CRISP: Characterizing Relative Impact of Scholarly Publications

Hannah Collison hgonza13@jhu.edu
Benjamin Van Durme vandurme@jhu.edu
Daniel Khashabi danielk@cs.jhu.edu

Johns Hopkins University

Abstract

Assessing a cited paper's impact is typically done by analyzing its citation context *in isolation* within the citing paper. While this focuses on the most directly relevant text, it prevents *relative* comparisons across all the works a paper cites. We propose CRISP, which instead jointly ranks *all* cited papers within a citing paper using large language models (LLMs). To mitigate LLMs' positional bias, we rank each list three times in a randomized order and aggregate the impact labels through majority voting. This joint approach leverages the full citation context, rather than evaluating citations independently, to more reliably distinguish impactful references. CRISP outperforms a prior state-of-the-art impact classifier by $+9.5\%$ accuracy and $+8.3\%$ F1 on a dataset of human-annotated citations. CRISP further gains efficiency through fewer LLM calls and performs competitively with an open-source model, enabling scalable, cost-effective citation impact analysis. We release our rankings, impact labels, and codebase to support future research.

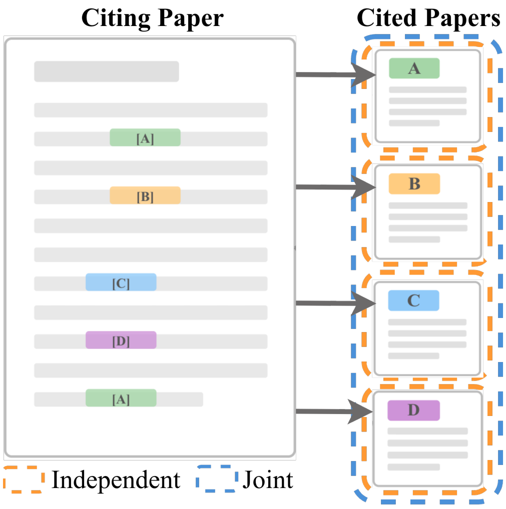


Figure 1: A citing paper p (left) references papers $\{A, B, C, D\}$ (right), with each arrow denoting a citation edge $(p \rightarrow \tilde{p})$ for $\tilde{p} \in \{A, B, C, D\}$. CRISP considers all cited papers *jointly* to assess the *relative* impact $\tilde{p}$ had on p, unlike prior work that considers only p and $\tilde{p}$ in isolation, evaluating impact independently.

1 Introduction

Funding agencies, hiring committees, and researchers routinely rely on measures of *scientific impact* to make consequential decisions; yet most such measures reduce a paper's influence to noisy proxy measures such as citation count (Garfield, 1972). Citation count alone provides a poor measure of impact because not all citations are equally important (Zhu et al., 2015; Aguinis et al., 2012). For example, a citation providing background information contributes less to the citing paper than one that adopts the cited paper's methodology (Hassan et al., 2017).

These challenges have motivated a line of research on citation intent and impact classification (e.g., Valenzuela et al., 2015; Jurgens et al., 2018; Arnaout et al., 2025), which aims to assess the impact and intent behind each citation to a prior work mentioned in a given paper. As illustrated in Figure 1, these approaches evaluate each citation edge independently, assessing each citation based on its surrounding text, capturing *how* and *why* a paper is cited — and to classify citations as meaningful or incidental. Focusing on each citation context *in isolation* is a natural choice: it concentrates on the most directly relevant text as highlighted in Figure 1, and keeps context length short to avoid the risk of positional biases and the cost of inference. Yet evaluating citations in isolation discards a valuable signal: the *relative importance* of a reference compared to the other works cited alongside it.

We propose CRISP, a method that *jointly* ranks *all* cited papers within a citing paper using LLMs to characterize their *relative* impact (§4). As shown in Figure 1, jointly analyzing all citation contexts within a single citing paper introduces this comparative dimension, even though the additional con-

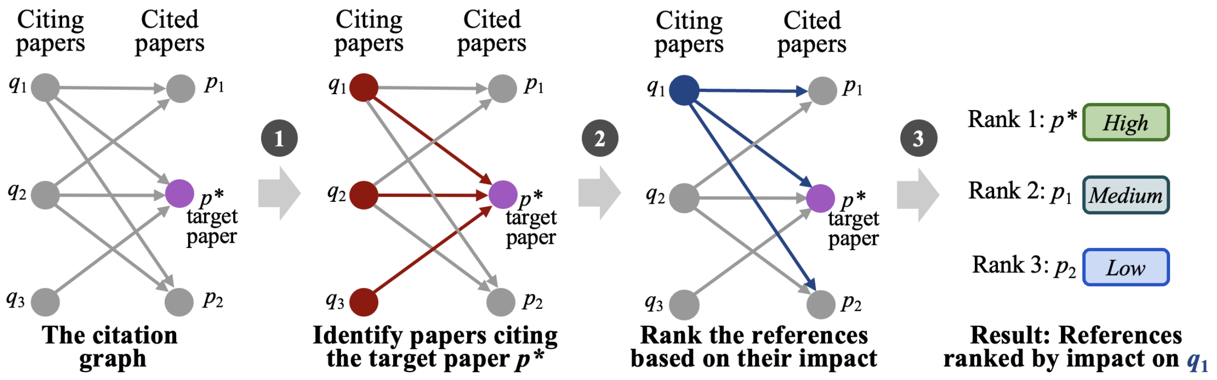


Figure 2: Overview of CRISP, which characterizes the relative impact of a target paper p^* on each of its citing papers, e.g. q_1. ❶ Identify all citing papers of p^* (q_1, q_2, q_3). ❷ For each citing paper, rank its references by impact (e.g., p_1, p_2, p^* for q_1). ❸ The result is an impact-based ranking of q_1's references. See §4 for details.

text risks distracting the model. To mitigate LLMs' known positional bias (Tang et al., 2024), we rank the list of cited papers three times, each with a random order, and aggregate the resulting impact labels via majority voting. Surprisingly, this joint ranking approach—leveraging comparative context rather than evaluating citations in isolation—more effectively distinguishes impactful references, outperforming a prior state-of-the-art impact classifier (Arnaout et al., 2025) by $+9.5\%$ in accuracy and $+8.3\%$ F1 on average across different LLMs on their dataset of citations with human-annotated impact labels (§6). We release rankings for 1,338 cited papers covering 442 citing papers and impact labels for citations from this dataset. CRISP also offers an efficiency advantage and performs competitively with an open-source model, enabling scalable, cost-effective citation impact analysis (§4.1).

In summary, our contributions are: **(a)** We propose CRISP, a method that *jointly* ranks all cited papers within a citing paper using LLMs to assess their *relative* impact. **(b)** We show that this joint ranking approach outperforms a prior state-of-the-art citation impact classifier. **(c)** We find that CRISP is more efficient and performs competitively with an open-source model. **(d)** We release rankings for 1,338 cited papers covering 442 citing papers, along with impact labels and our code. [1]

2 Notation and Terminology

Citation graph: We consider a universe of papers $\mathcal{P}$, where each element $p \in \mathcal{P}$ corresponds to a unique scholarly document. We model citations as a directed graph over $\mathcal{P}$, where a directed edge $(p \to \tilde{p})$ indicates that paper p cites paper $\tilde{p}$.

Given this convention, we define two neighborhood operators. The outgoing neighborhood of a paper p is *the set of papers cited by* p:

$$N_{\text{out}}(p) \ := \ \{\tilde{p} \in \mathcal{P} \ : \ (p \to \tilde{p})\} \,. \qquad (1)$$

Conversely, the incoming neighborhood of a paper p is *the set of papers that cite* p:

$$N_{\text{in}}(p) \ := \ \{\tilde{p} \in \mathcal{P} \ : \ (\tilde{p} \to p)\} \,. \qquad (2)$$

Citation contexts: We assume that each citation edge $(p \to \tilde{p})$ is accompanied by one or more *citation contexts*, i.e., spans of text in p that refer to $\tilde{p}$. We denote by $\mathrm{Ctx}(p, \tilde{p})$ the set of all such contexts:

$$\mathrm{Ctx}(p \to \tilde{p}) \ := \ \left\{ c_1^{(p \to \tilde{p})}, \ldots, c_k^{(p \to \tilde{p})} \right\}, \qquad (3)$$

where each $c_j^{(p \to \tilde{p})}$ is a sentence- or paragraph-level excerpt from p containing an in-text citation to $\tilde{p}$.

We also define the collection of citation contexts of *all* references made within a citing paper p:

$$\mathrm{Ctx}_{\text{all}}(p) \ := \ \{\mathrm{Ctx}(p \to q) \ : \ q \in N_{\text{out}}(p)\} \,.$$

Intuitively, $\mathrm{Ctx}_{\text{all}}(p)$ provides a paper-level context which we use to provide better calibrated judgments of impact for any particular citation $(p \to q)$.

3 Prior Work and Broader Context

Citation Content Analysis: Early studies conduct *manual* analysis to assess citation quality. For example, Moravcsik and Murugesan (1975) propose a four-dimensional framework assessing citation function and quality. We define citation impact by building upon this framework (refer to §5 for details). However, manual analysis does not scale

[1] Codebase and data will be released.

as the literature continues to grow rapidly. Natural language processing tools have thus been applied to address this challenge. Kinney et al. (2023) release a platform that extracts citation contexts from papers. This extraction enables downstream citation analysis tasks, such as citation intent classification.

Citation Intent Classification: Prior work focuses on classifying citation intents based on their function. Jurgens et al. (2018) and Cohan et al. (2019) classify each citation context $\mathrm{Ctx}(p \to \tilde{p})$ with a single label, while Lauscher et al. (2022) extend this to multiple labels. Similarly, Arnaout et al. (2025) propose a method that generates a sentence describing how a paper p references a paper $\tilde{p}$, considering all the sentences where p cites $\tilde{p}$. They then use this intent with an LLM-judge to predict citation impact. However, these approaches evaluate each citation *independently* based on $\mathrm{Ctx}(p \to \tilde{p})$, without comparing it to other references in the same paper. Our approach addresses this by considering $\mathrm{Ctx}_{\mathrm{all}}(p)$, the full set of citation contexts within a citing paper, which enables calibrated judgments that reflect the relative importance of each reference within its specific citation environment.

Applications of Citation Content Analysis: Identifying impactful citations is an important task, as it can enhance other applications, such as assessing the novelty of scientific ideas (Shahid et al., 2025), improving the retrieval of papers for solving research problems (Garikaparthi et al., 2025), and tracing the sources of key contributions across publications (Zhang et al., 2024). It also enables more nuanced research evaluation by distinguishing substantive intellectual influence from perfunctory citations (Manchanda and Karypis, 2021).

4 Evaluating Relative Impact via CRISP

We present CRISP, our approach for characterizing the relative impact of citations. As illustrated in Figure 2, the method proceeds as follows:

(1) Downstream corpus retrieval: Given a target paper p^*, we use the Semantic Scholar API (Kinney et al., 2023) to retrieve all papers that cite it, denoted $N_{\mathrm{in}}(p^*)$. In Figure 2, $N_{\mathrm{in}}(p^*) = \{q_1, q_2, q_3\}$.

(2) Citation context extraction: For each downstream paper $q \in N_{\mathrm{in}}(p^*)$, we extract its full reference list $N_{\mathrm{out}}(q)$ and their associated citation contexts $\mathrm{Ctx}_{\mathrm{all}}(q)$ via the same API. Figure 2 shows this process for q_1, where $N_{\mathrm{out}}(q_1) = \{p_1, p^*, p_2\}$.

(3) Citation-calibrated impact labeling: We assign an *impact label* to each citation edge $(q \to p)$ for every cited paper $p \in N_{\mathrm{out}}(q)$. Formally, we define an impact labeling function as:

$$f(q \to p) \ := \ f\big(\mathrm{Ctx}(q \to p) | \mathrm{Ctx}_{\mathrm{all}} q)\big) \ \in \ \mathcal{L},$$

where $\mathcal{L}$ is a discrete label space.

In the simplest setting, $\mathcal{L} = \{0, 1\}$ indicates whether the citation $(q \to p)$ reveals impact or not. As shown in Figure 2, we denote our impact labels as $\mathcal{L} = \{0, 1, 2\}$ for low, medium, and high impact.

In practice, f is instantiated by an LLM-based judge (e.g. Vital et al. (2024), Ikoma and Matsubara (2023), Lahiri et al. (2023) and Koloveas et al. (2025)). The key modeling choice in (4) is that the impact label for a citation $(q \to p)$ is predicted *relative to the full citation environment of the citing paper* q. This enables calibration against paper-specific citation conventions, such as whether q tends to cite prior work superficially or relies heavily on a small number of core references.

In our approach, the LLM judge simultaneously ranks q's references by their impact, as shown in Figure 3. Prior work by (Wang et al., 2024; Tang et al., 2024) demonstrates that LLMs exhibit position bias that can affect listwise ranking. We mitigate this bias with the Permutation Self-Consistency (PSC) approach proposed in Tang et al. (2024). We randomize the order of references and perform three independent runs of the classification, each with a different randomized reference order. We then determine the impact category of the target paper p^* through majority voting, selecting the most frequent label across the three runs.

Additionally, we propose an alternative approach to majority voting in which we aggregate the three generated ranking files with Reciprocal Rank Fusion (Cormack et al., 2009) and predict citation impact using an ordinal regression model (Pedregosa-Izquierdo, 2015). Please refer to §C.

4.1 Computational Cost

One might expect that jointly processing all citation contexts within a paper incurs greater computational cost than scoring each citation independently. We argue the opposite is true.

Number of LLM calls: Consider a citation graph with n citing papers and m citation edges. CRISP makes three LLM calls per citing paper, resulting in $O(n)$ calls in total. The UKP approach, by contrast, scores each citation edge independently, requiring $O(m)$ calls. Since each paper typically cites many others, $m \gg n$ in practice, making CRISP asymp-

Model	Method	Acc.	P	R	F1
Baseline	Random	49.9±1.2	33.1	50.9	40.1
GPT-5.1	UKP	66.7±1.1	49.6	63.5	55.7
	CRISP	78.6±1	72.2	63.7	**67.7**
	Δ	+11.9	+22.6	+0.2	+12.0
o4-mini	UKP	63.0±1.1	46.1	73.6	56.7
	CRISP	65.4±1.1	76.3	57.1	65.3
	Δ	+2.4	+30.2	−16.5	+8.6
Qwen3	UKP	60.8±1.2	44.5	76.7	56.3
	CRISP	75.1±1	70.0	53.3	60.5
	Δ	+14.3	+25.5	−23.4	+4.2

Table 1: Comparison of model performance between our approach (CRISP) and prior state-of-the-art method UKP (Arnaout et al., 2025) for identifying impactful citations. Reported accuracies (in percentages) include standard error (± SE). Green cells show positive $\Delta :=$ CRISP− UKP. **Our approach outperforms the prior method across all models as indicated by the positive gains in Δ.** The model with the highest F1 is **bold**.

totically more efficient in terms of LLM calls and thus more scalable for large collections.

Token budget: The total number of tokens processed by each approach is roughly the same because both ultimately consume the same citation contexts. However, CRISP has lower prompt overhead than UKP: because CRISP makes $O(n)$ calls compared to UKP's $O(m)$, the prompt is repeated far fewer times. In practice, this overhead is further reduced by prompt caching, which renders repeated system prompts essentially free.

5 Experimental Setup

Data: Our experiments use a human-labeled dataset of citation contexts, $\mathrm{Ctx}(p, \tilde{p})$, in which each instance is annotated with a binary impact label $\mathcal{L} = \{0, 1\}$ denoting whether the citation is *impact-revealing* or *other* (Arnaout et al., 2025). We augment this dataset with the Semantic Scholar API (Kinney et al., 2023). For every cited paper in the dataset whose title is non-empty, we obtain its Semantic Scholar identifier; for every citing paper p with a valid Semantic Scholar identifier in Arnaout et al. (2025), we retrieve the full reference list and all associated citation contexts $\mathrm{Ctx}_{\text{all}}(p)$. We discard citing papers whose Semantic Scholar API response contains no references, as well as duplicate entries and repeated annotations of the same citation context. After filtering, the dataset comprises 442 citing papers and 1,338 cited papers.

Baseline: We include a random classifier as a baseline for comparison. We also compare against the prior state-of-the-art method for identifying impactful citations proposed by Arnaout et al. (2025).

LLM-based judge: We use an LLM-based judge to assign the impact labels to citation contexts $\mathrm{Ctx}(p, \tilde{p})$ and rank the cited papers by impact for a given citing paper p. Our definition of impact is provided in the LLM prompt 3.

We compare several LLMs of different families and sizes for this task. Specifically, we use GPT-5.1 and o4-mini as our closed-source models, and Qwen3-30B-A3B-Instruct-2507-FP8 (Team, 2025) as our open-source model. We adopt the recommended temperature and top-p settings for each model (0.7 and 0.8 for Qwen; defaults for GPT-5.1 and o4-mini). These models all support large context windows of at least $200K$ tokens.

Evaluation Measures: We evaluate our model using the ground-truth, human-labeled dataset released by Arnaout et al. (2025), which contains impactful and non-impactful citations. For the system of Arnaout et al. (2025), the predicted labels are compared directly with the ground-truth categories of *impact-revealing* and *other*. For our system, we define the label set $\mathcal{L} = \{0, 1, 2\}$, corresponding to low, medium, and high impact. We map low and medium to *other* and high to *impact-revealing*, reducing $\mathcal{L}$ to $\{0, 1\}$ for direct comparison.

We compute model accuracy as the fraction of citations whose predicted labels match the ground truth. We also report precision, recall, and F1 score for the *impact-revealing* class in Table 1.

6 Results and Conclusion

We evaluate on 442 citing papers and 1338 cited papers in Table 1. Our method outperforms prior state-of-the-art impact classifier by $+9.5\%$ in accuracy and $+8.3\%$ in F1 score on average across *all* three models. We also find that Qwen3-30B is a strong alternative to GPT-5.1, offering competitive performance at lower cost. Figure 5 shows the qualitative results and reveals that o4-mini and Qwen3-30B achieve higher recall on UKP only because they tend to favor the impact-revealing class.

In conclusion, we show that considering all references *jointly* within a citing paper provides richer signal than classifying each citation independently. Combined with its efficiency and strong open-source model performance, CRISP enables scalable, cost-effective citation impact analysis.

Limitations

This work analyzes citing papers from psychology, medicine, and computer science. While this captures variations in the number of references in papers and citing practices across different fields, it does not comprehensively represent all disciplines. Additionally, only scientific papers written in English were analyzed.

In practice, as the list of references grows, models struggle to rank the complete list despite having sufficiently large context windows. The number of unranked references differs across models, with GPT-5.1 showing the best performance, suggesting this limitation will diminish as models improve. Further analysis can be found in §B.

It is also worth noting that if the Semantic Scholar API (Kinney et al., 2023) fails to retrieve information for a given paper (some responses for reference details are empty), the subsequent steps of our pipeline cannot be completed.

Finally, an open problem is that authors sometimes cite work differently from what they perceive as impactful, as revealed by interviews in our pilot study. More details can be found in §D.

Ethics Statement

The dataset used in CRISP consists of publicly available scholarly data obtained through the Semantic Scholar API (Kinney et al., 2023), used in accordance with its terms of service. This data contains author names as part of the standard bibliographic record, but no sensitive personal information. Therefore, no additional anonymization was performed.

CRISP is intended for research purposes only. Potential future applications include generating impact reports for authors to review how their work is cited over time. We note, however, that a potential risk of characterizing impact from citation contexts is that it could incentivize authors to strategically frame their citations to influence impact scores.

References

Herman Aguinis, Isabel Suarez-Gonzalez, Gaël Lannelongue, and Hari Joo. 2012. Scholarly impact revisited. *Academy of Management Perspectives*, 26(2):105–132.

Allen Institute for AI. 2025. Asta: A scholarly research assistant. Accessed: 2026-03-14.

Hiba Arnaout, Noy Sternlicht, Tom Hope, and Iryna Gurevych. 2025. In-depth research impact summarization through fine-grained temporal citation analysis. *ArXiv*, abs/2505.14838.

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 3586–3596.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Eugene Garfield. 1972. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, 178(4060):471–479.

Aniketh Garikaparthi, Manasi Patwardhan, Aditya Sanjiv Kanade, Aman Hassan, Lovekesh Vig, and Arman Cohan. 2025. Mir: Methodology inspiration retrieval for scientific research problems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28614–28659.

Saeed-Ul Hassan, Anam Akram, and Peter Haddawy. 2017. Identifying important citations using contextual information from full text. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–8.

Tomoki Ikoma and Shigeki Matsubara. 2023. On the use of language models for function identification of citations in scholarly papers. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 130–135, Bali, Indonesia. Association for Computational Linguistics.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, and 29 others. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Paris Koloveas, Serafeim Chatzopoulos, Thanasis Vergoulis, and Christos Tryfonopoulos. 2025. Can llms predict citation intent? an experimental analysis of in-context learning and fine-tuning on open llms. In

International Conference on Theory and Practice of Digital Libraries, pages 207–224. Springer.

Avishek Lahiri, Debarshi Kumar Sanyal, and Imon Mukherjee. 2023. Citeprompt: Using prompts to identify citation intent in scientific papers. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–55. IEEE.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1875–1889.

Saurav Manchanda and George Karypis. 2021. Evaluating scholarly impact: Towards content-aware bibliometrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6041–6053, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael J Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social studies of science*, 5(1):86–92.

Fabian Pedregosa-Izquierdo. 2015. *Feature extraction and supervised learning on fMRI: from practice to theory*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. 2025. Literature-grounded novelty assessment of scientific ideas. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 96–113.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Türe. 2024. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: human language technologies (volume 1: long papers)*, pages 2327–2340.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI workshop: Scholarly big data*, volume 15, page 13.

Adilson Vital, Filipi N. Silva, Osvaldo N. Oliveira, and Diego Raphael Amancio. 2024. Predicting citation impact of research papers using gpt and other text embeddings. *ArXiv*, abs/2407.19942.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, and Jie Tang. 2024. Pst-bench: Tracing and benchmarking the source of publications. *arXiv preprint arXiv:2402.16009*.

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427.

A Experiment Details

Figure 3 shows our definition of citation impact and the instruction used to rank a citing paper's references, with a sample output in Figure 4. Figure 8 displays the aggregated rankings from three independent runs using GPT-5.1. We use these aggregated ranking files to predict impact labels via an ordinal regression model (§C).

B Analysis of Missing References

As noted in §5, we rank references by their impact on the citing paper. We find that as the number of references grows, models struggle to rank the complete list despite having large context windows of at least 200k tokens. Figure 6 shows this effect.

When the citing paper contains fewer than 40 references, all three models successfully rank nearly every reference. However, as the reference list grows, the number of missing references increases sharply. GPT-5.1 and Qwen3-30B degrade more gracefully, with GPT-5.1 omitting roughly 70 references on average for papers with 200–240 references. In contrast, o4-mini exhibits the steepest decline, failing to rank over 170 references on average in the same range.

These results suggest that, although the models can technically ingest long contexts, they struggle to rank all items as the reference list grows.

C Alternative Impact Label Assignment Approach

While majority voting performs strongly as an aggregation technique for impact classification (§4), a limitation of this approach is that the resulting labels are not guaranteed to decrease monotonically with rank. For example, the paper ranked fifth could receive a *Low* label, while the paper ranked sixth could receive a *Medium* label. This violates the assumption that impact categories follow the order *High* > *Medium* > *Low*.

To address cases where an aggregated ranking with references sorted by impact is needed, we propose an alternative method based on ordinal regression that ensures labels respect this ordering.

Aggregating Ranks:

We aggregate the three rankings generated by the LLM judge for each citing paper using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). Following Cormack et al. (2009), we compute the RRF score as

$$\text{RRF-score}(p) = \sum_{i=1}^{N} \frac{1}{k + \text{rank}_i(p)},$$

where p denotes a cited paper, $N = 3$ since each citing paper produces three rankings, k is a constant (we set $k = 60$), and $\text{rank}_i(p)$ is the rank of paper p in the i-th ranking.

As noted in §B, some references do not appear in all ranked lists. For such cases, we compute the mean of the non-empty values among $\text{rank}_1(p)$, $\text{rank}_2(p)$, and $\text{rank}_3(p)$, and use this value to impute the missing ranks. If a paper does not appear in any of the three rankings, it is excluded from the final aggregated list. In addition, hallucinated references that are not present in the original reference list are discarded during aggregation. Furthermore, if a cited paper appears multiple times within the same ranking, we retain only its lowest (best) rank. For example, if a paper p is ranked both 1 and 3 in the same ranking r_i, we use 1 as $\text{rank}_i(p)$.

We release the dataset of aggregated rankings produced by the LLM judges using the models GPT-5.1, o4-mini, and Qwen3-30B-A3B-Instruct-2507-FP8 (Team, 2025). Specifically, the dataset contains the aggregated rankings for each of the 442 citing papers per model we generated in our experiments (§5), for a total of 1,326 files. A sample aggregated ranking is shown in Figure 8.

Predicting Impact Labels with an Ordinal Regression Model:

Training Data Construction. For each of the 442 citing papers, we collect the ranked reference lists produced by the three independent LLM runs described in §4. For each cited reference, we construct a feature vector consisting of the raw ranks from each run (r_1, r_2, r_3), the normalized ranks $(r_i/N$, where N is the total number of references in the citing paper), the standard deviation across the three runs, and the mean rank. If a reference is missing a rank from one or two runs, we impute the missing value(s) with the median of the available ranks. Normalizing by the reference list length ensures comparability across citing papers with different numbers of references. The impact label (*Low*, *Medium*, or *High*) for the training data is determined by majority vote over the three runs' impact category assignments.

Citation Impact Classifier	Model	Acc.	P	R	F1
Ours (Ord. Reg.)	GPT-5.1	**78.9**±1	73.6	62.7	**67.7**
	o4-mini	64.3±1.1	76.4	54.5	62.2
	Qwen3 30B	75.9±1	75.6	48.6	59.1

Table 2: Results using an ordinal regression model to predict impact labels. Reported accuracies include the standard error (± SE).

To prevent data leakage, we exclude all (citing paper, cited paper) pairs that appear in our held-out test set prior to model fitting. The held-out test set consists of the 1,338 cited papers for which we have ground-truth impact annotations from the dataset released by Arnaout et al. (2025). Therefore, the training data comprises only the remaining references that were ranked alongside the test pairs during the per-paper ranking step but do not themselves appear in the evaluation set. Importantly, the normalization factor N is computed from the full reference list before exclusion, preserving the original scale of the citing paper.

Model Training. We train a single global ordinal regression model pooled across all citing papers, rather than fitting per-paper models, to maximize training signal. We use the Immediate-Threshold variant of ordinal logistic regression (Pedregosa-Izquierdo, 2015) from the mord library, which respects the natural ordering of the three impact categories through shared threshold parameters. The model is regularized with an L2 penalty ($\alpha = 1.0$).

Prediction. At inference time, we apply the trained global model to the aggregated ranking produced by RRF (§C) for each citing paper. For every reference in the aggregated list, we construct the same feature vector used during training from its per-run ranks and predict an impact category (*Low*, *Medium*, or *High*). The predicted label is then assigned to the corresponding entry in the RRF-aggregated ranking, yielding a final output in which each cited reference has both an aggregated rank and a predicted impact category. A sample output is shown in Figure 8.

Results. Table 2 reports the performance of the ordinal regression model. These results are evaluating the same 1,338 cited papers described in §5. Our method outperforms the approach of Arnaout et al. (2025) across all models. Compared to our majority-vote approach (Table 1), ordinal regression yields similar overall performance: GPT-5.1 improves slightly in accuracy (78.9 vs. 78.6) while maintaining the same F1 of 67.7. For o4-mini and Qwen3-30B, the model trades recall for higher precision, resulting in slightly lower F1 scores. These results suggest that ordinal regression provides a lightweight post-processing refinement over majority vote without requiring additional LLM calls, with the largest benefit when the underlying LLM judge is already strong. The qualitative results are shown in Figure 7.

D Pilot Study

Prior to running the experiments described in §5, we conducted a pilot study to evaluate the effectiveness of the prompt shown in Figure 3 for ranking references by their impact on their citing papers. Using a custom annotation interface we designed, six annotators ranked references from a paper they co-authored according to our predefined impact criteria. The annotators were PhD students from our lab who volunteered to participate; no monetary compensation was offered and all names are anonymized.

Figures 9–12 illustrate the custom annotation interface we designed for the study. Figure 9 shows the main annotation task layout. Figure 10 displays the complete list of references for an annotator's paper. Figure 11 demonstrates the feature we added to reveal cited papers' citation context. Lastly, Figure 12 presents the final ranked list generated within the interface. As shown in Figure 12, the impact definition for each category was available to the annotators throughout the task, and citations were color-coded according to their corresponding impact category.

We compute the Spearman rank correlation between each annotator's ranking and the ranking generated by GPT-5.1 using the same prompt from our experiments (Figure 3). As shown in Figure 13, all correlation values exceeded 0.7, indicating strong agreement. These results suggest that our prompt produces rankings that closely align with how authors would rank their citations, though we defer a more comprehensive evaluation to future work.

E Disclosure on the Use of Generative Assistants

The authors adhered to ACL's guidelines for appropriate use of generative assistance in authorship. In particular, we used generative assistants to polish our original writing. Additionally, we used LLM-powered tools, such as (Allen Institute for AI, 2025), for literature search.

Impact label assignment to every cited paper in $\mathbf{N_{out}}$(p)

The task is to rank all of the references R (where R is a list of papers r_1, r_2, ...) of a given paper P based on how impactful and influential each r_n was on P.

The paper P you will be analyzing is:
Title: {main_paper_title}
Abstract: {main_paper_abstract}

The list of references is given below (each reference includes paper ID, title, and context of citation): {references_text}

Impact categories:

1. High-impact citations: These are the papers without which your own work would not have been possible. They supply essential conceptual, methodological, or operational ingredients.
- Conceptual or operational indispensability: The reference provides a unique conceptual insight, methodological innovation, dataset, or technique that is directly instrumental to your paper. Examples: a specific algorithm your method extends; a benchmark or dataset your study critically depends on; a theoretical formulation your contribution builds on.
- Organic necessity: The reference is uniquely and genuinely required for a reader to understand how your paper works or how its core logic unfolds. Without this citation, the intellectual lineage of your method would be opaque or incomplete.
- Typical quantity: 1–5 papers (or even 1).

2. Medium-impact citations: These are papers that helped you write your paper, but were not fundamentally irreplaceable. You could have used an alternative prior work or formulation, but you chose this one because it was particularly useful, clear, or canonical.
- Conceptual or operational contribution (non-unique): The reference conveys an idea, dataset, or model family that meaningfully helped your setup, but other comparable alternatives exist. Examples: selecting LLaMA-1 vs LLaMA-2; choosing one evaluation protocol among several similar ones; relying on one of several formulations of a known concept.
- Organic helpfulness: The reference is genuinely helpful for understanding your paper, but not uniquely necessary. It situates your work clearly, but your contribution does not hinge on this specific citation.
- Typical quantity: roughly 5–15 papers.

3. Low-impact citations: These citations provide background, context, or perfunctory acknowledgement, but the core contribution of your paper is not dependent on them in any strong way.
- Background or definitional citations: References used to define a task (e.g., Question Answering), introduce a general problem area, or acknowledge standard terminology. The same role could have been fulfilled by many other papers.
- Perfunctory or field-signaling citations: The reference mainly signals that prior work exists in the broad area. The citing paper does not substantively depend on the specific ideas of the cited work.
- Typical quantity: the majority of citations.

Please output the ranked references as a **JSON array**, where each entry has:
- "rank": integer
- "paperId": string
- "title": string
- "contexts": string (all citation contexts)
- "reason": string (why this rank was assigned)
- "category": string ("High", "Medium", "Low")

Return **valid JSON only**, without any extra text. DO NOT wrap the output in ``` or ```json.
DO NOT include any explanation, commentary, or text outside the JSON array.

Figure 3: Shows the prompt used in our system to define impact and rank cited papers based on their impact category.

```
[
  {
    "rank": 1,
    "paperId": "5507d267bbf0b4cdb9f893c3c0960a45016f7010",
    "title": "Deep Leakage from Gradients",
    "contexts": "For DLG [1], as described by the authors, we start the procedure with the randomly initialized dummy
          data and outputs ( x (cid:48), y (cid:48) ), then iteratively update them to minimize the gradient matching
          objective. | Recent work by Zhu et al. [1] presents an approach (DLG) to steal the proprietary data protected
          by the participants in distributed learning from the shared gradients. | In this section, we empirically
          demonstrate the advantages of our (iDLG) method over DLG [1]. | However, recent work by Zhu et al., \"Deep
          Leakage from Gradient\" (DLG) [1] showed the possibility to steal the private training data from the shared
          gradients of other participants. | ...for 300 iterations, and evaluate the performance in terms of (i) the
          accuracy of the extracted labels c (cid:48), and (ii) the fidelity of the extracted Dataset DLG iDLG MNIST
          89.9% 100.0% CIFAR-100 83.3% 100.0% LFW 79.1% 100.0% Table 1: Accuracy of the extracted labels for DLG [1] and
           iDLG. | Following the settings in [1], we use the randomly initialized LeNet for all experiments. | - We
          empirically demonstrate the advantages of iDLG over DLG [1] via comparing the accuracy of extracted labels and
           the fidelity of extracted data on three datasets. | This enables us to always extract the ground-truth labels
           and significantly simplify the objective of DLG [1] in order to extract good-quality data.",
    "reason": "The entire contribution of iDLG is explicitly positioned as an improvement over DLG. The problem setting,
           core optimization objective (gradient matching), baseline method, experimental protocol (including LeNet and
          evaluation metrics), and even the paper's name are derived from or defined relative to this work. The logic of
           iDLG--both conceptually (what it improves) and empirically (what it compares against)--is unintelligible
          without DLG. Thus it is conceptually and operationally indispensable, making it a high-impact citation.",
    "impactCategory": "High"
  },
  {
    "rank": 2,
    "paperId": "6a6ad9eb495739f4c80e7c09598720c3d5c5dff7",
    "title": "Federated Learning: Collaborative Machine Learning without\nCentralized Training Data",
    "contexts": "In multi-node distributed learning systems such as Collaborative Learning [2, 3, 4] and Federated
          Learning [5, 6, 7], it is widely believed that sharing gradients between nodes will not leak the private
          training data.",
    "reason": "This is a canonical paper defining the federated learning paradigm, which provides the main application
          context where gradient sharing occurs. iDLG's motivation--privacy leakage in distributed / federated setups--
          relies on this paradigm. However, any of several federated learning introductions could have served a similar
          role; the iDLG method itself does not technically depend on this specific paper. Hence it is important context
           but replaceable, so medium impact.",
    "impactCategory": "Medium"
  },
  {
    "rank": 3,
    "paperId": "7fcb90f68529cbfab49f471b54719ded7528d0ef",
    "title": "Federated Learning: Strategies for Improving Communication Efficiency",
    "contexts": "In multi-node distributed learning systems such as Collaborative Learning [2, 3, 4] and Federated
          Learning [5, 6, 7], it is widely believed that sharing gradients between nodes will not leak the private
          training data.",
    "reason": "This work further characterizes federated learning systems and communication strategies, reinforcing the
          setting where gradients are shared. It supports the real-world relevance of the attack scenario but is not
          directly used in the method or experiments. Other federated learning references could substitute it, so it is
          context-setting and thus medium rather than high impact.",
    "impactCategory": "Medium"
  },
% omitted entries for brevity
  {
    "rank": 7,
    "paperId": "f2f8f7a2ec1b2ede48cbcd189b376ab9fa0735ef",
    "title": "Privacy-preserving deep learning",
    "contexts": "In multi-node distributed learning systems such as Collaborative Learning [2, 3, 4] and Federated
          Learning [5, 6, 7], it is widely believed that sharing gradients between nodes will not leak the private
          training data.",
    "reason": "This paper offers an early privacy-preserving framework for deep learning and is cited to indicate that
          prior work assumed gradient/parameter sharing can preserve privacy. It supports the contrast between perceived
           and actual privacy guarantees that iDLG exposes. Nonetheless, the iDLG algorithm does not depend on any of
          its mechanisms or theory; the citation is mainly motivational and background, so medium impact.",
    "impactCategory": "Medium"
  },
  {
    "rank": 8,
    "paperId": "5d90f06bb70a0a3dced62413346235c02b1aa086",
    "title": "Learning Multiple Layers of Features from Tiny Images",
    "contexts": "We perform experiments on the classification task over three datasets: MNIST [8], CIFAR-100 [9], and
          LFW [10] with 10, 100, and 5749 categories respectively.",
    "reason": "This technical report defines the CIFAR-100 dataset, which is one of the main benchmarks used to evaluate
           iDLG's label extraction and data reconstruction performance. The choice of CIFAR-100 helps demonstrate
          scalability across many classes and more complex images, but in principle another comparable dataset could
          have been used. It operationally supports experiments but is not uniquely necessary, making it a low-to-medium
           impact citation; given the categories, it best fits low impact as a standard dataset reference.",
    "impactCategory": "Low"
  },
% omitted entries for brevity
{
    "rank": 11,
    "paperId": "1267fe36b5ece49a9d8f913eb67716a040bbcced",
    "title": "On the limited memory BFGS method for large scale optimization",
    "contexts": "L-BFGS [11] with learning rate 1 is used as the optimizer.",
    "reason": "This optimization paper is referenced to justify the use of L-BFGS for matching gradients when
          reconstructing data and labels. While L-BFGS may influence convergence behavior in practice, the conceptual
          contribution of iDLG--the analytic label recovery from gradients--does not depend on this specific optimizer.
          Many gradient-based optimizers could fill this role with similar effect. Thus it is an implementation-level,
          replaceable choice and so low impact.",
    "impactCategory": "Low"
  }
]
```

Figure 4: Shows a sample output using Prompt 3 with GPT-5.1. This is the impact label assignment and ranked references of the citing paper *iDLG: Improved Deep Leakage from Gradients.*

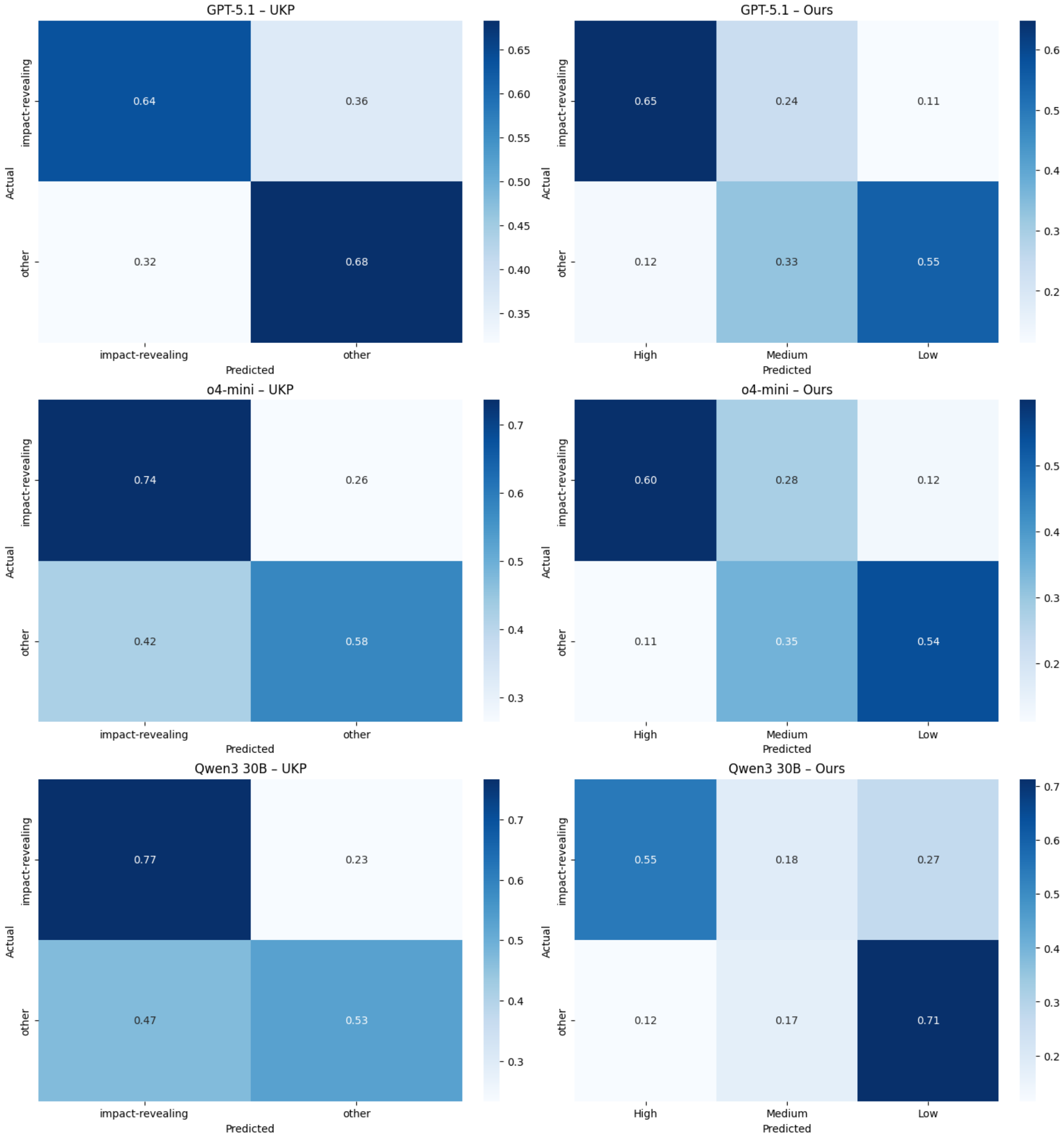


Figure 5: Confusion matrices per model for assigning impact labels through majority vote. Rows represent actual labels; columns represent predicted categories. To compare both approaches (collapsing High to impact-revealing and Medium/Low to other, as described in 5), we find that our method produces more discriminative boundaries, substantially reducing false positives across all three models. Notably, GPT-5.1 achieves this while maintaining comparable recall (0.65 vs. 0.64), demonstrating that our method improves precision without sacrificing sensitivity.

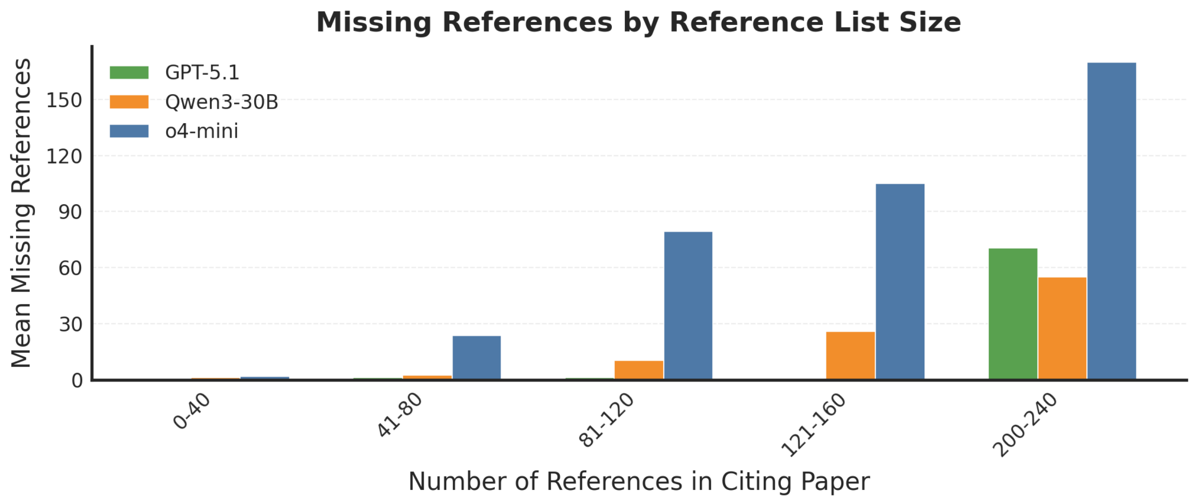


Figure 6: Mean number of missing references in the aggregated ranks per model as a function of the number of references in the citing paper. All three models rank nearly all references when the reference list is less than 40. However, they omit more references as the list grows longer. o4-mini exhibits the steepest degradation, while GPT-5.1 and Qwen3-30B are more robust.

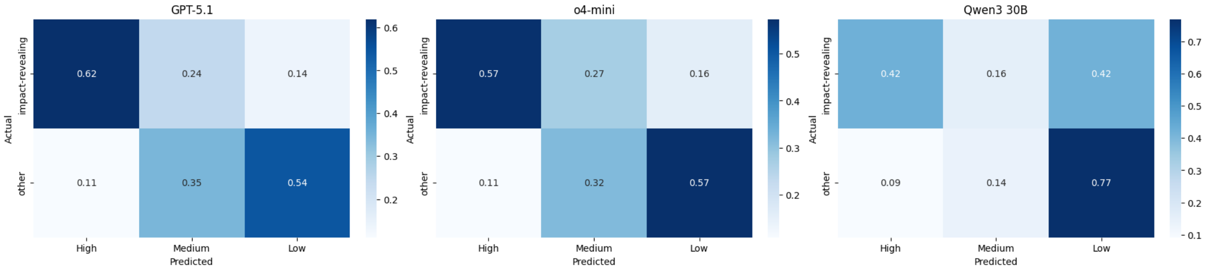
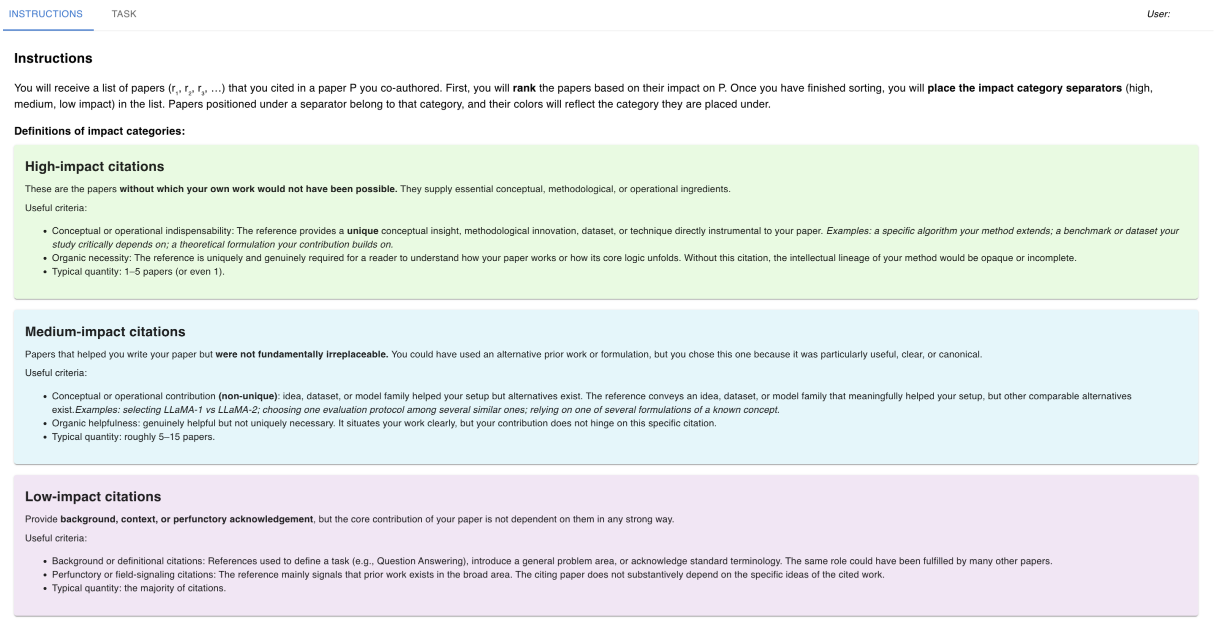


Figure 7: Confusion matrices per model for assigning impact labels with an ordinal regression model. Rows represent actual labels; columns represent predicted categories. GPT-5.1 achieves the best performance in identifying impactful citations and reducing false positive. However, we see a slight decrease in performance with Qwen3-30b.

```
[
  {
    "rank": 1,
    "paperId": "5507d267bbf0b4cdb9f893c3c0960a45016f7010",
    "title": "Deep Leakage from Gradients",
    "rrf_score": 0.04918032786885246,
    "num_rankings_found": 3,
    "predicted_impact": "High"
  },
  {
    "rank": 2,
    "paperId": "6a6ad9eb495739f4c80e7c09598720c3d5c5dff7",
    "title": "Federated Learning: Collaborative Machine Learning without\nCentralized Training Data",
    "rrf_score": 0.04788306451612903,
    "num_rankings_found": 3,
    "predicted_impact": "Medium"
  },
  {
    "rank": 3,
    "paperId": "7fcb90f68529cbfab49f471b54719ded7528d0ef",
    "title": "Federated Learning: Strategies for Improving Communication Efficiency",
    "rrf_score": 0.047619047619047616,
    "num_rankings_found": 3,
    "predicted_impact": "Medium"
  },
  {
    "rank": 4,
    "paperId": "8a564ee07fa930ebc1176019deacdc9951063a99",
    "title": "Collaborative Learning for Deep Neural Networks",
    "rrf_score": 0.046153846153846156,
    "num_rankings_found": 3,
    "predicted_impact": "Medium"
  },
  {
    "rank": 5,
    "paperId": "49bdeb07b045dd77f0bfe2b44436608770235a23",
    "title": "Federated Learning: Challenges, Methods, and Future Directions",
    "rrf_score": 0.04595588235294118,
    "num_rankings_found": 3,
    "predicted_impact": "Medium"
  },
  {
    "rank": 6,
    "paperId": "8bdf6f03bde08c424c214188b35be8b2dec7cdea",
    "title": "Inference Attacks Against Collaborative Learning",
    "rrf_score": 0.045228403437358664,
    "num_rankings_found": 3,
    "predicted_impact": "Medium"
  },
  {
    "rank": 7,
    "paperId": "f2f8f7a2ec1b2ede48cbcd189b376ab9fa0735ef",
    "title": "Privacy-preserving deep learning",
    "rrf_score": 0.04500226142017187,
    "num_rankings_found": 3,
    "predicted_impact": "Medium"
  },
  {
    "rank": 8,
    "paperId": "1267fe36b5ece49a9d8f913eb67716a040bbcced",
    "title": "On the limited memory BFGS method for large scale optimization",
    "rrf_score": 0.04429804634257156,
    "num_rankings_found": 3,
    "predicted_impact": "Low"
  },
  {
    "rank": 9,
    "paperId": "5d90f06bb70a0a3dced62413346235c02b1aa086",
    "title": "Learning Multiple Layers of Features from Tiny Images",
    "rrf_score": 0.04390451832907076,
    "num_rankings_found": 3,
    "predicted_impact": "Low"
  },
  {
    "rank": 10,
    "paperId": "c6b3ca4f939e36a9679a70e14ce8b1bbbc5618f3",
    "title": "Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments",
    "rrf_score": 0.0432712215320911,
    "num_rankings_found": 3,
    "predicted_impact": "Low"
  },
  {
    "rank": 11,
    "paperId": "162d958ff885f1462aeda91cd72582323fd6a1f4",
    "title": "Gradient-based learning applied to document recognition",
    "rrf_score": 0.04265593561368209,
    "num_rankings_found": 3,
    "predicted_impact": "Low"
  }
]
```

Figure 8: Shows a sample output of the ordinal regression model predicting impact labels for the ranked references of the citing paper *iDLG: Improved Deep Leakage from Gradients*.



Instructions

You will receive a list of papers (r_1, r_2, r_3, …) that you cited in a paper P you co-authored. First, you will **rank** the papers based on their impact on P. Once you have finished sorting, you will **place the impact category separators** (high, medium, low impact) in the list. Papers positioned under a separator belong to that category, and their colors will reflect the category they are placed under.

Definitions of impact categories:

High-impact citations

These are the papers **without which your own work would not have been possible.** They supply essential conceptual, methodological, or operational ingredients.

Useful criteria:

- Conceptual or operational indispensability: The reference provides a **unique** conceptual insight, methodological innovation, dataset, or technique directly instrumental to your paper. *Examples: a specific algorithm your method extends; a benchmark or dataset your study critically depends on; a theoretical formulation your contribution builds on.*
- Organic necessity: The reference is uniquely and genuinely required for a reader to understand how your paper works or how its core logic unfolds. Without this citation, the intellectual lineage of your method would be opaque or incomplete.
- Typical quantity: 1–5 papers (or even 1).

Medium-impact citations

Papers that helped you write your paper but **were not fundamentally irreplaceable.** You could have used an alternative prior work or formulation, but you chose this one because it was particularly useful, clear, or canonical.

Useful criteria:

- Conceptual or operational contribution **(non-unique)**: idea, dataset, or model family helped your setup but alternatives exist. The reference conveys an idea, dataset, or model family that meaningfully helped your setup, but other comparable alternatives exist.*Examples: selecting LLaMA-1 vs LLaMA-2; choosing one evaluation protocol among several similar ones; relying on one of several formulations of a known concept.*
- Organic helpfulness: genuinely helpful but not uniquely necessary. It situates your work clearly, but your contribution does not hinge on this specific citation.
- Typical quantity: roughly 5–15 papers.

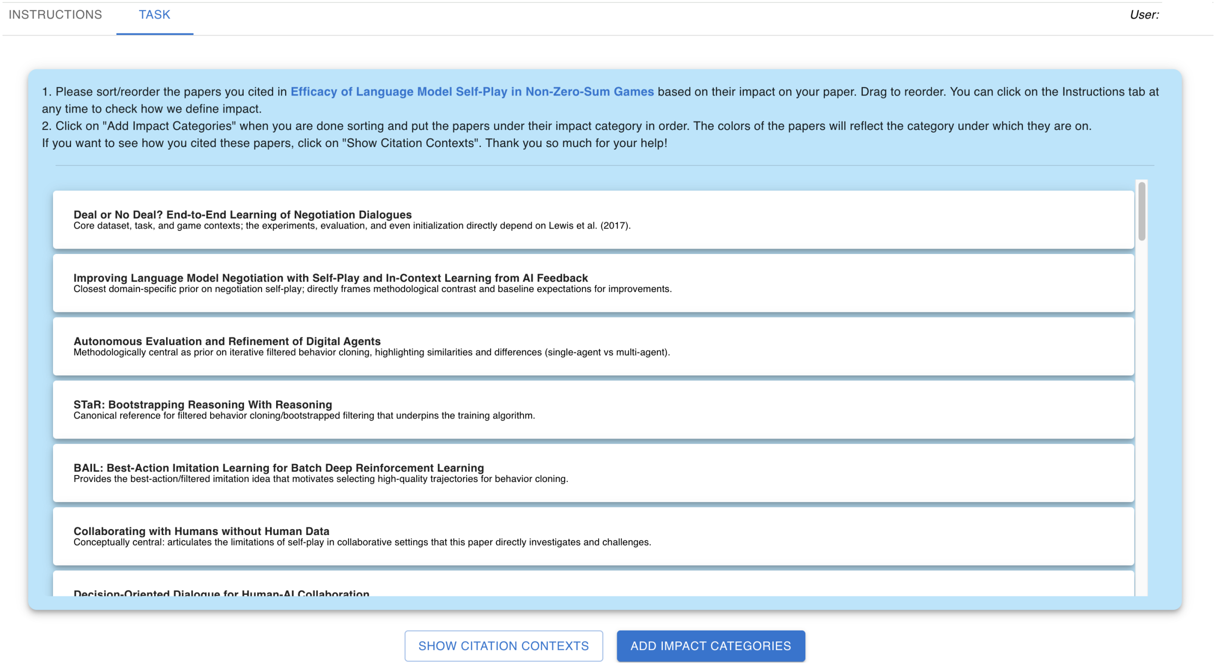
Low-impact citations

Provide **background, context, or perfunctory acknowledgement**, but the core contribution of your paper is not dependent on them in any strong way.

Useful criteria:

- Background or definitional citations: References used to define a task (e.g., Question Answering), introduce a general problem area, or acknowledge standard terminology. The same role could have been fulfilled by many other papers.
- Perfunctory or field-signaling citations: The reference mainly signals that prior work exists in the broad area. The citing paper does not substantively depend on the specific ideas of the cited work.
- Typical quantity: the majority of citations.

Figure 9: The main layout of the custom annotation interface used in the pilot study, showing the overall task setup and instructions for ranking references by impact.



1. Please sort/reorder the papers you cited in Efficacy of Language Model Self-Play in Non-Zero-Sum Games based on their impact on your paper. Drag to reorder. You can click on the Instructions tab at any time to check how we define impact.
2. Click on "Add Impact Categories" when you are done sorting and put the papers under their impact category in order. The colors of the papers will reflect the category under which they are on.

If you want to see how you cited these papers, click on "Show Citation Contexts". Thank you so much for your help!

Deal or No Deal? End-to-End Learning of Negotiation Dialogues
Core dataset, task, and game contexts; the experiments, evaluation, and even initialization directly depend on Lewis et al. (2017).

Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback
Closest domain-specific prior on negotiation self-play; directly frames methodological contrast and baseline expectations for improvements.

Autonomous Evaluation and Refinement of Digital Agents
Methodologically central as prior on iterative filtered behavior cloning, highlighting similarities and differences (single-agent vs multi-agent).

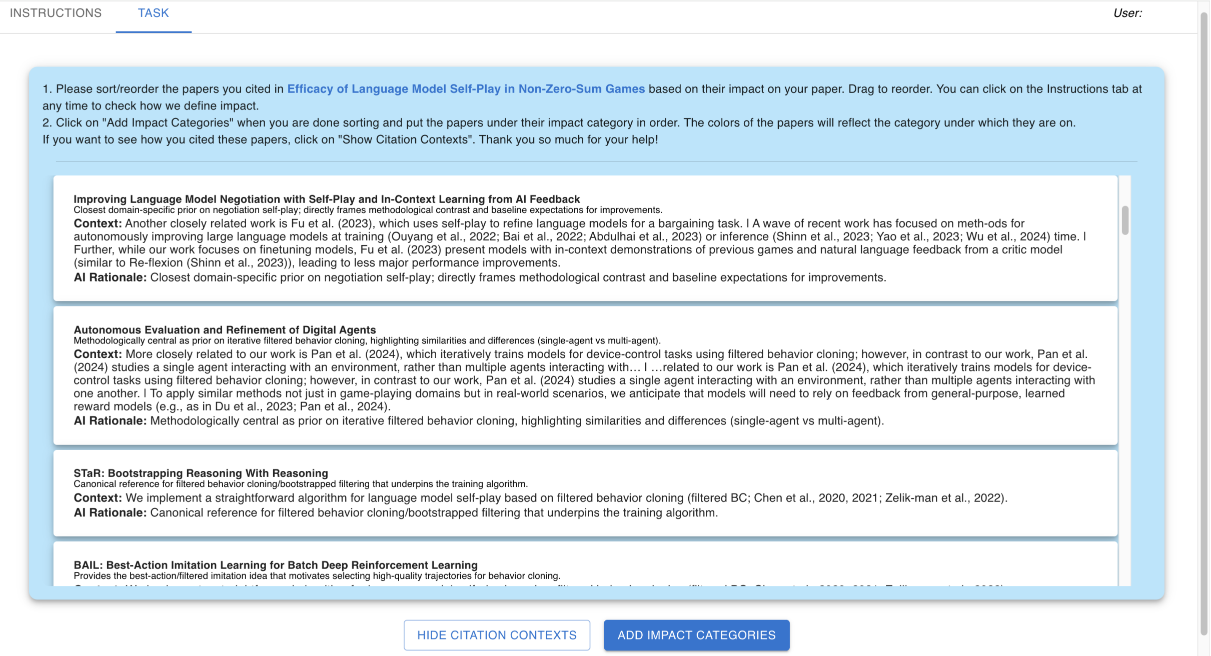
STaR: Bootstrapping Reasoning With Reasoning
Canonical reference for filtered behavior cloning/bootstrapped filtering that underpins the training algorithm.

BAIL: Best-Action Imitation Learning for Batch Deep Reinforcement Learning
Provides the best-action/filtered imitation idea that motivates selecting high-quality trajectories for behavior cloning.

Collaborating with Humans without Human Data
Conceptually central: articulates the limitations of self-play in collaborative settings that this paper directly investigates and challenges.

Decision-Oriented Dialogue for Human-AI Collaboration

Figure 10: The full list of references from an annotator’s paper displayed within the interface, allowing users to rank their references by dragging and reordering them according to the impact criteria.



1. Please sort/reorder the papers you cited in **Efficacy of Language Model Self-Play in Non-Zero-Sum Games** based on their impact on your paper. Drag to reorder. You can click on the Instructions tab at any time to check how we define impact.
2. Click on "Add Impact Categories" when you are done sorting and put the papers under their impact category in order. The colors of the papers will reflect the category under which they are on.
If you want to see how you cited these papers, click on "Show Citation Contexts". Thank you so much for your help!

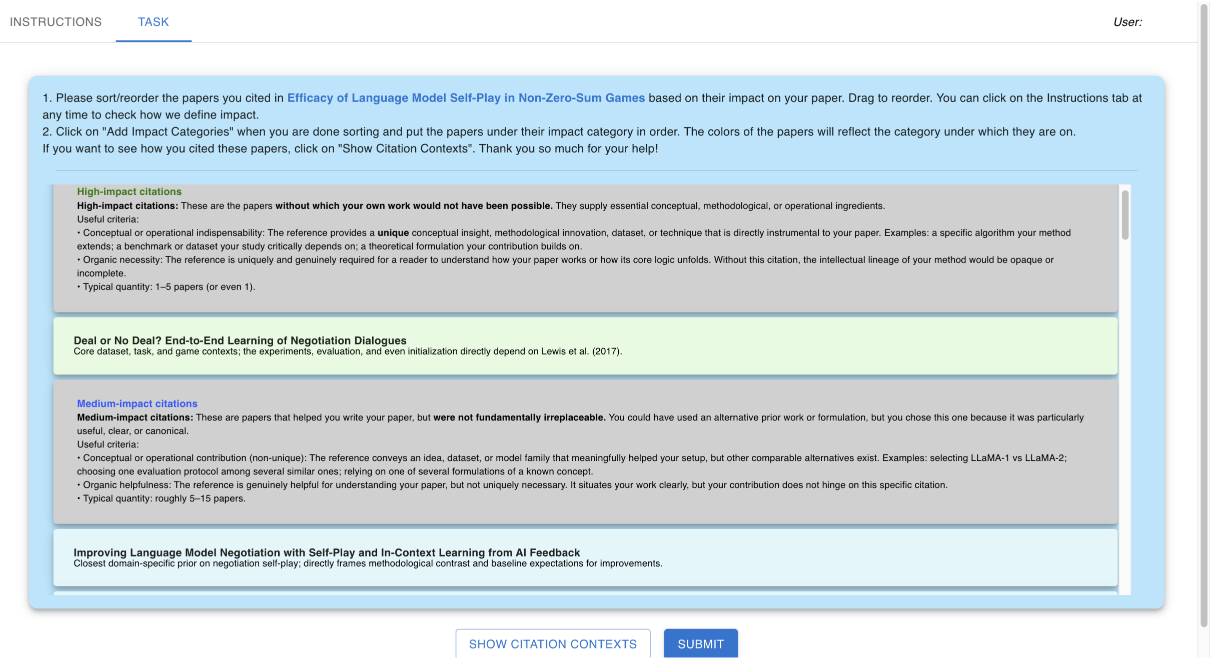
Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback
Closest domain-specific prior on negotiation self-play; directly frames methodological contrast and baseline expectations for improvements.
Context: Another closely related work is Fu et al. (2023), which uses self-play to refine language models for a bargaining task. | A wave of recent work has focused on meth-ods for autonomously improving large language models at training (Ouyang et al., 2022; Bai et al., 2022; Abdulhai et al., 2023) or inference (Shinn et al., 2023; Yao et al., 2023; Wu et al., 2024) time. | Further, while our work focuses on finetuning models, Fu et al. (2023) present models with in-context demonstrations of previous games and natural language feedback from a critic model (similar to Re-flexion (Shinn et al., 2023)), leading to less major performance improvements.
AI Rationale: Closest domain-specific prior on negotiation self-play; directly frames methodological contrast and baseline expectations for improvements.

Autonomous Evaluation and Refinement of Digital Agents
Methodologically central as prior on iterative filtered behavior cloning, highlighting similarities and differences (single-agent vs multi-agent).
Context: More closely related to our work is Pan et al. (2024), which iteratively trains models for device-control tasks using filtered behavior cloning; however, in contrast to our work, Pan et al. (2024) studies a single agent interacting with an environment, rather than multiple agents interacting with… | …related to our work is Pan et al. (2024), which iteratively trains models for device-control tasks using filtered behavior cloning; however, in contrast to our work, Pan et al. (2024) studies a single agent interacting with an environment, rather than multiple agents interacting with one another. | To apply similar methods not just in game-playing domains but in real-world scenarios, we anticipate that models will need to rely on feedback from general-purpose, learned reward models (e.g., as in Du et al., 2023; Pan et al., 2024).
AI Rationale: Methodologically central as prior on iterative filtered behavior cloning, highlighting similarities and differences (single-agent vs multi-agent).

STaR: Bootstrapping Reasoning With Reasoning
Canonical reference for filtered behavior cloning/bootstrapped filtering that underpins the training algorithm.
Context: We implement a straightforward algorithm for language model self-play based on filtered behavior cloning (filtered BC; Chen et al., 2020, 2021; Zelik-man et al., 2022).
AI Rationale: Canonical reference for filtered behavior cloning/bootstrapped filtering that underpins the training algorithm.

BAIL: Best-Action Imitation Learning for Batch Deep Reinforcement Learning
Provides the best-action/filtered imitation idea that motivates selecting high-quality trajectories for behavior cloning.

Figure 11: A feature of the interface that reveals the citation context when the "Show Citation Contexts" button is clicked, illustrating how each reference was cited in the paper.



1. Please sort/reorder the papers you cited in **Efficacy of Language Model Self-Play in Non-Zero-Sum Games** based on their impact on your paper. Drag to reorder. You can click on the Instructions tab at any time to check how we define impact.
2. Click on "Add Impact Categories" when you are done sorting and put the papers under their impact category in order. The colors of the papers will reflect the category under which they are on.
If you want to see how you cited these papers, click on "Show Citation Contexts". Thank you so much for your help!

High-impact citations
High-impact citations: These are the papers **without which your own work would not have been possible.** They supply essential conceptual, methodological, or operational ingredients.
Useful criteria:
- Conceptual or operational indispensability: The reference provides a **unique** conceptual insight, methodological innovation, dataset, or technique that is directly instrumental to your paper. Examples: a specific algorithm your method extends; a benchmark or dataset your study critically depends on; a theoretical formulation your contribution builds on.
- Organic necessity: The reference is uniquely and genuinely required for a reader to understand how your paper works or how its core logic unfolds. Without this citation, the intellectual lineage of your method would be opaque or incomplete.
- Typical quantity: 1–5 papers (or even 1).

Deal or No Deal? End-to-End Learning of Negotiation Dialogues
Core dataset, task, and game contexts; the experiments, evaluation, and even initialization directly depend on Lewis et al. (2017).

Medium-impact citations
Medium-impact citations: These are papers that helped you write your paper, but **were not fundamentally irreplaceable.** You could have used an alternative prior work or formulation, but you chose this one because it was particularly useful, clear, or canonical.
Useful criteria:
- Conceptual or operational contribution (non-unique): The reference conveys an idea, dataset, or model family that meaningfully helped your setup, but other comparable alternatives exist. Examples: selecting LLaMA-1 vs LLaMA-2; choosing one evaluation protocol among several similar ones; relying on one of several formulations of a known concept.
- Organic helpfulness: The reference is genuinely helpful for understanding your paper, but not uniquely necessary. It situates your work clearly, but your contribution does not hinge on this specific citation.
- Typical quantity: roughly 5–15 papers.

Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback
Closest domain-specific prior on negotiation self-play; directly frames methodological contrast and baseline expectations for improvements.

Figure 12: The final ranked list of references generated within the interface, showing the outcome of the annotation task with color-coded impact categories. Annotators submit their final ranked list of references.

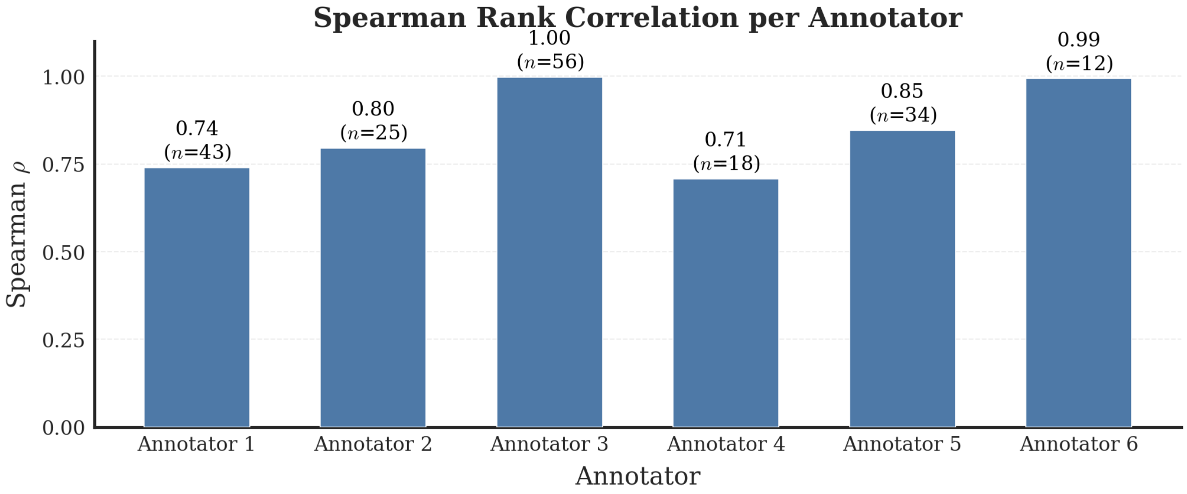


Figure 13: Shows the Spearman correlation between each annotator's ranking and the ranking generated by GPT-5.1 using the same prompt from our experiments. The n value above each bar represents the total number of references ranked in that paper. All correlation values exceeded 0.7, indicating strong agreement.