Crystal: Characterizing Relative Impact of Scholarly Publications

Anonymous ACL submission

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 Crystal, 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. Crystal outperforms a prior state-of-the-art impact classifier by $+9.5\%$ accuracy and $+8.3\%$ F1 on a dataset of human-annotated citations. Crystal further gains efficiency through fewer LLM calls and outperforms prior baselines using an open-weight model, enabling scalable, cost-effective citation impact analysis. In a case study of ACL Test-of-Time award-winning papers, we find that Crystal's impact characterizations align closely with long-term scientific recognition. We release Crystal-Bank, a 46.8k-paper dataset with rankings and impact labels, along with code.

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 proxies such as citation count (Garfield, 1972). Citation count alone is a poor measure of impact since 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

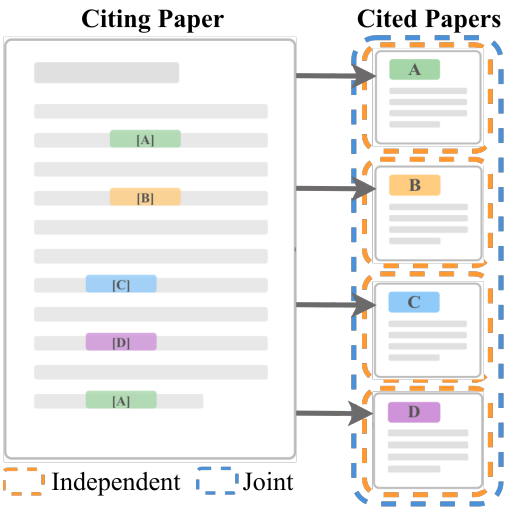

Figure 1: A citing paper p (left) references papers $\{A, B, C, D\}$ (right), with each arrow denoting a citation edge $(p \rightarrow \tilde{p})$ for $\tilde{p} \in \{A, B, C, D\}$. Crystal considers all cited papers *jointly* to assess the *relative* impact $\tilde{p}$ had on p, unlike prior work that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

(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 Crystal, 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 context risks distracting the model.

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 (§5). Crystal also offers an efficiency advantage through fewer LLM calls and outperforms prior baselines using an open-weight model, enabling scalable, cost-effective citation impact analysis (§4.1). We conduct a case study of ACL Test-of-Time award-winning papers (§7), analyzing 46.4k citations of ACL 1996 and ACL 2000 publications. We find that Crystal's rankings align with long-term scientific recognition. We release Crystal-Bank, a dataset with impact labels and rankings (§I), along with our codebase.

In summary, our contributions are: **(a)** We propose Crystal, a method that *jointly* ranks all cited papers within a citing paper using LLMs to assess their *relative* impact. **(b)** We find that this joint ranking approach outperforms a prior state-of-the-art citation impact classifier; ablations confirm the joint formulation drives the gains. **(c)** We show that Crystal is more scalable. **(d)** We observe that Crystal's impact characterizations align closely with long-term scientific recognition. **(e)** We release Crystal-Bank, a dataset of 46,890 papers with rankings and impact labels, and our codebase.[1]

2 Notation and Terminology

Citation graph: We consider a universe of scholarly works $\mathcal{P}$, where each element $p \in \mathcal{P}$ corresponds to a unique paper. 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:

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

[1]Codebase and data will be released.

Citation contexts: We assume 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 for citation function and quality, which we use to define citation impact (§5). However, manual analysis does not scale with the rapid growth of the literature. 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 classification.

Citation Impact Classification: Several prior works, including Jurgens et al. (2018), Cohan et al. (2019), and Lauscher et al. (2022), classify *citation intents* based on function (e.g., background, method, or result). In contrast, our work characterizes *citation impact*, i.e., the degree to which a cited work influences the citing paper. While related, intent and impact are fundamentally distinct, and mapping between them is non-trivial (Appendix A). Motivated by this distinction, prior work has proposed direct approaches to characterize citation impact. For example, Valenzuela et al. (2015) use a feature-based supervised model, and Arnaout et al. (2025) use an LLM judge to classify 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 uses $\mathrm{Ctx}_{\text{all}}(p)$, the full set of citation contexts in a paper, to calibrate each reference against others in the same citing paper.

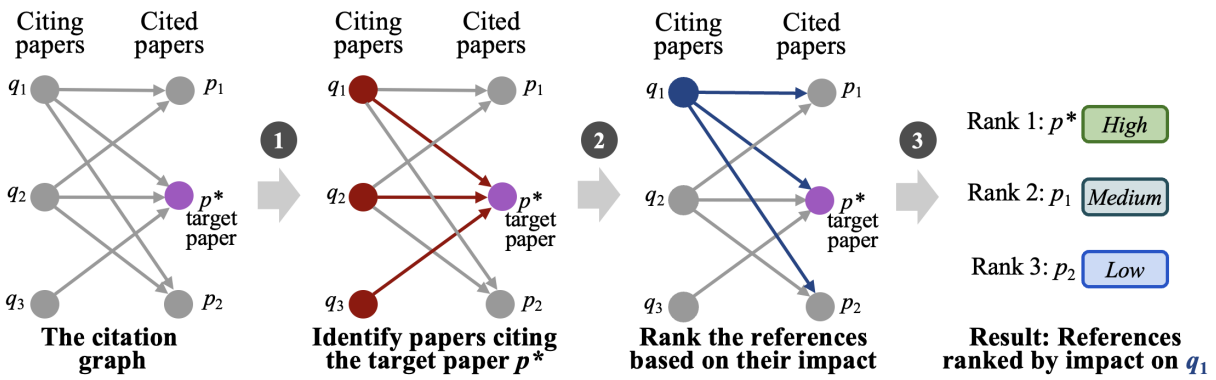

Figure 2: Overview of Crystal, which characterizes the relative impact of a target paper p^* on each of its citing papers, e.g. q_1. ❶ Identify all citing papers of p^* (q_1, q_2, q_3). ❷ For each citing paper, rank its references by impact (e.g., p_1, p_2, p^* for q_1). ❸ The result is an impact-based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

Applications of Citation Content Analysis: Identifying impactful citations is an important task, as it can enhance other applications, such as assessing scientific novelty (Shahid et al., 2025), improving retrieval (Garikaparthi et al., 2025), and tracing key contributions (Zhang et al., 2024). It also enables more nuanced evaluation by distinguishing substantive from perfunctory citations (Manchanda and Karypis, 2021), and informs benchmarks for multi-dimensional impact prediction (Zhu et al., 2026) and prior work forecasting (Ajith et al., 2026).

4 Evaluating Relative Impact via Crystal

We present Crystal, 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 $N_{\text{in}}(p^*)$. In Figure 2, $N_{\text{in}}(p^*) = \{q_1, q_2, q_3\}$.

(2) Citation context extraction: For each downstream paper $q \in N_{\text{in}}(p^*)$, we extract its full reference list $N_{\text{out}}(q)$ and their associated citation contexts $\text{Ctx}_{\text{all}}(q)$ via the same API. Figure 2 shows this process for q_1, where $N_{\text{out}}(q_1) = \{p_1, p^*, p_2\}$. The API returns the parsed sentence containing each citation; when multiple references are co-cited in a single sentence, each is assigned that full sentence as its context (Kinney et al., 2023).

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

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

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

In practice, f is implemented as an LLM-based judge (Vital et al., 2024; Ikoma and Matsubara, 2023; Lahiri et al., 2023; Koloveas et al., 2025, inter alia) and in the simplest setting, it uses binary labels $\mathcal{L} = \{0, 1\}$ for *impactful* and *not impactful*. However, we adopt $\mathcal{L} = \{0, 1, 2\}$ for *low*, *medium*, and *high* impact, as ablations show this finer granularity improves reliability in LLM predictions by reducing high-impact false positives (§6).

The key modeling choice in (4) is that the impact label for a citation $(q \to p)$ is predicted *relative to all other references in* q. This enables calibration against paper-specific citation conventions, such as whether q tends to cite prior work superficially or relies heavily on a few core references.

In our approach, the LLM judge simultaneously ranks q's references by their impact, as shown in Figure 4. 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 (Byerly and Khashabi, 2025), 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 §E for details.

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. Crystal 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 Crystal asymptotically more efficient in terms of LLM calls. This advantage becomes critical at scale. For example, OpenAlex (Priem et al., 2022), comprises 248M papers and 1.9B citation edges, a scale at which Crystal's $O(n)$ complexity offers a decisive computational advantage over edge-level approaches.

Token budget: Token budgets are comparable, as both approaches consume the same citation contexts. However, Crystal incurs lower prompt overhead due to its $O(n)$ calls versus UKP's $O(m)$, meaning the prompt repeats far fewer times. In practice, this overhead is further reduced by prompt caching, making repetitions 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 the evaluation 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 this filtering, the evaluation dataset (Arnaout et al., 2025) comprises 442 citing papers, 1,338 cited papers, and 1,783 citation contexts, each annotated by human labelers as either *impact-revealing* or *other*. There are 1,196 instances (67.1%) of the *other* class and 587 (32.9%) of the *impact-revealing* class. Refer to §8 for details on the type of papers in the data.

Baselines: We compare against five baselines. We include a random classifier, a model always predicting "impact-revealing," and one always predicting "other." We further compare against the supervised classifier of Valenzuela et al. (2015) and UKP (Arnaout et al., 2025), the prior state-of-the-art LLM-based method for identifying impactful citations. We replicate UKP (Arnaout et al., 2025) directly following the authors' method, and obtain Valenzuela et al. (2015) impact predictions from the Semantic Scholar API's `isInfluential` field (Kinney et al., 2023), which Semantic Scholar documents as being generated by this classifier.[2]

We do not compare to citation *intent* classifiers (Cohan et al., 2019; Jurgens et al., 2018; Lauscher et al., 2022), as mapping *intent* labels (e.g., background, method, result) to *impact* labels inevitably introduces assumptions that may unfairly favor or penalize particular methods. A more detailed discussion of this distinction is provided in §A.

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 Figure 4, which also details all inputs to the model. We validate the LLM prompt in a pilot study §F, where we asked annotators to rank their own paper's references by impact; rankings produced by our prompt closely align with these author judgments, achieving Spearman $\rho > 0.7$ across all annotators.

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-weight 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 on the human-annotated dataset released by Arnaout et al. (2025), which labels citations as *impact-revealing* or *other*. To enable comparison across baselines, we map each system's label set to these two categories at evaluation time; inference is performed in each system's original label space, with mapping applied only post hoc (details are in §A).

We assess performance using standard classifi-

[2] https://www.semanticscholar.org/faq/influential-citations

Model	Method	Accuracy(↑)	Bal. Acc.(↑)	MCC(↑)	P(↑)	R(↑)	F1(↑)	F1$_{-}$(↑)
Baselines	Random	$49.9_{\pm 1.2}$	50.2	0.00	33.1	50.9	40.1	57.0
	All Imp-Rev	$32.9_{\pm 1.1}$	50.0	0.00	32.9	100.0	49.5	0.0
	All Non-Imp	$67.1_{\pm 1.1}$	50.0	0.00	0.0	0.0	0.0	80.3
	Valenzuela	$70.6_{\pm 1.1}$	67.1	0.35	58.0	53.2	55.5	79.5
GPT-5.1	UKP	$66.7_{\pm 1.1}$	65.9	0.30	49.6	63.5	55.7	73.4
	Crystal	$78.6_{\pm 1}$	75.8	0.54	72.2	63.7	67.7	85.5
	Δ	**+11.9**	**+9.9**	**+0.24**	**+22.6**	**+0.2**	**+12.0**	**+12.1**
o4-mini	UKP	$63.0_{\pm 1.1}$	65.7	0.29	46.1	73.6	56.7	67.7
	Crystal	$65.4_{\pm 1.1}$	74.2	0.53	76.3	57.1	65.3	86.0
	Δ	**+2.4**	**+8.5**	**+0.24**	**+30.2**	−16.5	**+8.6**	**+18.3**
Qwen3	UKP	$60.8_{\pm 1.2}$	64.9	0.28	44.5	76.7	56.3	64.5
	Crystal	$75.1_{\pm 1}$	71.1	0.46	70.0	53.3	60.5	83.9
	Δ	**+14.3**	**+6.2**	**+0.18**	**+25.5**	−23.4	**+4.2**	**+19.4**

Table 1: Comparison of our approach Crystal against prior state-of-the-art method UKP (Arnaout et al., 2025) and four baselines: random label selection, always predicting "impact-revealing," always predicting "other," and the method of Valenzuela et al. (2015). Reported accuracies (in percentages) include standard error (± SE). Bal. Acc. = balanced accuracy; MCC = Matthews Correlation Coefficient; F1$_{-}$ = F1 on the negative (non-impact-revealing) class. Green cells show positive Δ := Crystal− UKP; **bold** green cells show Crystal's gains. **Our approach outperforms the prior state-of-the-art method and all other baselines across all models.**

cation metrics. Specifically, we compute accuracy as the fraction of citations whose predicted labels match the ground truth, and report precision, recall, and F1 score for the *impact-revealing* class in Table 1. We additionally report the following balance aware metrics: balanced accuracy—the average of per-class recalls, ensuring equal contribution from each class—and Matthews Correlation Coefficient (MCC), which accounts for all four quadrants of the confusion matrix and yields 0.0 for trivial single-class predictors. For completeness, we also report the F1$_{-}$ score for the *other* class in Table 1.

Results: As shown in Table 1, Crystal consistently outperforms prior state-of-the-art impact classifier, which we refer to as UKP (Arnaout et al., 2025), improving accuracy by +9.5% and F1 by +8.3% on average across all three models. Crystal achieves MCC scores of 0.46–0.54 and balanced accuracy of 71.1%–75.8%, compared to UKP's MCC of 0.28–0.30 and balanced accuracy of 64.9%–65.9%, **indicating far stronger discrimination across both classes.** Crystal consistently achieves higher precision 70.0%–76.3% precision versus UKP's 44.5%–49.6%, meaning **our predictions for impactful citations are considerably more reliable.** Importantly, Table 1 reports the standard error (SE) of accuracy, which remains low (≈±1.0%–1.2%) across all settings, indicating that **the estimates are stable.** Moreover, the observed improvements over prior work are consistently larger than the corresponding SE (e.g. +14.3%), suggesting that the gains are robust.

Compared to the method of Valenzuela et al. (2015), Crystal achieves higher F1 across all three models (+5.0%–+12.2%), as well as higher MCC (+0.11–+0.19) and balanced accuracy (+4.0%–+8.7%), demonstrating stronger and more consistent discrimination.

We also find that Qwen3-30B is a strong alternative to GPT-5.1, achieving a comparab performance at lower cost. As shown in Figure 6, o4-mini and Qwen3-30B achieve higher recall on UKP mainly due to a tendency to over-predict the impact-revealing class. We additionally observe that models sometimes omit references from their rankings when papers have long reference lists. We treat such cases as misclassifications in Table 1, and even under this strict evaluation, Crystal outperforms all baselines. As described in §C, this limitation is expected to diminish as models' long-context capabilities improve. Overall, **Crystal outperforms prior state-of-the-art method and all baselines** with substantial gains across all settings.

Qualitative Error Analysis: A domain expert (author) manually analyzed the 45 false positives (FP) and 166 false negatives (FN) that all three models (QWEN3-30B, GPT-5.1, and O4-MINI) misclassified in Crystal, identifying commonalities in these errors. FP contexts tend to describe the cited work in detail and credit its authors as

the source of a specific method or finding, often through verbs such as *proposed* or *introduced* or through comparative phrasing such as *similar to* or *variant of*, and they are substantially longer than true negatives (median 954 vs. 238 characters). FN contexts show two recurring patterns. The first is contexts that are substantially shorter than true positives (median 280 vs. 1,110 characters), offering little surrounding text for the model to draw on. The second is contexts that group the cited work with several other references in the same citation. **Crystal may associate the citation with its neighbors, diluting and downgrading its impact label.** Length statistics and examples for both error categories (FP and FN) are provided in §D.

6 Ablation Studies

We perform three ablation studies to isolate the contributions of Crystal's key design choices.

1) Three vs. Two Impact Labels We ablate Crystal's three-class *High*/*Medium*/*Low* formulation against the binary *impact-revealing* vs. *other* convention of Arnaout et al. (2025), using the same Qwen3-30B model and experimental setup as §5. This isolates whether performance gains stem from label granularity or from Crystal's joint ranking formulation. Figure 7 shows the prompt used.

As shown in Table 4, Crystal outperforms UKP even under the binary setting ($+8.5\%$ accuracy, $+4.0\%$ balanced accuracy, $+0.09$ MCC), confirming that *the joint ranking formulation, not label count, is the primary driver of improvement.* Moving to three labels yields further gains ($+12.6\%$ in precision and $+2.0\%$ F1). Analysis of the confusion matrices (Figure 8) indicates that the *Medium* bin absorbs borderline cases that the binary model incorrectly assigns to *impact-revealing*, reducing false positives and increasing precision from 57.1% to 70%. Thus, **finer label granularity gives the model a more calibrated decision space**, enabling it to reliably identify truly high-impact citations.

2) Independent Variant vs. Joint Ranking: We ablate Crystal's joint ranking formulation against an independent-input variant, where each citation context Ctx_p is assessed in isolation. To match Crystal's majority voting procedure, we run the independent variant three times per citation and take the majority label. Impact definitions, shown in Figure 10, are identical across both settings.

As shown in Table 5, the ablation reveals a clear trade-off between recall and calibration. Using independent input leads to a strong bias toward predicting the *High* class, yielding very high recall (91.7%) but poor precision (46.8%) and lower overall agreement (MCC 0.40). This is reflected in the confusion matrix (Figure 9), where both *Impact-Revealing* and *Other* instances are frequently assigned to *High*. In contrast, joint input substantially improves class discrimination, increasing precision ($+23.2\%$), accuracy ($+12.1\%$), and MCC ($+0.06$), while producing a more balanced prediction distribution; particularly by correctly assigning most *Other* cases to *Low*. Although recall decreases (-38.4%), the joint formulation yields better-calibrated and more reliable predictions overall, suggesting that **shared context is critical for mitigating the overprediction bias observed in the independent setting**.

3) Permutation-based majority voting versus a single-pass ranking: We ablate Crystal's permutation-based majority voting procedure vs a single-pass ranking variant that uses a different randomized reference ordering per pass. This isolates the contribution of voting from the underlying model quality, and tests whether majority voting over permuted orderings robustly mitigates positional bias.

As shown in Table 6, individual runs produce noticeably different results across orderings for all three models, confirming that positional bias is a real source of instability. Majority voting consistently matches or exceeds the best individual run across nearly all metrics. The benefit is most pronounced for o4-mini (+11.9% accuracy, $+2.8\%$ F1) and Qwen3 ($+8.5\%$ accuracy, $+1.8\%$ F1), both of which are more sensitive to reference ordering in the single-pass setting. For GPT-5.1, where individual runs are already tightly clustered, voting still yields modest gains ($+1.5\%$ accuracy, $+0.2\%$ F1) without hurting any metric. Thus, **permutation-based majority voting helps most when the underlying model is less consistent, and does not hurt when it is already stable**—making it a robust default across model families.

To summarize, our ablations explain Crystal's gains: **(1)** the *joint* ranking formulation is the main driver of improvement **(2)** a three-class formulation further improves precision via better calibration; and **(3)** permutation-based majority voting mitigates positional bias, especially for weaker models.

7 Case Study: ACL Test-of-Time Awards

We study whether Crystal's high-impact impact citations are correlated to long-term scientific recognition, as measured by the ACL Test-of-Time (ToT) Paper Award, which annually recognizes up to four NLP and computational linguistics papers (two published 25 years prior and two published 10 years prior) for their enduring impact on the field.[3]

Dataset	#Papers	Actual Citing	S2 API	API + Scrape
ACL 1996	577	29,308	4,216	18,451
ACL 2000	1,184	40,681	6,639	27,997

Table 2: Coverage statistics for ACL 1996 and 2000 papers and their citing papers. "#Papers" denotes the number of papers published at ACL in each year. "Actual Citing" is the total number of papers citing these publications. "S2 API" is the subset of citing papers with non-empty reference data returned by the Semantic Scholar API. "API + Scrape" includes additional citing papers recovered via website scraping, which substantially improves coverage of reference information.

Data: We begin by selecting three ACL papers that received the Test-of-Time (ToT) Award: Berger et al. (1996), Carletta (1996), and Gildea and Jurafsky (2002). We choose these papers because they were recognized 25 years after publication and were honored in different announcement years. Berger et al. (1996) and Carletta (1996) were the first recipients in 2021, while Gildea and Jurafsky (2002) received the award in 2025.

We follow the approach in §4 to construct the citation graph. Using the ACL-OCL dataset (Rohatgi et al., 2023), we collect all ACL papers published in 1996 and 2000. We then identify all papers that cite them using the Semantic Scholar API (Kinney et al., 2023) and retrieve their citation contexts. At scale, about 85% of API responses lack reference details due to licensing restrictions (e.g., "Notice: The following paper fields have been elided by the publisher"). Although these details are often unavailable through the API, they are typically visible on the Semantic Scholar website. We thus supplement the API data by scraping the site to recover missing reference data and citation contexts. Table 2 displays the coverage statistics of the citing papers considered in the analysis for each year.

[3] https://www.aclweb.org/portal/content/announcement-2025-acl-test-time-paper-award

Experiment: As described in §4, we apply Crystal to characterize the impact of all references in papers that cite publications from ACL 1996 and ACL 2000. Concretely, we run Crystal on 18,451 papers that cite ACL 1996 publications and on 27,997 papers that cite ACL 2000 publications, as summarized in Table 2.

We use the same Qwen3-30B model described in §5 and the same prompt §4 when running Crystal. This ensures that the citation context extraction and downstream scoring are consistent with the setup used in the main experiment §5.

Results: We aggregate the impact labels produced by Crystal for each paper published in ACL 1996 and ACL 2000. Specifically, for each paper in these two sets, we consider all papers that cite it and collect the impact labels assigned by Crystal to those citations. We then summarize these labels by counting, for each paper, how many of its citations are classified as high-, medium-, or low-impact. As shown in Table 8, the label distribution is highly consistent across both years. Approximately 50–51% of citations are labeled as low impact, 30–34% as medium impact, and 17–19% as high impact. This indicates that **Crystal assigns the high-impact label selectively**, reserving it for a relatively small fraction of citations.

High-Impact Citation Counts Analysis: We rank ACL 1996 and ACL 2000 publications by the number of high-impact citations they receive. We then examine the top 10 papers for each year. As shown in Figure 3, the ToT awarded paper Berger et al. (1996) ranks first with 712 high-impact citations, making it the top-ranked paper under Crystal in ACL 1996. Comparing this ranking to one based on raw citation counts leads to a different conclusion about which papers are most impactful. The second-ranked paper, Levelt et al. (1999), has a higher total number of citations but fewer high-impact citations. This difference shows that citation count alone is a weak proxy for impact, as it does not distinguish between citations that reflect substantive use and those that are merely incidental. However, **high-impact citation counts are not sufficient on their own to explain long-term recognition.** This becomes clear when comparing against ToT awards. While Berger et al. (1996) is both top-ranked and a ToT recipient, this alignment does not generalize. For example, Carletta (1996) ranks fourth with 190 high-impact citations yet received the ToT award, whereas Ratnaparkhi (1996)

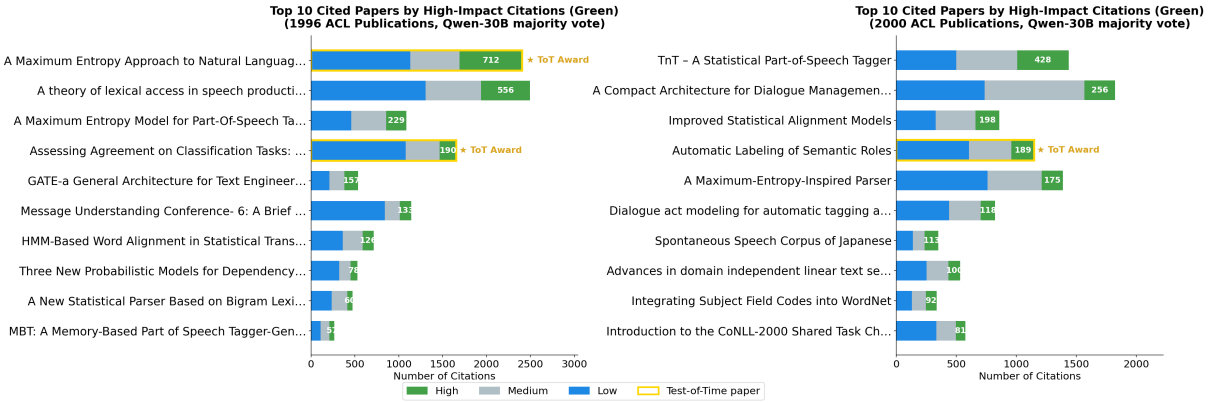

Figure 3: Impact distribution of ACL 1996 (left) and ACL 2000 (right) publications, ranked by the number of high-impact citations as predicted by Crystal, showing the top 10 in each year with the top-ranked paper at the top of each plot. Each bar shows the breakdown of high, medium, and low impact citations per paper, with the high-impact count annotated. Papers awarded the ACL Test-of-Time (ToT) Award are highlighted with a yellow border: Berger et al. (1996) and Gildea and Jurafsky (2002). In both years, ToT-winning papers rank among the top in high–impact citations, suggesting that Crystal's impact characterization aligns with long-term community recognition of foundational contributions.

has more high-impact citations (229) but was not selected. A similar pattern appears in ACL 2000, where Gildea and Jurafsky (2002) ranks fourth despite being the ToT recipient. These cases indicate that even impact-aware aggregation does not fully capture the factors behind long-term recognition.

Temporal Dynamics of Citation Impact: To better understand these differences, we examine the temporal distribution of citation impact. Figures 17 and 18 show how high-, medium-, and low-impact citations are distributed year by year for the top four papers in ACL 1996 and ACL 2000, respectively. The top four papers in each year are selected by ranking all papers according to their number of high-impact citations and then taking the highest four; the ToT papers are included in this subset for both conferences. Berger et al. (1996) has the largest volume of high-impact citations among the ACL 1996 papers, *sustained* across more than two decades. Carletta (1996) shows a stable but smaller number of high-impact citations, with most of its citation volume concentrated in the low-impact category. For ACL 2000, Gildea and Jurafsky (2002) maintains a stable level of high-impact citations through 2025, despite not ranking first by aggregate counts. In contrast, the **non-ToT papers** in the top four for ACL 2000 **show high-impact citations that peak earlier and then decline more steeply.**

In summary, these observations highlight three distinct points. (**1**) Citation counts do not capture *how* a paper is cited. (**2**) Impact characterization addresses this limitation but ignores *when* that influence occurs. (**3**) Temporal patterns reveal whether impact is sustained, concentrated, or diffuse over time, which helps explain differences in long-term recognition.

Overall, Crystal offers a more informative view of scholarly influence than raw citation counts, though it does not fully predict ToT selections. Nevertheless, **Crystal's impact characterization aligns with long-term community recognition of foundational contributions**, as all three ToT-awarded papers in our analysis rank among the top four in their respective ACL conference years by high-impact citations.

8 Conclusion

As shown in Table 1, Crystal consistently outperforms prior state-of-the-art impact classifiers, improving accuracy by $+9.5\%$ and F1 by $+8.3\%$ on average across all three models. We also find that Qwen3-30B is a strong alternative to GPT-5.1, achieving a strong performance at lower cost. Beyond benchmark performance, our case study (§7) shows that high-impact citations identified by Crystal align with long-term scientific recognition, as signaled by the ACL Test-of-Time Paper Award.

In conclusion, these results show that *jointly* modeling all references within a citing paper provides richer signal than independent citation classification, and Crystal offers a scalable and cost-effective solution for 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, the dataset is limited to English-language scientific papers.

Despite these limitations, Crystal is broadly applicable to scholarly work containing citations. It is most naturally suited for research and position papers, where citations vary meaningfully in how central they are to the paper's argument, making both ranking and characterizing their impact informative. For survey papers, a full ranking of references may be less informative, as surveys aim for comprehensive coverage rather than building on a few key works. Nevertheless, Crystal's impact categorization (high, medium, low) remains valuable for distinguishing landmark works from those included primarily for completeness.

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 §C.

Also, if the Semantic Scholar API (Kinney et al., 2023) fails to retrieve information for a given paper, the next 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 §F.

Ethics Statement

The dataset used in Crystal 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. Thus, no further anonymization was applied.

Crystal 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.

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C Kozlowski, Oyvind Tafjord, James Evans, Daniel S Weld, Tom Hope, and Doug Downey. 2026. Prescience: A benchmark for forecasting scientific contributions. *arXiv preprint arXiv:2602.20459*.

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.

Adam Berger, Stephen A Della Pietra, and Vincent J Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Adam Byerly and Daniel Khashabi. 2025. Self-consistency falls short! the adverse effects of positional bias on long-context problems.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

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.

Daniel Gildea and Dan Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

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.

Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38.

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.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on empirical methods in natural language processing*.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The acl ocl corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361.

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.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. 2017. Breaking the softmax bottleneck: A high-rank rnn language model. *arXiv preprint arXiv:1711.03953*.

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*.

Hangxiao Zhu, Yuyu Zhang, Ping Nie, and Yu Zhang. 2026. Sciimpact: A multi-dimensional, multi-field benchmark for scientific impact prediction. *arXiv preprint arXiv:2604.17141*.

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 Citation Intent vs. Citation Impact

The notions of citation intent and impact are fundamentally distinct, and mapping intent labels to impact labels is inherently non-trivial. A direct comparison between intent-classification methods and impact-classification methods requires manually defining a mapping from intent categories to impact categories. However, such mappings inevitably introduce assumptions that may unfairly favor or penalize particular methods. For example, one may map *background* citations to non-impactful and *uses*/*extends* citations to impactful, following prior work (Arnaout et al., 2025). However, this simplification can lead to unstable or misleading labels. To illustrate, consider the paper by Yang et al. (2017) in the MultiCite (Lauscher et al., 2022) test set.[4] Its annotations are evenly split across the categories *background*, *uses*, and *extends* (5 annotations each). Under a majority-vote mapping scheme with random tie-breaking, the resulting impact label has 33% chance of being mislabeled as non-impactful.

More importantly, citation intent does not uniquely determine citation impact. A citation unanimously labeled as *background* may still be highly impactful if it introduces a foundational concept on which the citing work critically depends. For these reasons, we do not include intent-based baselines (Cohan et al., 2019; Jurgens et al., 2018; Lauscher et al., 2022) in our main comparisons and instead focus on methods explicitly designed to assess citation impact.

Evaluation Label Mappings: All systems are evaluated against the ground-truth categories of *impact-revealing* and *other* from Arnaout et al. (2025). Table 3 shows how each system's predicted labels are mapped to these categories for evaluation purposes only; each system predicts over its own native label set during inference. In Crystal, we map *High* to *impact-revealing*, and both *Medium* and *Low* to *other*. This mapping reflects the definition of impact in Arnaout et al. (2025), where *impact-revealing* citations explicitly emphasize the significance or influence of prior work. While *Medium* citations may indicate relevance or usefulness, they do not consistently foreground such impact, and are therefore grouped with *other* for evaluation. Notably, we retain the three-label design (*High*, *Medium*, *Low*), as ablations show that this finer granularity yields a more calibrated decision space, enabling more reliable identification of truly high-impact citations (§6).

System	Predicted	Ground-Truth
Arnaout et al.	Impact-revealing Other	Impact-revealing Other
Valenzuela et al.	Meaningful Non-meaningful	Impact-revealing Other
Crystal (ours)	High Medium Low	Impact-revealing Other Other

Table 3: Evaluation-time label mappings to ground-truth categories from Arnaout et al. (2025), applied only at evaluation time.

Sample Outputs: Figure 4 shows our definition of citation impact and the instruction used to rank a citing paper's references, with a sample output in Figure 5. Figure 16 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 (§E).

B Results from ablation studies

We perform three ablation studies to isolate the contributions of Crystal's key design choices: label granularity (Table 4, Figure 8), joint ranking formulation (Table 5, Figure 9), and permutation-based majority voting (Table 6).

C 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 11 shows this effect.

When the citing paper contains fewer than 40 references, all three models successfully rank nearly

[4] https://github.com/allenai/multicite/blob/master/data/classification_gold_context/test.json

The task is to rank all of the references R (where R is a list of papers $r_1, r_2, \ldots$) 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 and impactCategory was assigned)
- `"impactCategory"`: 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. **DO NOT** rank more references than the ones on the list of references given.

Figure 4: Main Crystal prompt used in all primary experiments. The model jointly ranks all references R of a given paper P by impact, assigning each to one of three categories (*High*, *Medium*, *Low*). Bracketed fields (`[main_paper_title]`, etc.) are filled at inference time.

Model	Method	Accuracy(↑)	Bal. Acc.(↑)	MCC(↑)	P(↑)	R(↑)	F1(↑)	F1$_{-}$(↑)
Qwen3	UKP	$60.8_{\pm 1.2}$	64.9	0.28	44.5	76.7	56.3	64.5
	Crystal (2 labels)	$69.3_{\pm 1.1}$	68.9	0.37	57.1	60.0	58.5	78.8
	Δ_{2L-UKP}	+8.5	+4.0	+0.09	+12.6	−16.7	+2.2	+14.3
	Crystal (3 labels)†	$75.1_{\pm 1.0}$	*71.1*	*0.46*	*70.0*	*53.3*	*60.5*	*83.9*
	Δ_{3L-2L}†	+5.8	+2.2	+0.09	+12.9	−6.7	+2.0	+5.1

Table 4: Results of Ablation 1, evaluated with Qwen3-30B, ablating Crystal's three-label formulation (high, medium, low) down to two labels (impact-revealing vs. other) following Arnaout et al. (2025). Δ_{2L-UKP} shows the gain of Crystal (2 labels) over UKP. Δ_{3L-2L} shows the further gain of three labels over two. Green cells indicate positive differences. †Full Crystal (3 labels) results shown for reference only. **The finer granularity of three labels gives the model a more calibrated decision space: the model more reliably identifies truly high-impact citations.**

Model	Method	Accuracy(↑)	Bal. Acc.(↑)	MCC(↑)	P(↑)	R(↑)	F1(↑)	F1$_{-}$(↑)
Qwen3	Independent	$63.0_{\pm 1.1}$	70.3	0.40	46.8	91.7	61.9	63.9
	Crystal(Joint)	$75.1_{\pm 1.0}$	71.1	0.46	70.0	53.3	60.5	83.9
	Δ	+12.1	+0.8	+0.06	+23.2	−38.4	−1.4	+20.0

Table 5: Shows the results of Ablation 2, which compares considering citation contexts **independently** against **jointly** ranking of all citations within a paper, evaluated with Qwen3-30B. Δ := Crystal − Independent. Green cells indicate gains from joint ranking. The precision gains show that considering contexts jointly improves the identification of truly high-impact citations and corrects the overprediction bias observed in the independent setting.

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.

D Qualitative Error Analysis Details

We provide additional material supporting the qualitative error analysis in §5. We include examples of both false positives (FP) and false negatives (FN), drawn from cases that were unanimously misclassified by all three models (Qwen3-30B, GPT-5.1, and o4-mini) under Crystal, identifying commonalities in these errors.

Figure 12 compares context lengths between unanimously misclassified cases and correctly classified instances. Figures 13 and 14 present representative examples for each error type.

E 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 §C, 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

Model	Method	Accuracy(↑)	Bal. Acc.(↑)	MCC(↑)	P(↑)	R(↑)	F1(↑)	F1$_{-}$(↑)
GPT-5.1	Run 1	$76.8_{\pm 1.0}$	75.3	0.53	71.7	62.7	66.9	85.2
	Run 2	$77.1_{\pm 1.0}$	75.3	0.52	71.1	63.2	66.9	85.1
	Run 3	$76.9_{\pm 1.0}$	75.7	0.53	71.1	64.2	67.5	85.2
	Crystal	$78.6_{\pm 1.0}$	75.8	0.54	72.2	63.7	**67.7**	85.5
	Δ	+1.5	+0.1	+0.01	+0.5	−0.5	+0.2	+0.3
o4-mini	Run 1	$53.5_{\pm 1.2}$	72.4	0.50	76.0	53.0	62.5	85.5
	Run 2	$52.7_{\pm 1.2}$	72.3	0.50	77.3	52.1	62.3	85.6
	Run 3	$52.6_{\pm 1.2}$	71.8	0.48	74.5	52.3	61.5	85.0
	Crystal	$65.4_{\pm 1.1}$	74.2	0.53	76.3	57.1	65.3	86.0
	Δ	+11.9	+1.8	+0.03	−1.0	+4.1	+2.8	+0.4
Qwen3	Run 1	$65.3_{\pm 1.1}$	68.3	0.40	66.9	48.2	56.0	82.6
	Run 2	$66.0_{\pm 1.1}$	70.1	0.46	73.1	49.1	58.7	84.3
	Run 3	$66.6_{\pm 1.1}$	69.1	0.43	69.7	48.6	57.2	83.4
	Crystal	$75.1_{\pm 1.0}$	71.1	0.46	70.0	53.3	60.5	83.9
	Δ	+8.5	+1.0	0.00	−3.1	+4.2	+1.8	−0.4

Table 6: Shows the results of Ablation 3, which studies the effect of **majority voting** across three randomized orderings vs. any single run. $\Delta :=$ Crystal $-$ best single run, where the best is taken column-wise across the three runs. Green cells show gains from majority voting. **Bold** indicates the highest F1 overall. **Majority voting consistently matches or exceeds even the strongest individual run**, demonstrating robustness to positional bias.

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 16.

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.

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 (§E) for each citing paper. For every reference in the aggregated list, we construct the same feature vector used during training from

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

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 16.

Results. Table 7 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 15.

F 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 4 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 19–22 illustrate the custom annotation interface we designed for the study. Figure 19 shows the main annotation task layout. Figure 20 displays the complete list of references for an annotator's paper. Figure 21 demonstrates the feature we added to reveal cited papers' citation context. Lastly, Figure 22 presents the final ranked list generated within the interface. As shown in Figure 22, 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 4). As shown in Figure 23, 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.

G ACL Case Study Details

Table 8 reports the number of high-, medium-, and low-impact citations identified by Crystal in the ACL case study. Figures 17 and 18 show the temporal distribution of these citation types for the top four papers in ACL 1996 and ACL 2000, respectively, ranked by their high-impact citation counts under Crystal.

Dataset	High	Medium	Low
ACL 1996	3,934 (19%)	6,190 (30%)	10,696 (51%)
ACL 2000	5,987 (17%)	12,184 (34%)	17,901 (50%)

Table 8: Distribution of impact labels assigned by Crystal across all citations of ACL 1996 and ACL 2000 publications. "High", "Medium", and "Low" report the number (and percentage) of citations assigned to each impact category.

H Details about Models Used in Experiments

As described in §5, 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-weight 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. We use these parameters for all experiments in §5, §6, and §7.

The total cost of running the experiments described in §5 was $77.51 for the closed-source models, including approximately $48 for the UKP experiments and $30 for Crystal experiments. All ex-

periments involving Qwen3-30B (in §5, §6, and §7) were conducted by hosting the model with vLLM on H100 GPUs. Although precise runtime measurements for all Qwen experiments are unavailable, based on our setup (H100 GPU with vLLM and FP8 inference), we estimate that the full set of experiments, i.e. running Crystal on Crystal-Bank (46,890 papers), can be completed within approximately 10 days of continuous runtime on a single instance. This runtime could be further reduced by using multiple GPUs.

I Crystal-Bank

Crystal-Bank contains 46,890 citing papers with rankings and impact labels produced by Crystal. These correspond to the citing papers used in §5 and §7, comprising 442 and 46,448 papers, respectively.

J 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.

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

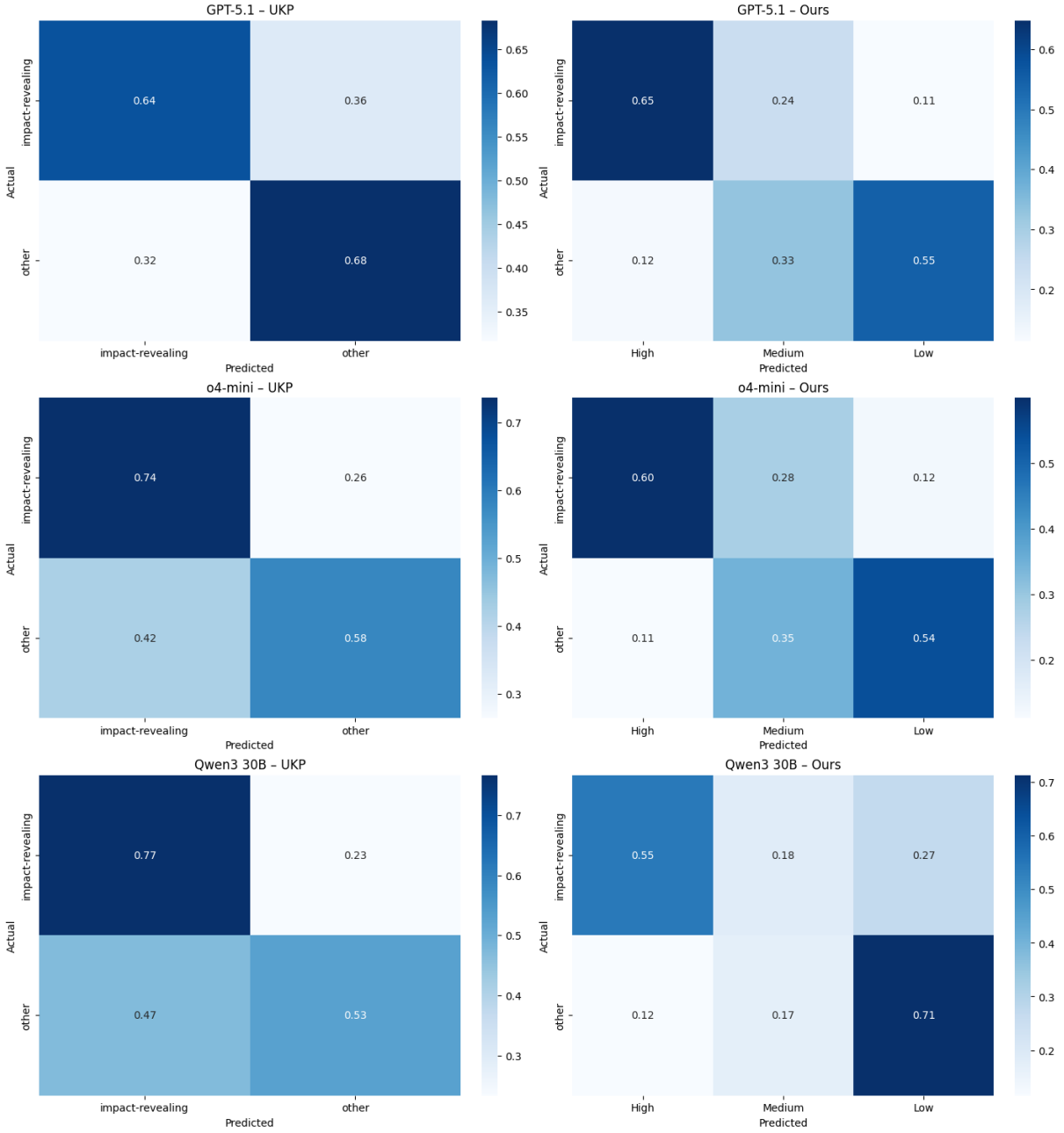

Figure 6: Confusion matrices per model for assigning impact labels through majority vote. Rows represent actual labels; columns represent predicted categories. To compare both approaches (collapsing High to impact-revealing and Medium/Low to other, as described in 5), we find that our method produces 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.

The task is to rank all of the references R (where R is a list of papers $r_1, r_2, \ldots$) 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, abstract, and context of citation): `[references_text]`

Impact categories:

1. "impact-revealing" 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. "other" 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. These citations could also provide background, context, or perfunctory acknowledgement, but the core contribution of your paper is not dependent on them in any strong way.

- *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.
- *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 and impactCategory was assigned)
- `"impactCategory"`: string (`"impact-revealing"`, `"other"`)

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. **DO NOT** rank more references than the ones on the list of references given.

Figure 7: Prompt used in Ablation 1, which collapses Crystal's three-label schema to two labels (impact-revealing vs. other) following Arnaout et al. (2025). Bracketed fields (`[main_paper_title]`, etc.) are filled at inference time.

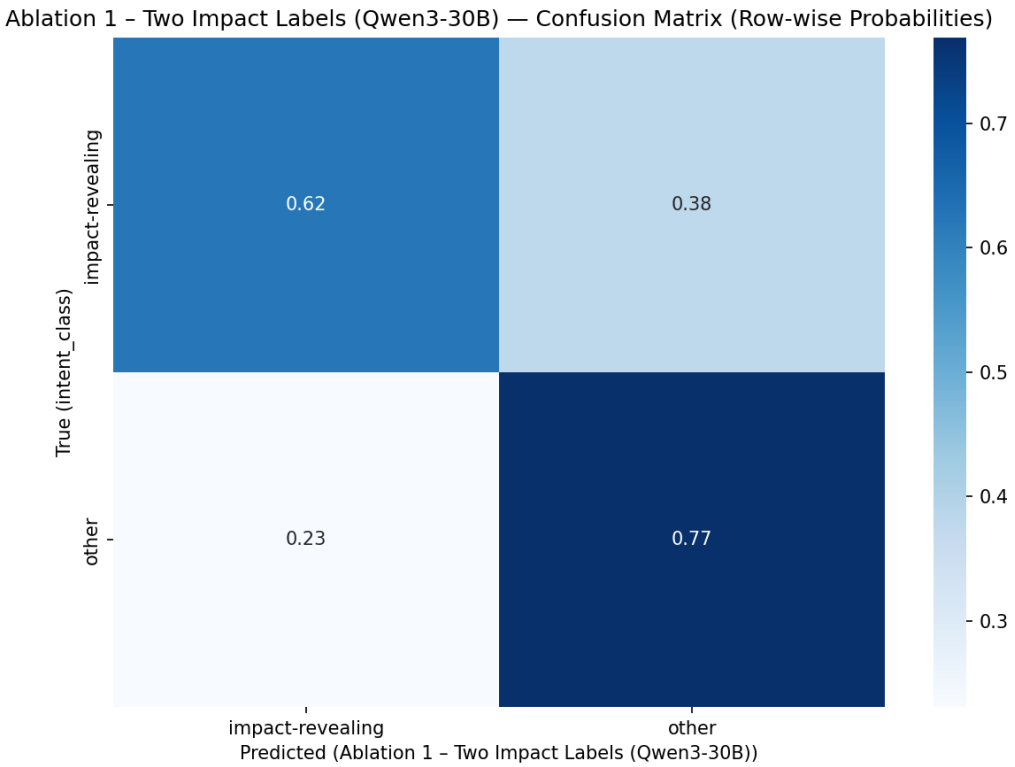

Figure 8: Confusion matrix (row-wise probabilities) for Ablation 1 (two-label variant). Compared to Figure 6 (Qwen3-30b), the three-label setting introduces a *Medium* bin that captures borderline cases the binary model misclassifies as *impact-revealing*, reducing false positives (precision: $57.1\% \rightarrow 70\%$, F1: $+2\%$). Thus, we conclude that three impact labels yield more true impact-revealing classifications and more calibrated judgements.

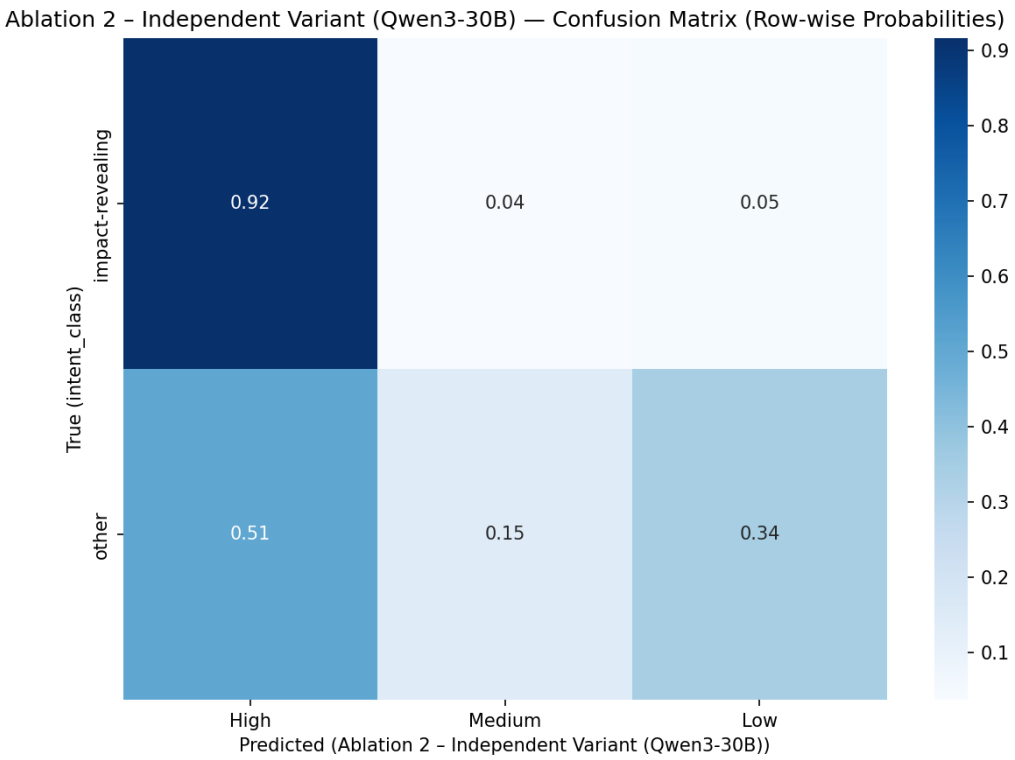

Figure 9: Confusion matrix (row-wise probabilities) for Ablation 2 (independent ranking). The independent variant shows a strong bias toward predicting *High*. In contrast, joint ranking (as shown in Figure 6) improves class discrimination and calibration, reducing overprediction and yielding more balanced, reliable predictions.

The task is to assess how impactful and influential a single cited reference was on a given paper P.

The paper P you will be analyzing is:

- Title: `[citing_title]`
- Abstract: `[citing_abstract]`

The cited reference to assess is:

- Paper ID: `[cited_id]`
- Title: `[cited_title]`
- Citation context from P: `[context]`

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:

- `"paperId"`: string
- `"title"`: string
- `"contexts"`: string (all citation contexts)
- `"reason"`: string (why this rank and impactCategory was assigned)
- `"impactCategory"`: 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. **DO NOT** rank more references than the ones on the list of references given.

Figure 10: Prompt used in Ablation 2, which scores each citation context **independently** (i.e., in isolation, without access to the full reference list). The model is queried three times per citation and the majority label is taken as the final prediction. Bracketed fields (`[citing_title]`, etc.) are filled at inference time.

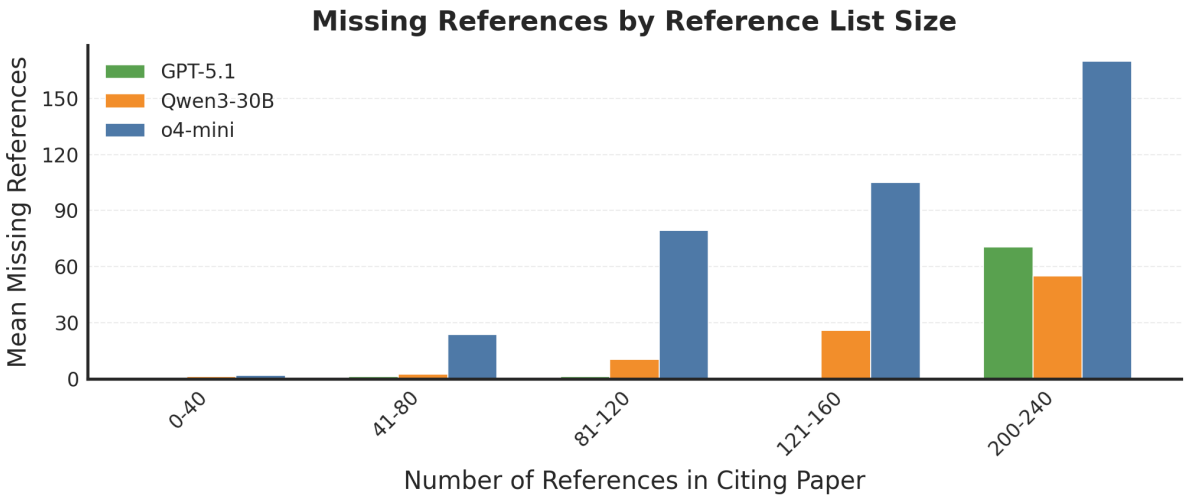

Figure 11: Mean number of missing references in the aggregated ranks per model as a function of the number of references in the citing paper. All three models rank nearly all references when the reference list is less than 40. However, they omit more references as the list grows longer. o4-mini exhibits the steepest degradation, while GPT-5.1 and Qwen3-30B are more robust.

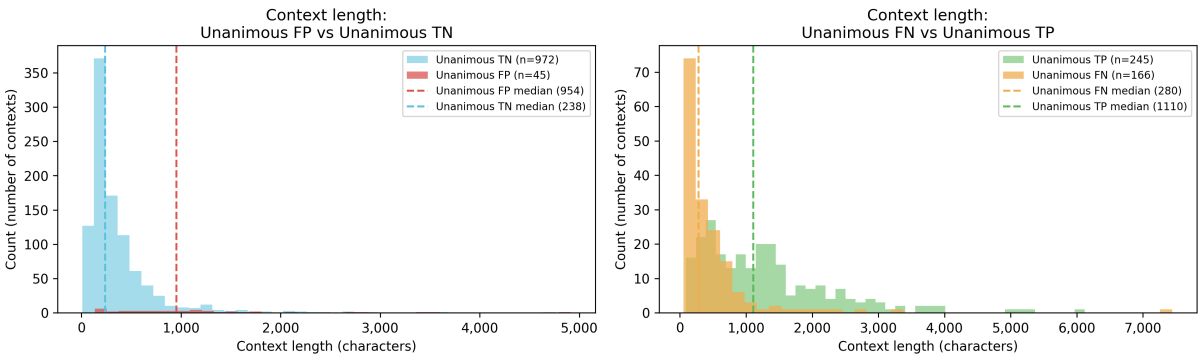

Figure 12: Distribution of citation context length (in characters) for cases unanimously classified by QWEN3-30B, GPT-5.1, and O4-MINI under Crystal. **Left:** unanimous false positives (FP, n=45) versus unanimous true negatives (TN, n=972), with medians of 954 and 238 characters respectively. **Right:** unanimous false negatives (FN, n=166) versus unanimous true positives (TP, n=245), with medians of 280 and 1,110 characters respectively. Dashed lines mark the medians for each group.

To automatically obtain a large training dataset, DS has been proposed (Mintz et al., 2009). Mintz et al. (2009) reports that distant supervision may lead to more than 30

(Ren et al. 2016) proposed a multimodal Long Short-Term Memory (LSTM) for speaker identification, which referred to locating a person who has the same identity with the ongoing sound in a certain video. Inspired by (Ren et al. 2016), we propose to build temporal dependency across visual and audio modalities through sharing weights for video caption, aiming at exploring whether temporal dependency across visual and audio modalities can capture the resonance information among them or not. Ren et al. (Ren et al. 2016) proposed a multimodal Long Short-Term Memory (LSTM) for speaker identification, which referred to locating a person who has the same identity with the ongoing sound in a certain video.

More recent studies (Brock et al., 2018; Shirakawa et al., 2018; Pham et al., 2018; Liu et al., 2019; Xie et al., 2019; Cai et al., 2019), on the other hand, optimize the weights and the architecture simultaneously within a single training by treating all possible architectures as subgraphs of a supergraph. We generalize the work by Shirakawa et al. (2018) to enable arbitrary types of architecture variables including categorical variables, ordinal (such as real or integer) variables, and their mixture. In the last direction, promising approaches transform a coupled optimization of weights and architectures into optimization of a differentiable objective by means of continuous relaxation (Liu et al., 2019; Xie et al., 2019) or stochastic relaxation (Shirakawa et al., 2018; Pham et al., 2018). . . . Shirakawa et al. (2018) has introduced it to model connections and types of activation functions in multi-layer perceptrons.

Figure 13: Examples of contexts that were misclassified as impact-revealing by QWEN3-30B, GPT-5.1, and O4-MINI under Crystal. In each case the cited work is described in detail and attributed to its authors, although the citing paper does not build on it.

[11] leveraged the reinforcement learning to automatically prune the convolution channels.

Our method also connects to PinSage [21] and GAT [15].

We run the 4 models on the datasets and respective splits from [Yang et al., 2016].

Discriminative models are known to suffer from catastrophic forgetting when learning sequentially from examples from a single class at a time, and specialized techniques are actively being developed to minimize this problem (Rusu et al., 2016; Kirkpatrick et al., 2017; Fernando et al., 2017).

We use a baseline Tacotron architecture specified in [8], where we use a GMM attention [9], LSTM-based decoder with zoneout regularization [10] and phoneme inputs derived from normalized text.

Figure 14: Examples of contexts that were misclassified as not impactful by QWEN3-30B, GPT-5.1, and O4-MINI under Crystal. The citing paper builds on or is inspired by the cited work, but the mention is brief or the cited paper is grouped with other references in the same citation.

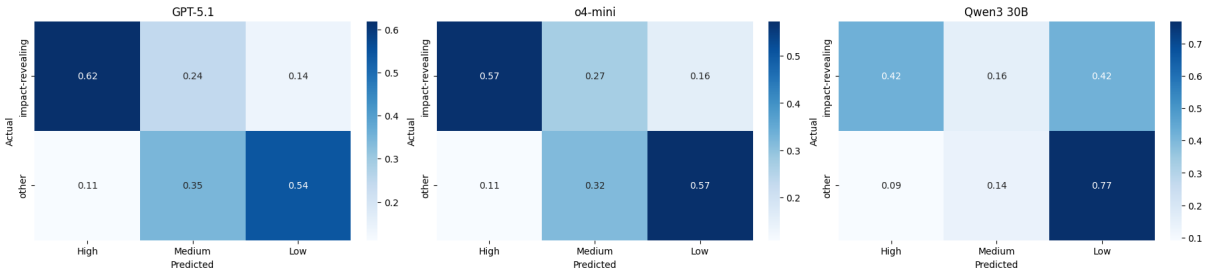

Figure 15: Confusion matrices per model for assigning impact labels with an ordinal regression model. Rows represent actual labels; columns represent predicted categories. GPT-5.1 achieves the best performance in identifying impactful citations and reducing false positive. However, we see a slight decrease 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 16: 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*.

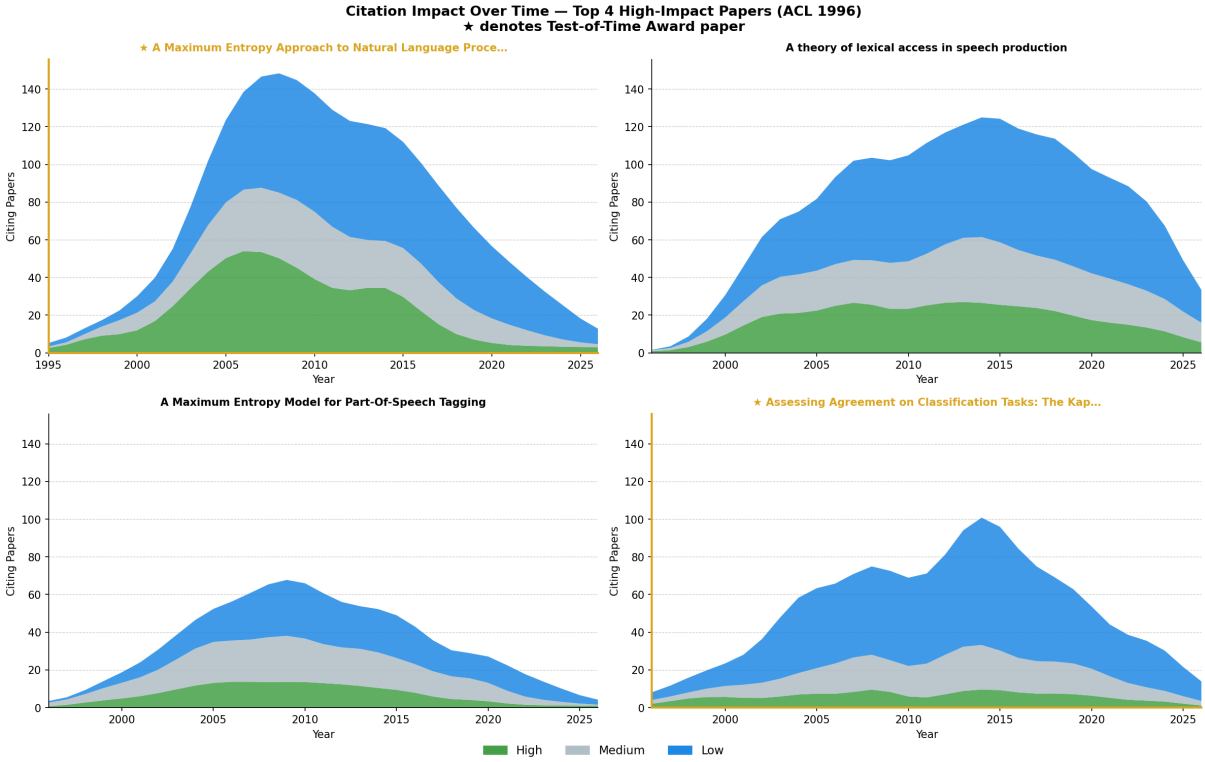

Figure 17: Yearly citation impact distribution for the top four ACL 1996 papers ranked by high-impact citations under Crystal. Each subplot stacks high-, medium-, and low-impact citations by year of the citing paper. Test-of-Time award winners (Berger et al. (1996), top-left; Carletta (1996), bottom-right) are marked with a ⋆ and highlighted with a yellow border. Berger et al. (1996) accumulates the largest volume of high-impact citations of the four papers, sustained across more than two decades. Carletta (1996) exhibits a stable but smaller stream of high-impact citations, with most of its citation volume concentrated in the low-impact category.

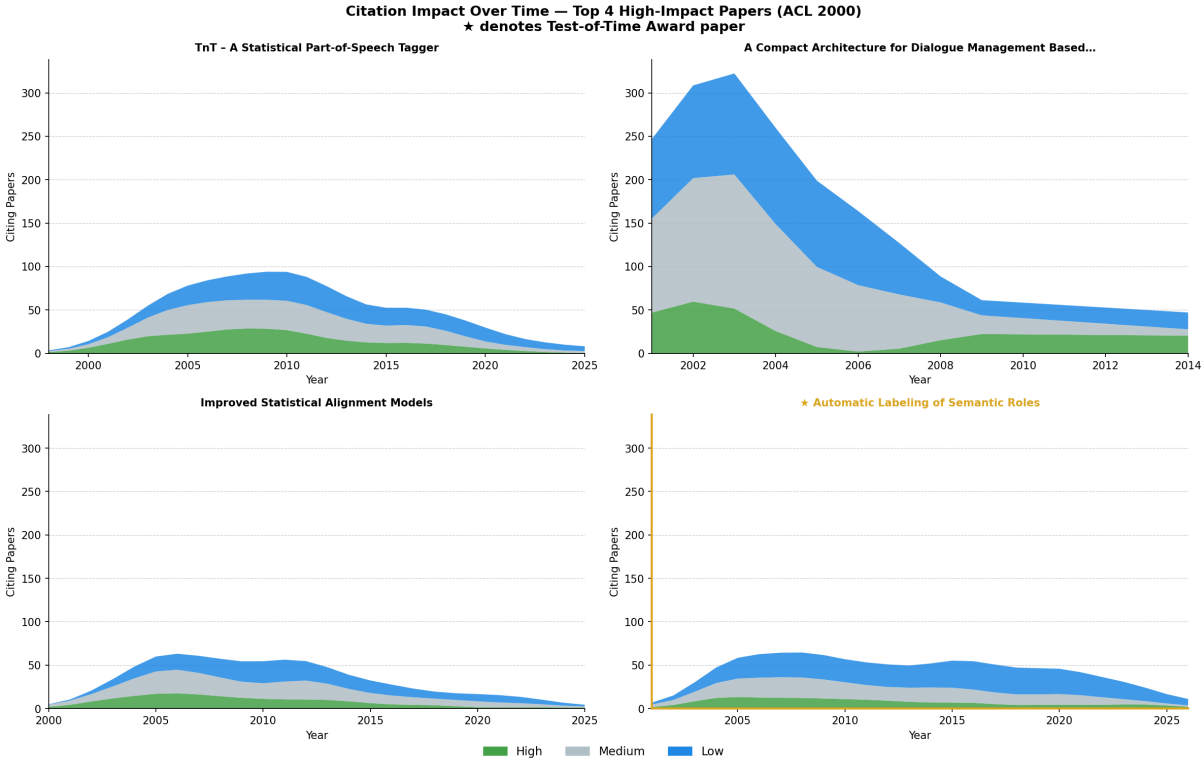

Figure 18: Yearly citation impact distribution for the top four ACL 2000 papers ranked by high-impact citations under Crystal (as shown in Figure 3). Each subplot stacks high-, medium-, and low-impact citations by year of the citing paper. The Test-of-Time award winner (Gildea and Jurafsky (2002), bottom-right) is marked with a ⋆ and highlighted with a yellow border. Its high-impact citation profile remains stable through 2025, while several non-awarded top-ranked papers peak earlier and decline more steeply.

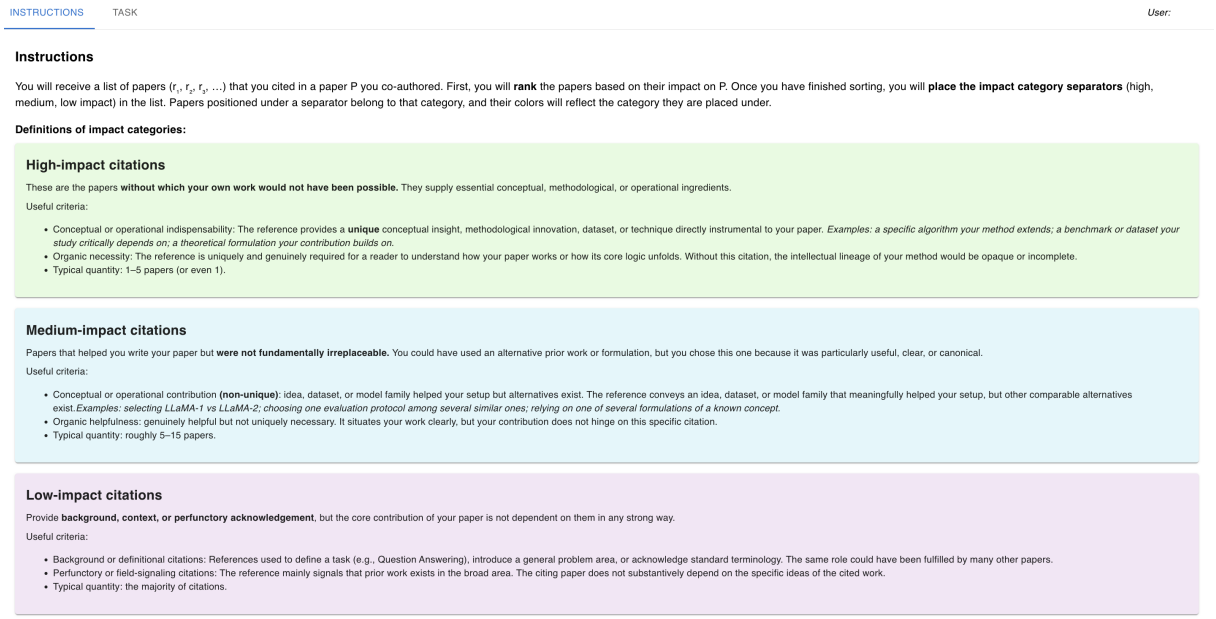

Figure 19: The main layout of the custom annotation interface used in the pilot study, showing the overall task setup and instructions for ranking references by impact.

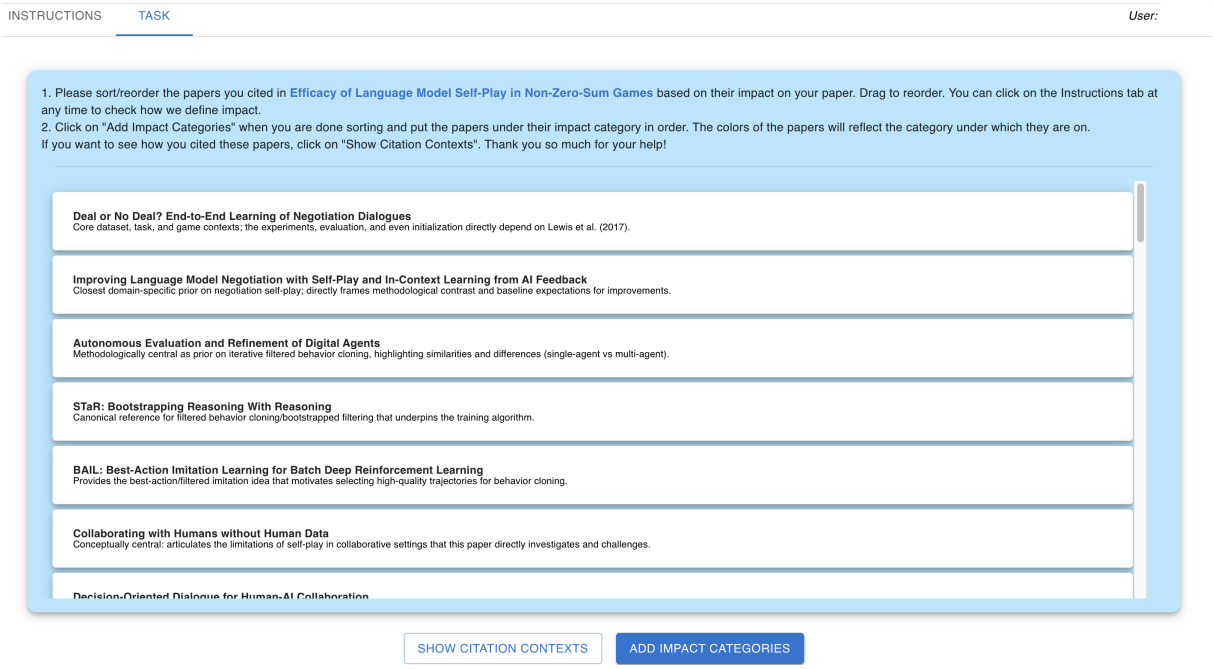

Figure 20: The full list of references from an annotator's paper displayed within the interface, allowing users to rank their references by dragging and reordering them according to the impact criteria.

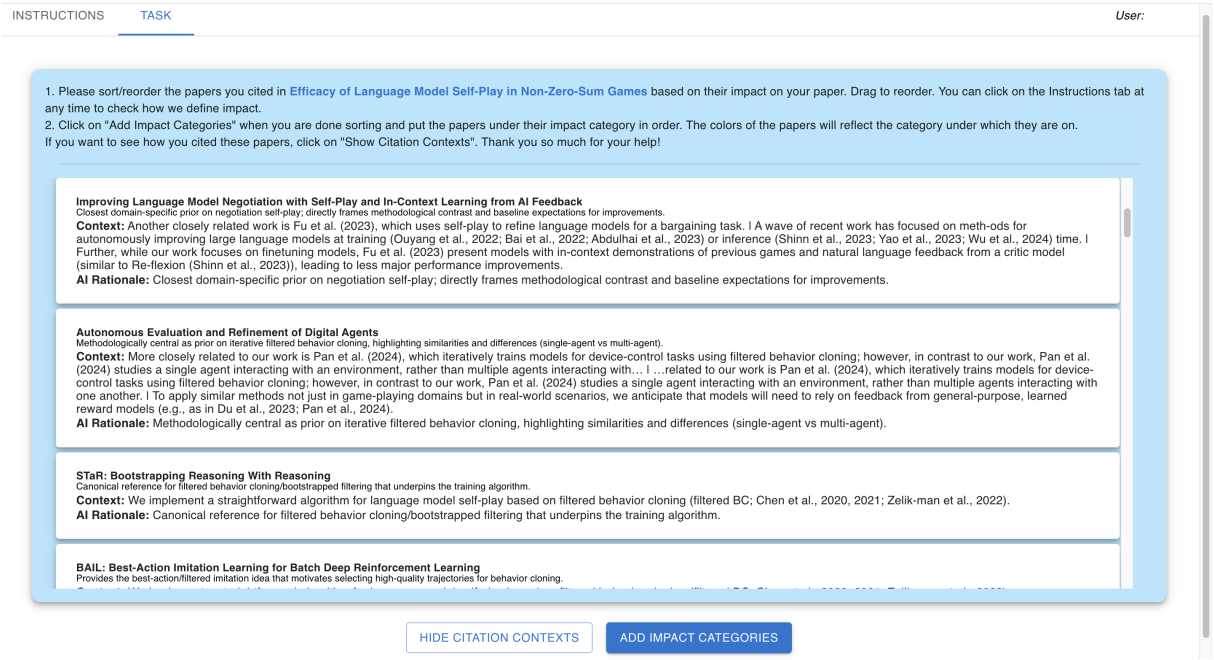

Figure 21: A feature of the interface that reveals the citation context when the "Show Citation Contexts" button is clicked, illustrating how each reference was cited in the paper.

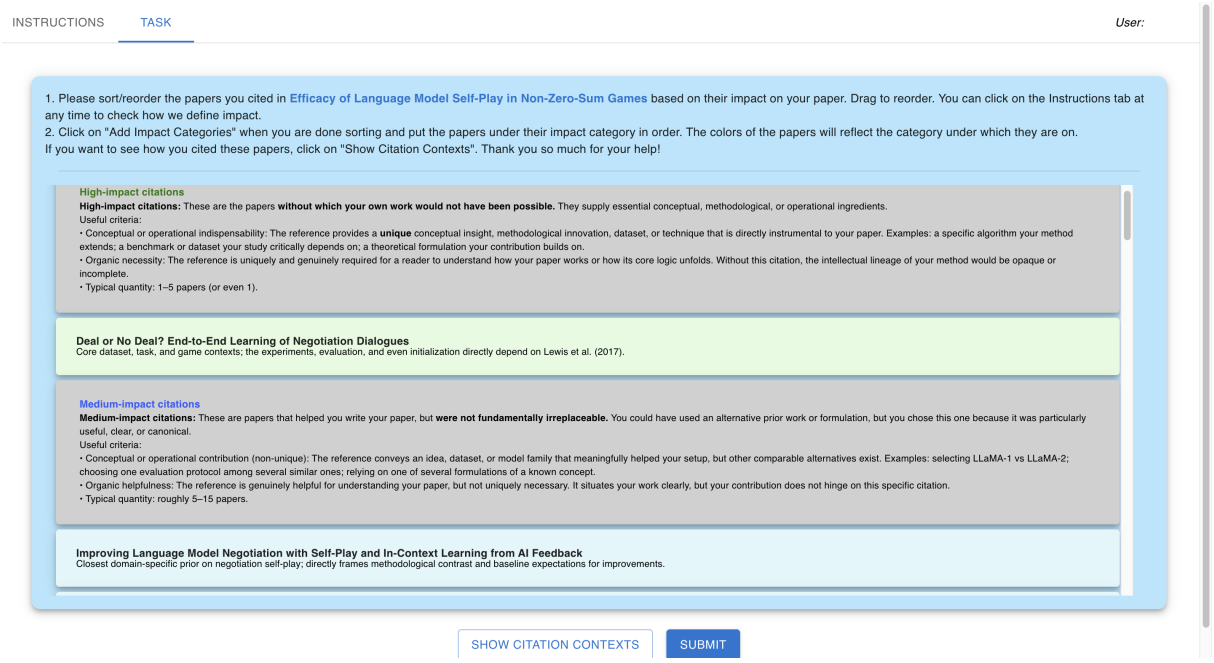

Figure 22: The final ranked list of references generated within the interface, showing the outcome of the annotation task with color-coded impact categories. Annotators submit their final ranked list of references.

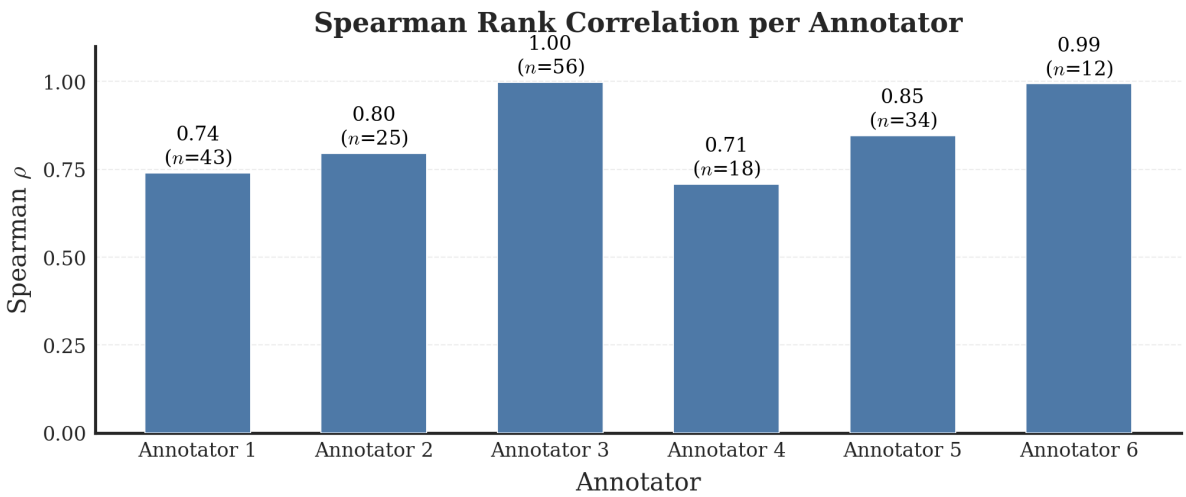

Figure 23: Shows the Spearman correlation between each annotator's ranking and the ranking generated by GPT-5.1 using the same prompt from our experiments. The n value above each bar represents the total number of references ranked in that paper. All correlation values exceeded 0.7, indicating strong agreement.