**
  - *(Definition)* this paper provides an operational definition of aspect (Section 3.1);
  - *(Description)* this paper describes the workflow regarding aspect identification and prediction;
  - *(Detail)* this paper reports detailed experimental settings and implementation specifics.
- **IJMV:**
  - *(Motivation)* this paper is motivated by the lack of a comprehensive set of aspects in the review guidelines for major NLP venues; the use of an LLM to identify aspects from reviews is motivated by evidence that shows the strong performance of LLMs in aspect identification (Section 2);
  - *(Validation)* Section 3.4 validate the method used for aspect identification.
- **Novelty:** this paper is the first attempt to derive a comprehensive set of aspects from NLP paper reviews; this paper introduces a new dataset of NLP paper reviews augmented with aspects.
- **Presentation:**
  - *(Clarity)* this paper describes the methods, experiments, and results clearly;
  - *(Figure)* this paper uses many figures to illustrate their findings.
- **Related work:** this paper reviews the opportunities and challenges of peer review in the era of LLMs, aspects in peer review, the quality of review writing, LLM-generated review detection, and scaffolding (Section 2).
- **Significance:** this paper provides a comprehensive set of aspects which benefits peer review in multiple ways, such as contributing to better review guidelines and LLM-generated review detection.

Given that our paper is both a resource and NLP application paper, for paper-dependent aspects, please pay special attention to the Data/Task and Methodology aspects (findings in Section 5.1):

- **Analysis:** this paper presents an analysis of the identification results (Section 3.3) and one in terms of track and review similarity (Section 5.1 and 5.2).
- **Comparison:** this paper compares model performance on aspect prediction (Section 4) and compares different methods for LLM-generated review detection (Section 5.3).
- **Data/Task:**
  - *(Data)* this paper uses paper reviews from both NLP and ML venues across different years, and this paper creates a new dataset of NLP paper reviews augmented with aspects (Section 3);
  - *(Task)* this paper proposes two tasks related to aspect (Section 4).
- **Evaluation:** this paper uses a range of metrics, including BERTScore, SentenceBERT, and Jaccard similarity.
- **Experiment:** this paper conducts many experiments (Section 3, 4, and 5), and all the experimental settings are reported.
- **Methodology:**
  - *(Method)* this paper proposes a workflow for identifying aspects using an LLM;
  - *(Model)* this paper uses GPT-4o, `BoW+RF`, and RoBERTa;
  - *(Framework)* this paper proposes a taxonomy of aspects;
  - *(Implementation)* all the implementation details are provided.
- **Result:**
  - *(Findings)* this paper finds that LLM-generated reviews are more generic than human-written reviews (Section 5.2);
  - *(Performance)* this paper shows that models trained using the aspect sets proposed in this work perform better than using previous ones (Section 5.3).

Return to main text.

B More on aspect identification

The EMNLP23 and ICLR data were obtained via the OpenReview API. We used the prompt in Table 7 to identify aspects from the reviews. We set `temperature=0` and `seed=2266`. We segmented the reviews into sentences, and removed entries that consist of only indices (e.g., "1.").

```
Identify the aspect(s) that each of the given sentences focuses on. Format the output in a json dictionary. For example, given a dictionary as follows:

{"1": "The methodology is convincing, and the improvement is noticeable.", "2": "A dataset is assembled."}

The output should be:

{"1": "Methodology, Improvement", "2": "Dataset"}
```

Table 7: The prompt used to identify aspects from the reviews. Return to main text.

B.1 More on results

Table 8 shows examples from the dataset we created. Each review sentence is accompanied by an LLM annotation, which is mapped to both the COARSE and FINE label sets. Table 9 shows the frequency of aspects of different levels of granularity in the dataset we created.

B.2 More on validity check

We experimented with a different prompt, where we replaced the word "aspect" in the prompt shown in Table 7 with a synonym "facet". This prompt is designed to determine whether the LLM truly understands the semantics of the prompt rather than relying on specific word choices. We also tested with `temperature=[0,1]`.

Table 10 shows the consistency between annotations generated under different settings. Table 11 and 12 show examples of how pairs of texts and sets correspond to different BERTScore and Jaccard similarities.

For the human annotation, we recruited annotators through Prolific. We applied screening conditions to ensure quality: participants are required to hold a graduate or doctorate degree in Computer Science, have English as their first language, and have an approval rate above 90% for previous Prolific submissions. We conducted several pilot studies and selected 3 annotators for the full study. See here for the annotation guidelines.

We sampled review sentences to maintain class balance as much as possible (the distribution of aspects in the annotation file is shown in Table 13). We sampled 100 reviews, which corresponds to 1852 sentences. We asked the annotators whether each review sentence addresses the aspects identified by GPT-4o. To mitigate anchoring bias (Tversky and Kahneman, 1974) and prevent participants from simply confirming all queries, we required them to provide explanations for their evaluations. Table 14 shows examples of the questionnaire entries we distributed. The total number of entries in the questionnaire is 3032.

The "yes" rates, i.e., the agreement between human and LLM annotations, are 85.19%, 90.11%, and 97.56% for the three annotators (with 100% indicating complete agreement with all LLM annotations). The Fleiss' Kappas (Fleiss, 1971), shown in Figure 4, indicate substantial inter-annotator agreement for the first 400 entries in the questionnaire, but it declined as the annotation progressed, suggesting that annotation quality may not have been consistent throughout. It is important to note that there should not be a significant difference in difficulty between the first 400 entries and the remaining ones, as all entries were randomly sampled.

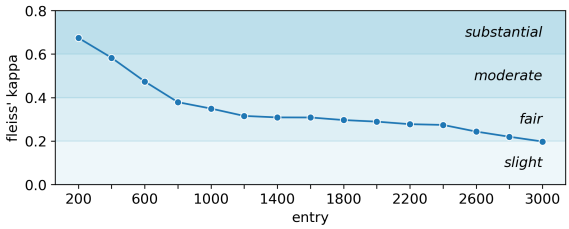


Figure 4: The Fleiss' Kappas for the annotations across different ranges. For example, "600" on the x-axis indicates that the first 600 annotations have a Fleiss' Kappa of 0.4739. The overall Fleiss' Kappa is 0.1944. Return to main text.

Table 15 shows the agreement between human and LLM annotations across different aspects. Agreement exceeds 0.9 for most aspects, with the highest agreement observed for Contribution and Ablation, while that for Significance and Presentation is the weakest.

review sentence	LLM **annotation**	COARSE	FINE
The proposed method demonstrates commendable innovation, standing apart from mere amalgamation of existing models.	Innovation, Method	Methodology, Novelty	Method, Novelty
The novel approach showcases tangible efficacy without introducing additional parameters.	Efficacy, Novel Approach, Parameters	Methodology, Novelty	Approach, Novelty, Parameter
1.The absence of a computational complexity analysis, coupled with marginal and non-significant experimental performance improvements, raises concerns regarding the practical significance.	Computational Complexity, Experimental Performance, Practical Significance	Experiment, Methodology, Result, Significance	Complexity, Experiment, Performance, Significance
If the proposed model introduces high computational complexity and yields only marginal gains, its real-world utility may be limited.	Computational Complexity, Real-world Utility, Marginal Gains	Methodology	Complexity
2.The paper lacks comparative analysis with some important baselines, such as P-tuning v2.	Comparative Analysis, Baselines	Analysis, Comparison	Analysis, Baseline
Additionally, the theoretical substantiation for the proposed method is insufficiently detailed.	Theoretical Substantiation, Method	Methodology, Theory	Method, Theory
The exclusive use of a single language model, T5-base, as the backbone prompts doubts about the general applicability of the proposed approach across a broader spectrum of models.	Language Model, General Applicability	Methodology	Application, Model
Furthermore, the paper lacks visual or interpretable analyses that incorporate concrete natural language statements.	Visual Analysis, Interpretability, Natural Language Statements	Analysis, DDDDEI	Analysis, Interpretation
Considering these points, I respectfully recommend that the authors thoroughly address these shortcomings to enhance the paper's overall quality and potential for contribution before reconsidering it for acceptance.	Recommendations, Paper Quality, Contribution	Contribution	Contribution

Table 8: Examples from the dataset we created. Each of the review sentence is augmented with aspects. See here for the complete dataset. Return to main text. Return to appendix.

COARSE	FINE	LLM annotation
Methodology (34.47%)	Model (26.29%), Method (19.38%), Training (11.77%), Approach (4.62%), Parameter (4.50%)	Methodology (7.03%), Method (5.00%), Model (2.63%), Training (2.60%), Generalizability/Generalization (2.12%)
Result (12.47%)	Result (36.12%), Performance (35.49%), Improvement (14.92%), Accuracy (4.85%), Robustness (2.75%)	Performance (23.20%), Results (20.06%), Improvement (9.32%), Experimental Results (4.59%), Accuracy (3.17%)
Data/Task (9.98%)	Data (53.89%), Task (38.26%), Benchmark (4.75%), Annotation (3.10%)	Dataset (11.64%), Datasets (9.08%), Tasks (6.14%), Task (2.83%), Data (2.07%)
Presentation (9.68%)	Clarity (37.40%), Presentation (19.24%), Figure (14.41%), Table (11.67%), Typo (6.11%)	Clarification/Clarity (30.37%), Writing Quality (12.12%), Presentation (5.09%), Figure/Visualization (4.00%), Readability (3.92%)
Comparison (6.00%)	Comparison (74.97%), Baseline (25.03%)	Comparability/Comparison (59.54%), Baselines (12.74%), Baseline (9.72%), Comparisons (5.02%), Model Comparison (1.44%)
Experiment (5.33%)	Experiment (100.00%)	Experiments (50.99%), Experimental Results (10.72%), Experiment (9.38%), Experimental Setup (1.83%), Experimentation (1.69%)
DDDDEI (4.57%)	Explanation (29.61%), Discussion (23.36%), Definition (14.23%), Description (13.57%), Detail (11.76%)	Explainability/Explanation (24.18%), Discussion (16.61%), Description (6.00%), Definition (5.76%), Interpretability/Interpretation (5.26%)
Related Work (3.93%)	Related Work (100.00%)	Reference (20.67%), Related Work (17.99%), References (13.59%), Previous Work (8.23%), Citation (6.70%)
Evaluation (3.80%)	Evaluation (73.90%), Metric (26.10%)	Evaluation (47.33%), Metrics (8.32%), Metric (5.05%), Human Evaluation (3.96%), Empirical Evaluation (2.97%)
IJMV (2.07%)	Motivation (53.27%), Justification (17.82%), Validation (15.09%), Intuition (13.82%)	Motivation (48.91%), Justification (13.45%), Intuition (12.91%), Validity/Validation (9.09%), Motivations (1.82%)
Analysis (1.99%)	Analysis (100.00%)	Analysis/Analytics (61.32%), Theoretical Analysis (4.15%), Error Analysis (3.96%), Qualitative Analysis (2.64%), Empirical Analysis (1.70%)
Novelty (1.74%)	Novelty (100.00%)	Novelty (76.67%), Innovation/Novelty/Originality (8.64%), Technical Novelty (3.67%), Novel Approach (3.02%), Technical novelty (0.86%)
Contribution (1.49%)	Contribution (100.00%)	Contribution (50.51%), Contributions (19.44%), Technical Contribution (4.55%), Main Contribution (3.79%), Core Contribution (1.77%)
Significance (1.00%)	Significance (37.45%), Importance (32.21%), Impact (30.34%)	Significance (35.96%), Importance (25.84%), Impact (20.60%), Problem Importance (3.37%), Performance Impact (2.25%)
Ablation (0.91%)	Ablation (100.00%)	Ablation Study (36.36%), Ablation Studies (23.97%), Ablations (12.81%), Ablation (7.44%), Ablation Experiments (4.55%)
Theory (0.55%)	Theory (100.00%)	Theory (18.37%), Theoretical Analysis (14.97%), Theoretical Results (6.12%), Theorem (5.44%), Theoretical results (3.40%)

Table 9: The aspects of different levels of granularity and their frequency in the dataset we created. We show the 5 most frequent aspects at each level. For example, Methodology appears in 35.30% of the review sentences, and the 5 most frequent FINE labels associated with Methodology are Model, Method, Training, Approach, and Parameter. Among all review sentences containing Methodology, the FINE label Model appears in 24.92% of the cases, and the LLM annotation Methodology appears in 6.54% of the cases. Return to main text. Return to appendix.

aspect	metric	$p=p_0$	$t=0$
raw	exact match	20.34%	15.31%
	BERTScore ≥ 0.9	67.14%	45.50%
COARSE	exact match	60.40%	53.43%
	Jaccard similarity	0.7112	0.6528

Table 10: The consistency scores between GPT-4o annotations obtained using different `temperature` ($p=p_0$, where p_0 refers to the prompt in Table 7), and those obtained using different prompts (`t=0`, where `t` refers to `temperature`). We calculated consistency scores using both the raw annotations and those mapped to the COARSE label set. Return to main text. Return to appendix.

BERTScore	annotation 1	annotation 2
≥ 0.90	Technique, Analysis, Task Writing Quality, Clarity Explanation, Settings, Obscurity Analysis, Tasks, Experiments, Findings Experiment Setup, Text Classification	Technique, Analysis, Task Suitability, Improvement Writing Clarity Explanation, Experimental details, Clarity Analysis, Experiments, Robustness Experiment Setup, Recency
< 0.90	Lack of definition, Context Clarity, Methodology Model Specification Generalization Performance, Frame Model Architecture	Definition, Paper Explanation Model, Structure, Clarity Experiment Hypothesis, Generalization Encoder, Decoder, Tensor Product Representation

Table 11: How BERTScores for different pairs look like. Return to main text. Return to appendix.

Jaccard similarity	set 1	set 2
0.9091	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}
0.9000	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 8, 9}
0.8182	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 8, 9, 11}
0.8000	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 8}
0.7500	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12}
0.6667	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 8, 11, 12}
0.5833	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 11, 12}
0.5385	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 7, 11, 12, 13}
0.4615	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 11, 12, 13}
0.4286	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 6, 11, 12, 13, 14}
0.3571	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	{1, 2, 3, 4, 5, 11, 12, 13, 14}

Table 12: How Jaccard similarity for different pairs look like. Return to main text. Return to appendix.

Method.	Result	Data/Task	Presentation	Comparison	DDDDEI	Experiment	Related Work
846	393	304	270	186	171	155	113

Analysis	Evaluation	IJMV	Novelty	Contribution	Theory	Ablation	Significance
100	85	77	72	71	65	63	61

Table 13: The distribution of aspects in the questionnaire we distributed. Return to appendix.

review	question	yes	no	explanation
I am also somewhat confused by the second set of experiments.	Does the review address Experiment?	□	□	
Proposes a novel approach by combining ideas from active learning and human-AI collaboration.	Does the review address Methodology?	□	□	
Proposes a novel approach by combining ideas from active learning and human-AI collaboration.	Does the review address Novelty?	□	□	

Table 14: Examples of the entries in the questionnaire we distributed. Note that a single review sentence appears multiple times in the questionnaire if it covers multiple aspects. Return to main text. Return to appendix.

Contribution	Ablation	Evaluation	Novelty	IJMV	Analysis	Comparison	DDDDEI
0.9953	0.9947	0.9843	0.9769	0.9740	0.9733	0.9695	0.9688

Theory	Experiment	Related Work	Data/Task	Methodology	Result	Significance	Presentation
0.9641	0.9505	0.9204	0.9145	0.8928	0.8617	0.8361	0.8225

Table 15: The agreement between human and LLM annotations across different aspects. Return to appendix.

C More on aspect prediction

The dataset is split into 90% for training and 10% for testing. We implemented bag-of-words models with random forest (BoW+RF) using `scikit-learn`. We used `RandomForestClassifier`, and we set `n_estimators=100`. For the RoBERTa models, we set `batch_size=16` and `learning_rate=3e-5`. The models were trained for 10 epochs. For GPT-4o experiments, we set `temperature=0`. We set `seed=2266` for all the experiments.

The equation for focal loss is given in 3, and we experimented with $\alpha = [0.1, 0.2, 0.3]$ and $\gamma = [1.5, 2.0, 2.5]$.

$$\mathcal{L}_{\text{focal}}(p) = -\alpha_t(1-p)^{\gamma}\log(p). \quad (3)$$

Table 16 shows the few-shot prompt. We selected exemplars from the training set.

```
Identify the aspect(s) that the given sentence
focuses on. Aspects: $COARSE/FINE ASPECTS$.
If a sentence focuses on none of these aspects,
mark it as "-". Format the output in a json
dictionary: {"Aspects": [...]}. Here are some
examples:

$EXAMPLES$
```

Table 16: The few-shot prompt used to predict aspects. Return to main text.

Table 17, 18, 19, and 20 show additional results of PAP and RAP. Table 21 shows a classification report of an RAP model.

model	precision	recall	f1	Jaccard
BoW+RF	0.5254	0.6179	0.5435	0.4928
RoBERTa	0.5187	0.6644	0.5764	0.5378
GPT-4o	0.6980	0.3892	0.4455	0.3319

(a) abstract

model	precision	recall	f1	Jaccard
BoW+RF	0.5483	0.6203	0.5481	0.4957
RoBERTa	0.5240	0.6599	0.5781	0.5327
GPT-4o	0.7324	0.6840	0.6561	0.5246

(b) title

Table 17: The highest precision, recall, F1 score, and Jaccard similarity of predictions made by models trained on PAP using the FINE label set. For GPT-4o, the evaluation of using the abstract as input is in the zero-shot setting, and that of using the title in the few-shot setting. Return to main text.

model	precision	recall	f1	Jaccard
BoW+RF	0.8694	0.9430	0.8909	0.7952
GPT-4o	0.8580	0.7215	0.7452	0.6813

(a) full paper

model	precision	recall	f1	Jaccard
BoW+RF	0.8139	0.9177	0.8527	0.7725
RoBERTa	0.8658	0.9586	0.9048	0.8339
GPT-4o	0.8700	0.5633	0.6392	0.5589

(b) abstract

model	precision	recall	f1	Jaccard
BoW+RF	0.8308	0.9114	0.8575	0.7625
RoBERTa	0.8944	0.9290	0.8864	0.8381
GPT-4o	0.8835	0.9684	0.9200	0.8471

(c) keywords

model	precision	recall	f1	Jaccard
BoW+RF	0.8139	0.9177	0.8528	0.7640
RoBERTa	0.8654	0.9645	0.9048	0.8266
GPT-4o	0.8913	0.9494	0.9171	0.8448

(d) title

Table 18: The highest precision, recall, F1 score, and Jaccard similarity of predictions made by models trained on PAP using the COARSE label set. For GPT-4o, the evaluation of using the full paper or abstract as input is in the zero-shot setting, and the rest is in the few-shot setting. Return to main text.

model	precision	recall	f1	Jaccard
BoW+RF	0.7392	0.3817	0.4689	0.3447
RoBERTa	0.7413	0.7146	0.7196	0.6402
GPT-4o	0.6426	0.7092	0.6309	0.5217

Table 19: The highest precision, recall, F1 score, and Jaccard similarity of model predictions on RAP using the FINE label set. For GPT-4o, the evaluation is in the few-shot setting. Return to main text.

label set	precision	recall	f1	Jaccard
COARSE	0.5526	0.5646	0.5145	0.4341
FINE	0.6174	0.6939	0.6073	0.5041

Table 20: The highest precision, recall, F1 score, and Jaccard similarity of model predictions on RAP using GPT-4o in the zero-shot setting. Return to main text.

category	precision	recall	f1	support
Ablation	0.85	0.92	0.88	12
Analysis	0.83	0.77	0.80	39
Comparison	0.81	0.74	0.77	125
Contribution	0.79	0.87	0.83	31
Data/Task	0.75	0.85	0.79	212
DDDDEI	0.71	0.71	0.71	168
Evaluation	0.87	0.75	0.81	114
Experiment	0.93	0.82	0.87	99
IJMV	0.73	0.76	0.74	46
Methodology	0.80	0.83	0.81	602
Novelty	0.81	0.79	0.80	28
Presentation	0.75	0.70	0.73	279
Related Work	0.79	0.70	0.74	150
Result	0.82	0.84	0.83	266
Significance	0.79	0.41	0.54	27
Theory	0.00	0.00	0.00	0
None	0.66	0.69	0.67	401
AVERAGE (W.)	0.77	0.77	0.77	-

Table 21: The classification report of the predictions made by the best performing model on RAP (in terms of F1 score). Weighted average scores (AVERAGE (W.)) are reported. This model was trained using the COARSE label set. Return to main text. Return to appendix.

D More on practical applications

D.1 More on aspect analysis

Figure 5 shows the Levenshtein similarity of the set of the 10 most frequent aspects in the submission tracks in EMNLP23. It shows that the 1th to 18th tracks are more similar to each other. The 1th to 18th tracks are: **(1)** *Question Answering*, **(2)** *Information Retrieval and Text Mining*, **(3)** *Phonology, Morphology, and Word Segmentation*, **(4)** *Human-Centered NLP*, **(5)** *Machine Learning for NLP*, **(6)** *Natural Language Generation*, **(7)** *Discourse and Pragmatics*, **(8)** *Speech and Multimodality*, **(9)** *Summarization*, **(10)** *Computational Social Science and Cultural Analytics*, **(11)** *Interpretability, Interactivity, and Analysis of Models for NLP*, **(12)** *NLP Applications*, **(13)** *Linguistic Theories, Cognitive Modeling, and Psycholinguistics*, **(14)** *Dialogue and Interactive Systems*, **(15)** *Resources and Evaluation*, **(16)** *Information Extraction*, **(17)** *Language Modeling and Analysis of Language Models*, **(18)** *Language Grounding to Vision, Robotics and Beyond*.

The 19th to 27th tracks are not so similar to most of the other tracks. The 19th to 27th tracks are: **(19)** *Sentiment Analysis, Stylistic Analysis, and Argument Mining*, **(20)** *Syntax, Parsing and their Applications*, **(21)** *Commonsense Reasoning*, **(22)** *Multilinguality and Linguistic Diversity*, **(23)** *Machine Translation*, **(24)** *Ethics in NLP*, **(25)** *Efficient Methods for NLP*, **(26)** *Semantics: Lexical, Sentence level, Document Level, Textual Inference, etc.*, **(27)** *Theme Track: Large Language Models and the Future of NLP*.

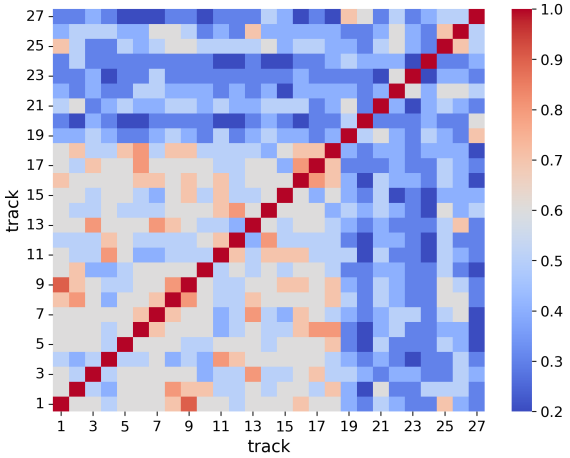


Figure 5: The Levenshtein similarity of the 10 most frequent aspects within each submission track in EMNLP23. Return to main text.

Figure 6 shows the 5 most frequent aspects in each of the submission tracks in EMNLP23.

Table 22 shows the frequency of Analysis and DDDDEI across all the submission tracks in EMNLP23.

D.2 More on review comparison

Table 23 shows the prompts used to generate reviews used in Section 5.2.

Figure 7 shows more results regrading the comparison of human-written and LLM-generated reviews.

D.3 More on LLM-generated review detection

We used PyPDF2 to convert PDFs to text files, and we only generate reviews for the main text (cut contents after reference). We used GPT-4o, set `temperature=0`, `seed=2266`, and `max_tokens=2048`.

Table 24 shows the zero-shot performance of GPT-4o on LLM-generated review detection.

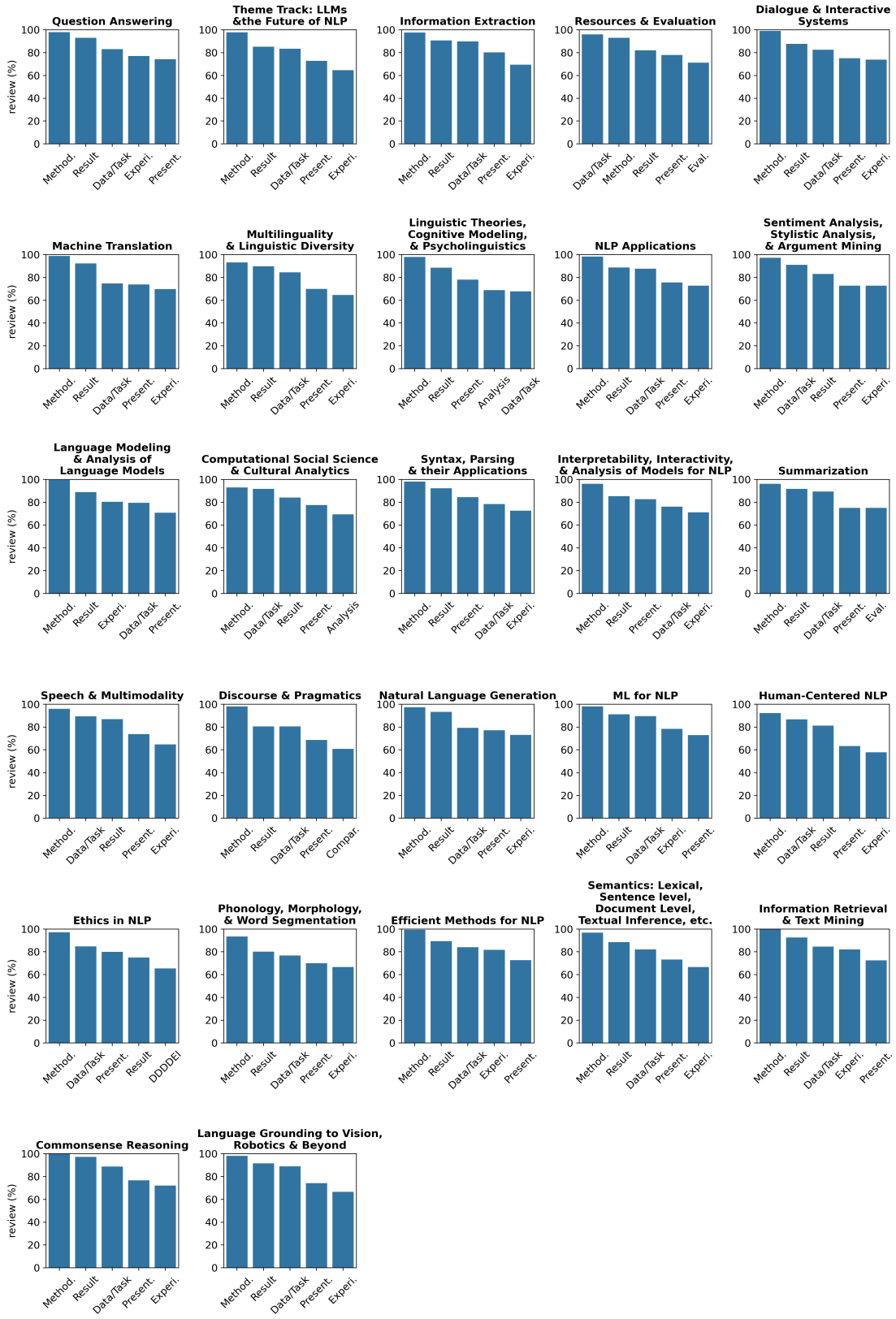


Figure 6: The 5 most frequent aspects in each of the submission tracks in EMNLP23. Return to main text. Return to appendix.

track	review (%)
Computational Social Science and Cultural Analytics	69.23
Linguistic Theories, Cognitive Modeling, and Psycholinguistics	68.75
Commonsense Reasoning	62.62
Multilinguality and Linguistic Diversity	60.68
Machine Learning for NLP	60.00
Machine Translation	57.38
Discourse and Pragmatics	56.86
Phonology, Morphology, and Word Segmentation	56.67
Interpretability, Interactivity, and Analysis of Models for NLP	56.37
Sentiment Analysis, Stylistic Analysis, and Argument Mining	55.66
Theme Track: Large Language Models and the Future of NLP	55.56
Information Retrieval and Text Mining	55.50
NLP Applications	54.48
Summarization	54.44
Resources and Evaluation	54.35
Efficient Methods for NLP	53.07
Information Extraction	53.06
Syntax, Parsing and their Applications	52.94
Language Modeling and Analysis of Language Models	51.85
Dialogue and Interactive Systems	51.23
Speech and Multimodality	51.01
Ethics in NLP	50.96
Semantics: Lexical, Sentence level, Document Level, Textual Inference, etc.	48.15
Language Grounding to Vision, Robotics and Beyond	47.91
Human-Centered NLP	47.78
Question Answering	45.58
Natural Language Generation	42.34

(a) Analysis

track	review (%)
Ethics in NLP	65.38
Linguistic Theories, Cognitive Modeling, and Psycholinguistics	60.42
Interpretability, Interactivity, and Analysis of Models for NLP	59.46
NLP Applications	57.17
Information Extraction	56.85
Resources and Evaluation	56.09
Semantics: Lexical, Sentence level, Document Level, Textual Inference, etc.	56.02
Human-Centered NLP	54.44
Machine Learning for NLP	53.56
Information Retrieval and Text Mining	53.5
Speech and Multimodality	52.02
Sentiment Analysis, Stylistic Analysis, and Argument Mining	51.89
Language Modeling and Analysis of Language Models	51.85
Language Grounding to Vision, Robotics and Beyond	50.95
Natural Language Generation	50.90
Computational Social Science and Cultural Analytics	50.30
Theme Track: Large Language Models and the Future of NLP	49.90
Summarization	49.44
Multilinguality and Linguistic Diversity	49.03
Dialogue and Interactive Systems	48.77
Discourse and Pragmatics	47.06
Efficient Methods for NLP	46.93
Question Answering	45.94
Commonsense Reasoning	42.99
Machine Translation	39.34
Syntax, Parsing and their Applications	39.22
Phonology, Morphology, and Word Segmentation	33.33

(b) DDDDEI

Table 22: The frequencies of Analysis and DDDDEI across all the submission tracks in EMNLP23. Return to main text. Return to appendix.

Du et al. (2024)

As an esteemed reviewer with expertise in the field of Natural Language Processing (NLP), you are asked to write a review for a scientific paper submitted for publication. Please follow the reviewer guidelines provided below to ensure a comprehensive and fair assessment:

Reviewer Guidelines: {review_guidelines}

In your review, you must cover the following aspects, adhering to the outlined guidelines:

Summary of the Paper: Provide a concise summary of the paper, highlighting its main objectives, methodology, results, and conclusions.

Strengths and Weaknesses: Critically analyze the strengths and weaknesses of the paper. Consider the significance of the research question, the robustness of the methodology, and the relevance of the findings.

Clarity, Quality, Novelty, and Reproducibility: Evaluate the paper on its clarity of expression, overall quality of research, novelty of the contributions, and the potential for reproducibility by other researchers.

Summary of the Review: Offer a brief summary of your evaluation, encapsulating your overall impression of the paper.

Correctness: Assess the correctness of the paper's claims; you are only allowed to choose from the following options: {Explanation on different correctness scores}

Technical Novelty and Significance: Rate the technical novelty and significance of the paper's contributions; you are only allowed to choose from the following options: {Explanation on different Technical Novelty and Significance scores}

Empirical Novelty and Significance: Evaluate the empirical contributions; you are only allowed to choose from the following options: {Explanation on different Empirical Novelty and Significance scores}

Flag for Ethics Review: Indicate whether the paper should undergo an ethics review [YES or NO].

Recommendation: Provide your recommendation for the paper; you are only allowed to choose from the following options: {Explanation on different recommendation scores}

Confidence: Rate your confidence level in your assessment; you are only allowed to choose from the following options: {Explanation on different confidence scores}

To assist in crafting your review, here are two examples from reviews of different papers:

Review Example 1: {review_example_1}

Review Example 2: {review_example_2}

Follow the instruction above, write a review for the paper below:

Liang et al. (2024b)

Your task now is to draft a high-quality review outline for the given submission.
======
Your task:
Compose a high-quality peer review of a paper.

Start by "Review outline:".
And then:
"**1. Significance and novelty**"
"**2. Potential reasons for acceptance**"
"**3. Potential reasons for rejection**", List multiple key reasons. For each key reason, use **>=2 sub bullet points** to further clarify and support your arguments in painstaking details. Be as specific and detailed as possible.
"**4. Suggestions for improvement**", List multiple key suggestions. Be as specific and detailed as possible.

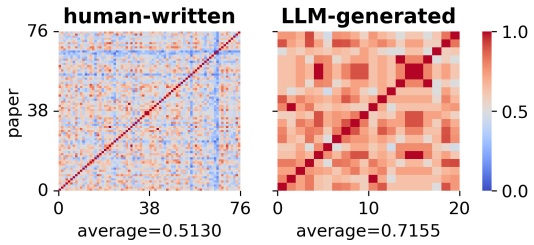
Be thoughtful and constructive. Write Outlines only.

---

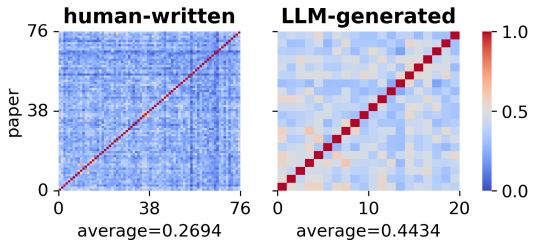
ours

Write a paper review for the following paper regarding its strengths and weaknesses.

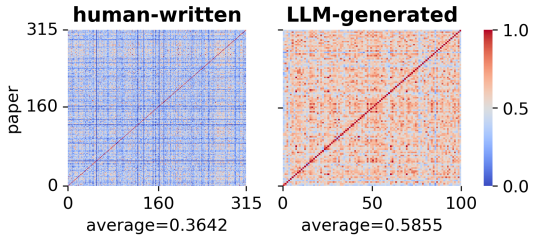
Table 23: The prompts used to generate reviews. Return to main text. Return to appendix.



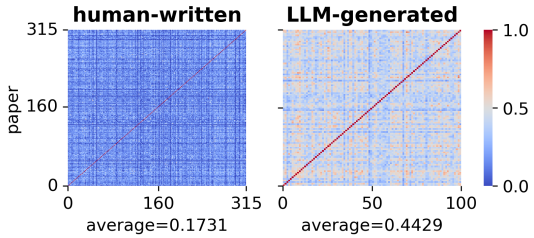
(a) Du et al. (2024), COARSE



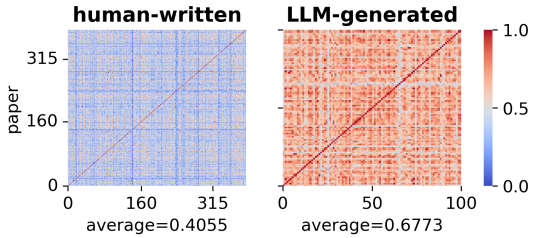
(b) Du et al. (2024), FINE



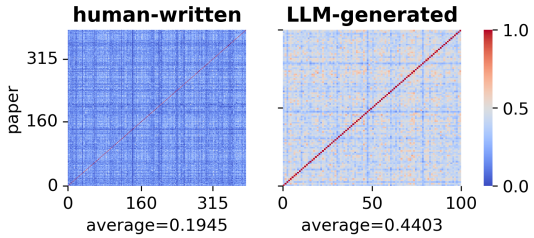
(c) ours, COARSE (EMNLP23)



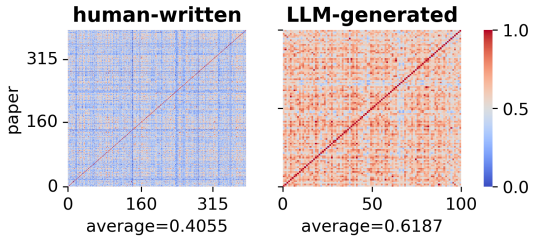
(d) ours, FINE (EMNLP23)



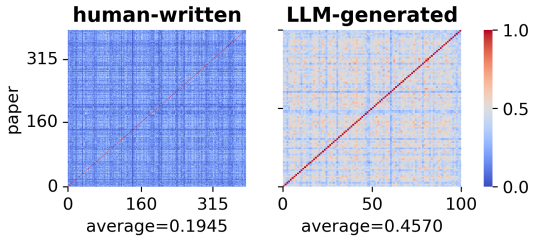
(e) Liang et al. (2024b), COARSE (ICLR24)



(f) Liang et al. (2024b), FINE (ICLR24)



(g) ours, COARSE (ICLR24)



(h) ours, FINE (ICLR24)

Figure 7: The heatmap of the Jaccard similarity between each pair of the human-written reviews and LLM-generated reviews using Du et al. (2024)'s, Liang et al. (2024b)'s, and our prompt. Return to main text. Return to appendix.

dataset	number of data	#reviews per paper	accuracy
Du et al. (2024)	20	4.80	0.80
Liang et al. (2024b)	200	4.52	0.75
ours	200	4.52	0.89

Table 24: The accuracy of GPT-4o in zero-shot LLM-generated review detection. Return to main text. Return to appendix.