In-depth Research Impact Summarization through Fine-Grained Temporal Citation Analysis

Hiba Arnaout[1], **Noy Sternlicht**[2], **Tom Hope**[2,3], **Iryna Gurevych**[1]

[1]UKP Lab, TU Darmstadt and Hessian Center for AI (hessian.AI)
[2]School of Computer Science and Engineering, Hebrew University of Jerusalem
[3]The Allen Institute for AI (AI2)
www.ukp.tu-darmstadt.de

Abstract

Understanding the impact of scientific publications is crucial for identifying breakthroughs and guiding future research. Traditional metrics based on citation counts often miss the nuanced ways a paper contributes to its field. In this work, we propose a new task: generating nuanced, expressive, and time-aware impact summaries that capture both praise (confirmation citations) and critique (correction citations) through the evolution of fine-grained citation intents. We introduce an evaluation framework tailored to this task, showing moderate to strong human correlation on subjective metrics such as insightfulness. Expert feedback from professors reveals a strong interest in these summaries and suggests future improvements. Data and code are made available. [1]

1 Introduction

Citation counts are a common proxy for measuring the impact of research papers, but they offer only a shallow view that fails to capture *how* a paper has influenced subsequent work. A raw citation count does not reveal whether a paper was foundational, extended, critiqued, or merely mentioned in passing. To truly understand the impact of a research paper, we must go beyond simple counts and examine the *context* in which it is cited, analyzing how its ideas have been discussed, applied, and evolved over time. Manually tracking how a paper is discussed across a very large number of publications and diverse domains is infeasible due to the scale and complexity of the task. To address this, we introduce a new task: **research impact summary generation**, which aims to generate concise, time-aware narratives that reflect a paper's evolving scientific influence. These summaries can support practical use cases such as enhancing future research directions by helping researchers quickly assess a paper's relevance and legacy, or assisting funding agencies, hiring committees, and research administrators in making more informed evaluations of scholarly contributions. Figure 1 shows an example of an impact summary about a research paper (Koller and Sahami, 1997), published at The International Conference on Machine Learning in 1997. This paper's impact has followed a dynamic trajectory, initially used for its method

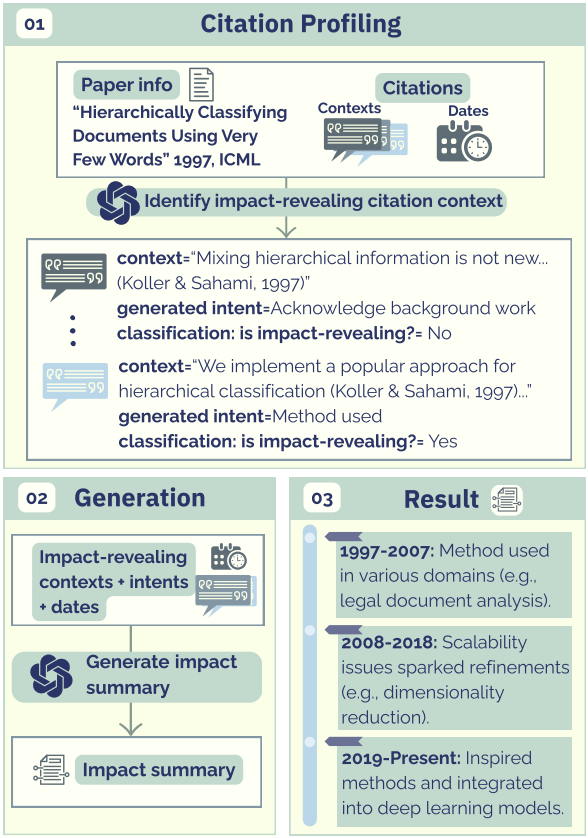

Figure 1: We propose a new task to summarize a paper's evolving impact over time, by analyzing *impact-revealing* citation contexts, reflecting both praise (confirming ideas) and critique (calls for correction). Our summary of Koller and Sahami (1997) reveals its impact trajectory—adaptation, critique, and rediscovery—offering deeper insight than its ∼1.4k cite count.

[1]https://ukplab.github.io/acl2026-generating-impact-summaries

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6749–6789
July 2-7, 2026

(1997–2007), later critiqued for scalability and accuracy limitations, then refined (2008–2018), and recently rediscovered as an inspiration for modern methods (2019–present). This example shows the need to look beyond the paper's ∼1.4k citations to understand the deeper story of how ideas gain relevance, face scrutiny, and find renewed significance.

We break the task of generating impact summaries for research papers into two subtasks. First, given all the citation contexts of a paper, we identify the "impact revealing" citations that directly interact with the paper, and their specific intents, whether in confirmation of its ideas (e.g., "method use") or in correction (e.g., "identifying limitations and proposing refinements"). Second, using *only* the impact-revealing citation contexts, their dates and intents, we generate an impact summary that captures the paper's influence over time. To the best of our knowledge, this is the first attempt to express scientific impact through time-aware textual summaries derived from citation context analysis. An overview is shown in Figure 1. To implement these subtasks, our method, in Section 2, uses in-context learning (ICL) with Large Language Models (LLMs) to generate fine-grained citation intents in the first stage. In the second stage, it filters for impact-revealing citation contexts based on the generated intents and feeds them to an LLM to generate semi-structured impact summaries under a pre-defined schema.

Existing work on scientific impact analysis predominantly focuses on citation and similar numerical counts as the main indicator of impact (Gu and Krenn, 2024), ignoring the exact reasons a paper has been cited, which can often be "unimpactful", e.g., citing a paper to acknowledge background work on a topic. Work on citation context analysis (Lauscher et al., 2022) studies the role citations play, but is based on coarse categories and is not directly intended for impact analysis. The goal of our new task is to create a narrative about a paper's impact using fine-grained citation intent analysis. Moreover, existing work typically emphasizes praise (confirmatory citations) (Zhang et al., 2024a; Valenzuela et al., 2015), but scientific progress relies on both confirmation and correction (Catalini et al., 2015), as ideas advance through adoption as well as critique and refinement. In this work, we ensure that our generated impact summaries capture both aspects of impact (when applicable), as shown in Figure 1. To support this broader view of impact, we extend the existing PST-Bench data set (Zhang et al., 2024a), which only focuses on "motivation" and "inspiration" intentions to consider an impact revealing citation, by introducing, in Section 2.5, a new ground-truth data set of 4000 citation contexts labeled for whether they reveal impact. Our extended dataset also covers not only additional intents of confirmatory citations (e.g., method use), but also correction citations (e.g., method refinement). For detailed comparisons with related work, see Appendix A and Section 5.

Finally, in the absence of gold-standard summaries for this novel impact generation task, we propose an automated evaluation framework to measure the trustworthiness and informativeness of our impact summaries (Section 2.4) and demonstrate a moderate to strong correlation with human assessments (Section 3.3). To further validate our approach, we collect expert feedback from professors on the usefulness of impact summaries about their own papers (Section 4), demonstrating the need for such summaries. For future research, we release the code and data used to develop this work.

2 Method

2.1 Definitions

We propose the task of generating time-aware impact summaries for research papers by identifying and analyzing impact-revealing citation contexts and their intents over time.

Definition 1. Citation context refers to the specific part of a paper p' where another paper p is mentioned, including the surrounding text that explains how p is relevant to p'.

Definition 2. Fine-grained citation intent refers to the specific, nuanced reason (the "why") behind citing paper p in paper p', as inferred from the citation context. Unlike predefined or categorical intent schemes, fine-grained citation intent is expressed in free-form text and does not rely on a fixed set of labels or a predefined taxonomy.

A few examples are in Table 1.

Definition 3. An **Impact-revealing citation intent** is a citation intent where the cited paper p has a direct influence on the citing paper p' through: (i) *Confirmation*, by adopting or building upon p; or (ii) *Critique*, by correcting or refining p. Citation intents that do not fall under these categories are considered non-impact-revealing.

Figure 1 shows the intent "acknowledge background work", for the context that simply refers to prior work in a general manner, without actively building upon it. Hence, this is not considered impact-revealing. In contrast, the intent "method used", describing the implementation of a popular approach is indeed impact-revealing.

Definition 4. A **scientific impact summary** of a research paper describes the impact paper p has had on subsequent research. In other words, how it has been directly used for both confirmation and critique. The summary is time-aware, taking into account the date of these citations, to track the evolution of its impact over time.

A sample impact summary is in Figure 1, complete examples are in Appendix G.6.

2.2 Identifying impact-revealing intents

We use ICL with LLMs to generate fine-grained intents and classify them as either "impact-revealing" or "other". To facilitate this, we manually craft a prompt that incorporates a set of examples covering various citation intents, and ensuring the model can learn from them to generate fine-grained intents and classify them correctly (prompt and examples are in Appendix B and Table 1, respectively). This prompt is designed to capture both the confirmatory and correction citations, allowing for a nuanced understanding of how the cited paper influences the citing paper. It aims to generate concise but expressive labels that clearly explain the reason behind the citation without being overly general or overly specific, e.g., "identifying knowledge gap in literature" (more examples in Appendix E.1). To examine the quality of our method, we conduct experiments in Sections 3.1 and 3.2.

2.3 Generating scientific impact summaries

To generate scientific impact summaries, we begin by identifying impact-revealing citation contexts through the generated fine-grained intents (as detailed in Section 2.2). To ensure the generation of consistent and easily comparable summaries by LLMs, we design a tailored prompt (see Appendix C) that incorporates our definition of scientific impact (see Definition 4). This prompt includes the impact-revealing citation contexts of a given paper, augmented with corresponding fine-grained intents and citation timestamps (i.e., publication years). We evaluate the effectiveness of this approach through an ablation study (Section 3.3) and an expert review (Section 4).

2.4 Evaluation metrics

Evaluating our summaries is challenging due to the lack of gold-standard datasets for this novel task. We explored sources like Wikipedia articles on influential papers [2] and Test of Time Award pages[3], but these were insufficient, as they cover few papers and focus on content or general praise rather than usage-based insights. Web searches (e.g., "*what is the impact of paper X*") mostly returned the original paper or explanatory blog posts. Given these limitations, we introduce a new automated, reference-free evaluation framework that measures trustworthiness and informativeness, and correlates moderately to strongly with human judgment (Appendix G.4).

Trustworthiness. Measuring summary correctness is challenging due to the large gap between input and output lengths, compressing thousands of citation contexts into a few sentences. Without a gold reference, determining what content should be included or excluded becomes difficult. Additionally, time-augmented impact descriptions must be accurate both in content and assigned time period. Inspired by evaluation frameworks in the RAG setting (Ru et al., 2024; ES et al., 2024; Asai et al., 2024a), which aligns with our query-based summarization approach, we define: **(1) Faithfulness:** Our time-aware faithfulness metric examines whether details in the impact summary, i.e., the impact description elaborating on the dominant intent of an impact period, is grounded in the paper's citation contexts (i.e., the LLM's input context). We first split the impact summaries into impact descriptions representing different time periods (as defined in the output schema in Appendix C). Next, we instruct an LLM to verify each against citations from the corresponding period. Note that a summary impact description is verified against many citations at once and not in a simple pairwise entailment, since a single impact description can encompass information from multiple citations or discuss trends, which are only possible to verify by inspecting many sources (e.g., "*the paper is frequently used for...*"). The evaluator LLM is instructed to assess the impact description against the provided

[2] E.g., Article for (Vaswani et al., 2017): https://en.wikipedia.org/wiki/Attention_Is_All_You_Need

[3] E.g., WSDM 2022 Test of Time Award Winner: https://www.wsdm-conference.org/2022/timetable/event/wsdm-awards-program-test-of-time-presentation/

Citation context	Intent	Class
..we apply a minimization process [1].	use of minimization methodology	impact-revealing
Chiu and Nichols (2016) introduced convolutional neural networks for NER	background about NER methods	other
Quirk and Poon (2017) and Peng et al. (2017) build two distantly supervised datasets without human annotation, which may make the evaluation less reliable.. we present .. a large-scale human-annotated document-level RE dataset..	criticizing existing datasets and proposing a better one	impact-revealing

Table 1: Training examples for fine-grained intent generation and classification. Full table in Appendix E.1.

citations, determine its faithfulness, and justify the decision with a list of supporting citations. This procedure resembles the task of a human annotator evaluating the faithfulness of a machine-generated text (Kim et al., 2024). **(2) Coverage:** Coverage is defined as the ratio of impact-revealing intents mentioned in the summary. For example, if a paper is cited for its method use, for inspiring new research, and for exposing limitations and refinement, the summary should reflect all these intents, including relevant details. Inspired by the evaluation rubrics for coverage proposed in (Asai et al., 2024a), which defines it as the topics (themes) mentioned (i.e., the diversity of impact-revealing intents in our case), we develop a two-step evaluation. In the first step, we cluster the list of fine-grained intents under similar topics, e.g., "*method used in the legal domain*". These labels might have slightly different granularity, which depends on how much details the members (i.e., intents) of that cluster offer. Next, the evaluator LLM takes both the impact summary and the list of cluster titles that this summary should ideally cover, and returns a list of topics that were actually covered. Finally, we divide the number of intents mentioned in the summary by the total number of impact-revealing intent clusters. Prompts for faithfulness and coverage are in Appendix D. **(3) Citation Year Compliance:** We implement this metric as a script that flags citations falling outside the target impact period.

Informativeness. We also assess the informativeness of impact summaries, as trustworthy ones may still be unhelpful if they lack meaningful insights. Grounded in our definitions (Section 2.1), the informativeness of an impact summary depends on its ability to clearly describe the paper's direct influence on other papers, track the types of influence (intents) over time, and provide details inferred from the citation contexts. We define the following metrics: **(1) Insightfulness** measures how well the impact summary articulates the paper's direct influence on other works. **(2) Trend Awareness** evaluates the extent to which the impact summary identifies and distinguishes between different periods of the paper's influence over time. **(3) Specificity** assesses whether the impact summary includes concrete examples, such as techniques, influenced by the paper. To systematically evaluate these criteria, we leverage G-Eval (Liu et al., 2023a), a framework that uses LLMs with chain-of-thought (CoT) reasoning to assess the impact summaries. By integrating structured evaluation steps, G-Eval enables LLMs to reason step-by-step, improving the reliability and depth of assessments. We transform our criteria into prompts in Appendix D.

2.5 A new dataset for identifying impact-revealing citation contexts

To construct a new dataset for classifying citation contexts as "impact-revealing" or not, we start by building on the PST-Bench dataset (Zhang et al., 2024a), which focuses on positive influence through "inspiration" and "motivation". We select 1k impact-revealing citations from PST-Bench. Next, we augment this by first manually crafting textual patterns for both confirmation and correction intents, and use GPT-4 to generate variations. After crawling 200k random citation contexts (from the Semantic Scholar Academic Graph (S2AG)[4]) and checking if they match any of the textual patterns, we randomly sample 1k instances of impact-revealing contexts (a total of 2k impact-revealing with the original selection from PST-Bench, covering both confirmatory and correct citations). Additionally, we sample 2k non-impact-revealing citations ("other") directly from the PST-Bench dataset (references that were *not* annotated as influential).

[4] https://www.semanticscholar.org/product/api

We also ensure none of these match any of our impact-revealing citation patterns. This results in a balanced 4k citation context dataset, which we release alongside this work. To validate the quality of our automatically collected impact-revealing examples, we manually reviewed a random sample of 100 instances and found that 90% were correctly labeled. The remaining 10% consisted of edge cases where surface patterns, such as the phrase "motivated by" (e.g., Voters are motivated by partisan social identities... (Greene 2004)), appeared to indicate impact but did not actually reflect the citation's true intent. Details on the construction and textual patterns are in Appendix F.1.

3 Experiments

3.1 In-context-learning for fine-grained intent generation

Zero-shot vs. ICL. To evaluate LLMs' ability to generate fine-grained intents and classify them as "impact-revealing" or "non-impact-revealing", we manually select 10 citation contexts (Table 1 and Appendix E.1). Our selection criteria was based on diversity, a set that contains impact-revealing (praise, critique) and non-impact-revealing ("other") intents. The "other" class includes incidental mentions or contexts lacking sufficient information to understand the reason behind it. We use our prompt (Appendix B) with and without these examples to generate and classify intents, using GPT4o-mini, of 200 randomly sampled citation contexts. Two annotators (a postdoc and a PhD student) assess: (i) whether the classification as impact-revealing is correct, (ii) whether the generated intent accurately reflects the citation reason, and (iii) whether the intent is concise, as LLM results can be highly verbose. Results, including per-field metrics and qualitative examples, are in Appendix E.2. We find that including examples significantly improves performance across all metrics, with gains of 29%, 55%, and 39% in precision, recall, and F1, respectively. Inter-annotator agreement shows a substantial Cohen's kappa score of 0.68 on impact-revealing intent classification.

Effect of different number of shots (K). We find that performance improves significantly as the number of shots increases, plateauing around K = 50 with strong metrics (recall: 0.94, F1: 0.92); see Appendix F for details.

How does citation intent vary across fields? Table 2 presents an analysis of citation intents

Cited Papers	Citation Intents (%)					
	PS		MD		CS	
All	65	35	53	47	34	66
Recent	65	35	57	43	58	42
Older	65	35	52	48	33	67
Highly cited	66	34	36	64	38	62
Less cited	64	36	56	44	32	68

Table 2: Recent = the last 5 years, Highly cited = top 20 % by citation count in our dataset; orange for `impact-revealing`, light blue for `other`; PS= psychology, MD = medicine, CS = computer science.

across 70k citation contexts from psychology, medicine, and computer science papers, with further breakdowns by recency and citation counts. Overall, psychology citations tend to lean towards impact-revealing, while computer science skews more toward non-impact-revealing ("other") citations—except in more recent papers—and medicine shows a more balanced distribution. We observe that citations in psychology papers often exhibit a stronger subjective tone, especially in correction citations, such as: "*some of those assumptions have been controversial*" and "*researchers disagree about whether the kinds of behaviors measured by particular implicit tests should be considered indicators of attitudes or something else*". In contrast, computer science shows a notable shift in recent years toward more impact-revealing citations, likely driven by the novelty and immediate influence of AI research before transitioning into the *legacy* phase, where citations become more background-oriented. This legacy effect is particularly evident in the highly cited subsets.

3.2 Can existing intent classifiers detect impact-revealing citations?

Comparison with existing methods. We compare our classifier against the following popular intent classification methods: (i) Meaningful Citations (Valenzuela et al., 2015): a supervised machine learning method to classify citation intents into meaningful or non-meaningful, by leveraging context and citation metadata; (ii) Structural Scaffolds (Cohan et al., 2019): a multi-task model which enhances the main task by integrating auxiliary tasks; (iii) Multi-cite (Lauscher et al., 2022): a multi-label classification method that predicts multiple intents using a fine-tuned SciBERT (Beltagy

Intent classifier	P	R	F1	Acc
baseline=random	0.54	0.51	0.52	0.50
baseline=always-impact-revealing	0.53	**1.0**	**0.69**	0.53
Structural Scaffolds (Cohan et al., 2019)	0.55	0.44	0.49	0.51
Meaningful Citations (Valenzuela et al., 2015)	0.72	0.46	0.56	0.62
Multi-cite (Lauscher et al., 2022)	0.59	0.41	0.48	0.53
Ours	**0.74**	0.65	**0.69**	**0.69**

Table 3: Results for intent classification ("impact-revealing" or "other"). Our classifier is the best at distinguishing between citations that help in understanding the impact of a paper, and those that serve a more general purpose, such as background references or standard acknowledgments.

Intent classifier	Confirmatory	Correction
Structural Scaffolds	0.83	0.50
Meaningful Citations	0.41	0.65
Multi-cite	0.81	0.58
Ours	**0.88**	**0.98**

Table 4: F1 scores for impact-revealing citations, broken down by confirmatory and correction categories.

et al., 2019). We map their intent classes to either "impact-revealing" or "other". More details are in Appendix F.2. For our method, we use the ICL prompt, with K=50 and LLM=GPT4o-mini. For every test instance, we run our prompt 3 times (shuffling the order of the shots), and take a majority vote for the classification. Note that the results of the 3 runs have a full agreement of 72%. By full agreement, we mean all three runs predicted the same class, indicating stability to in-context example order.

Results. Numerical results are reported in Table 3. Our method demonstrates the best performance, with the most significant improvement in recall, outperforming the second-best method by 19 percentage points. This is particularly crucial for our goal in generating impact summaries from impact-revealing citations because the highest recall in this context means that our method ensures that fewer meaningful citations are missed. The qualitative results are shown in the appendix F.2. For a deeper analysis on the performance of these classifiers on confirmatory vs. correction citations, we report the F1 scores on these splits in Table 4. The table shows that our strong performance is largely driven by the intent classifier's effectiveness at identifying impact-revealing citations with correction-related intents (e.g., method refinement or highlighting research gaps). To ensure a fairer comparison with prior methods, we restrict the results in Table 4 to citations from computer science papers, aligning with the domain used to train those models.

3.3 Ablation study for impact summary generation

Data: papers, citations, intents. We select 105 papers from the Semantic Scholar Academic Graph (S2AG)[5], from psychology, medicine and computer science (35 each), published between 1974 and 2022 at top venues in their respective fields. The citation count for each article ranges from $\sim$ 500 to $\sim$ 5000. We collect citations and their contexts using the citation lookup feature in the S2AG API. We first run our fine-grained intent generator (with K=50) and classifier for every citation context. This amounts to 70k citation contexts with their generated intents and their classes.

Prompt variants. We experiment with different inputs to assess their effect on the summary's quality, focusing on: (1) Citation contexts: including all, only impact-revealing, or none, to test what the LLM infers without citations and how varying context types affect the summary. Although full citation contexts can enrich summaries, they can also introduce noise from incidental references (Liu et al., 2024). (2) Intents: whether to include the intents alongside citation contexts. Table 5 lists all variants. In total, we generate 945 impact summaries[6]. Appendix C shows the prompt. For intent generation, we choose GPT-4o-mini due to its lower cost. For both impact generation and evaluation, the more advanced tasks, we use GPT-4o.

Results. Table 5 shows the results for all metrics. Even though it shows a good understanding of how the paper was cited, likely due to the LLM's training data having access to a large corpus of scholarly articles, the baseline (no-knowledge) variant

[5] https://www.semanticscholar.org/product/api

[6] 1 variant without citations + (4 citation-based variants * 2 orderings) * 105 papers

Prompt variant		Trustworthiness				Informativeness		
Citations	**Intents**	**Faith.**	**Cov.**	**Cov.@3**	**Cyc.**	**Insi.**	**Trend.**	**Spec.**
None	✗	0.77	0.25	0.58	n/a	0.70	0.94	0.75
All	✗	0.83	0.32	0.74	0.55	0.80	0.95	0.85
All	✓	0.84	0.32	0.73	0.48	0.80	0.97	0.86
Impact-rev.	✗	0.87	0.33	0.73	**0.59**	0.80	0.96	0.87
Impact-rev.	✓	**0.88**	**0.34**	**0.75**	0.56	**0.83**	**0.98**	**0.88**

Table 5: Ablation results - Faithfulness: **Faith.**, Coverage: **Cov.**, Citation Year Compliance: **Cyc.**, Specificity: **Spec.**, Insightfulness: **Insi.**, Trend Awareness: **Trend.** These results show that adding impact-revealing contexts and their intents has an improvement on almost all metrics.

is the least faithful, and we observe that prompts with only impact-revealing citations generate more faithful summaries. This suggests that longer contexts might induce hallucinations in the generation process. Interestingly, providing the citation intents also gives a slight boost to the faithfulness of the summaries, possibly because the intents encourage wording that aligns with the input paper's citations. Coverage is especially challenging, requiring a balance of completeness, relevance, and conciseness. Our best variant still outperforms the baseline by 9%, showing that the LLM can capture key themes. Some inputs include over 100 distinct themes, explaining modest overall scores. Focusing only on the 3 most frequent themes (largest clusters), coverage rises significantly, as our best variant reaches 0.75 (+17%), highlighting its strength in capturing core impact. To assess the coverage of *meaningful low-frequency themes*, we analyze long-tail impact across 10 papers; Appendix G.2 shows up to 50% coverage when intents are included. Variants using only impact-revealing citations show higher citation year compliance; adding citation intents lowers it, likely due to greater focus on intent over timing. See Appendix G.1 for examples and Section 7 for discussion. All variants show high trend awareness, with a slight edge for the impact-revealing + intents variant, showing LLMs can recognize distinct impact periods. This variant also improves insightfulness and specificity by 13% over the no-knowledge baseline. Qualitative examples are in Appendix G.1. G-Eval, via DeepEval[7], also offers reasons for each score (Appendix G.5). We visualize sample summaries in Appendix G.6. To examine whether results change depending on the paper's field, we report results per field in Appendix G.3. Finally, Appendix G.7 show insights into how intent misclassification can impact the quality of the generated summaries.

Beyond GPTs. We also generate impact summaries using Qwen and Gemini, to ensure model-agnosticism. Quantitative and qualitative results are provided in Appendix G.8. Although GPT-4o served as our primary model, both Qwen and Gemini demonstrated strong performances, with Gemini in particular standing out in citation year compliance (0.93) and faithfulness (0.96).

LLM-Human correlation We compute the human-LLM correlation, which shows moderate to strong relationships for the Spearman and Kendall-Tau metrics. Details are provided in Appendix G.4.

3.4 Practical applications

We showcase the usability of our work in two practical applications. First, by analyzing 10 diverse research topics, such as LLMs for code generation, open information extraction, commonsense knowledge mining, and comparing papers with similar citation counts, it shows that papers can have very different types of impact. For example, in open information extraction, some papers are cited to highlight limitations or motivate new work (Cui et al., 2018), while others are cited mainly for their methods (Gashteovski et al., 2017; Han et al., 2019). More details and examples are in Appendix I.1. Second, we explore generating author-level impact summaries by aggregating LLM-generated paper-level summaries for an author's top-cited papers. Examples for two senior NLP researchers are shown in Appendix I.2. This method could be extended to institutions and venues, which we leave for future work.

Comprehensive details of the models and their configurations for each task/component are provided in Appendix H to support reproducibility and attribution.

[7] https://github.com/confident-ai/deepeval

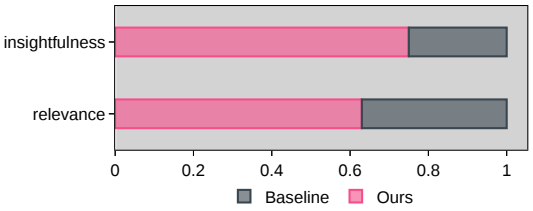

Figure 2: Pairwise comparison. *Relevance*: Which summary better reflects the paper's actual impact? *Insightfulness*: Which summary offers more valuable or novel information about the paper's usage? Professors consider our summaries more relevant and insightful.

4 Human Evaluation

Setup. To gain deeper insights into the quality of our impact summaries and collect expert feedback for potential future improvements, we conduct a user study with 9 university professors from diverse backgrounds [8]. Each expert reviews impact summaries of *their own papers* through a two-part evaluation: (1) a **pairwise comparison**, where they choose the better summary based on *relevance* and *insightfulness*, comparing the no-knowledge variant against our best variant. To prevent bias, we alternate the positions of the impact summaries. (2) a **perceived usefulness assessment**, where experts rate their agreement with given statements about an impact summary (generated by our best variant), using a 1-5 Likert scale. The professors evaluated the impact summaries for 82 papers and offered open-ended feedback.

Results. Figure 2 shows results of the pairwise comparison. Impact summaries generated by our variant[9] demonstrate greater relevance (63%) and insightfulness (75%) than the baseline[10]. Results of the perceived usefulness, in Figure 3, show that approximately 60% of professors agreed that the summaries provided an appropriate level of detail (clarity) and offered novel insights into how their papers were used, i.e., information not readily available elsewhere. Notably, for papers in the top 10% based on impact-revealing citations, agreement on clarity and informativeness rose to 75%. Finally, we ask the evaluators for open feedback. A couple professors found certain summaries too generic, failing to highlight specific strengths and limitations. They also pointed out that summaries based on citations from lesser-known conferences may not fully reflect a paper's true influence, and the quality of summaries often correlates with the size of impact-revealing citation context (a point further validated in Figure 3 (b)). Additionally, one professor noted that for some papers, it was difficult to assess impact because the work was collaborative, and they were only familiar with certain aspects of the paper's contributions, while their co-authors might be more familiar with other aspects. Other professors suggested improvements in both content and structure: they recommended incorporating citation counts within the text, which we plan to implement in the future, and enhancing the structure of our semi-structured summaries by adding elements like bullet points to improve readability.

[8] Nationalities: 2 German, 2 British, 2 Chinese, 1 American, 1 Czech, 1 Albanian; genders: 5 male, 4 female; research areas: AI, NLP, knowledge graphs, databases and information systems, computational social sciences, psychology and brain sciences.

[9] Impact-revealing citations with intents.

[10] no-knowledge variant

5 Related Work

Scientific impact analysis. Studying scientific impact is vital for guiding research and recognizing contributions. While prior work relies on citation counts and related metrics (Hutchins et al., 2016; Bos and Nitza, 2019; Siudem et al., 2020; Min et al., 2021; Wahle et al., 2023; Gu and Krenn, 2024), impact cannot be reduced to a single number (Zhu et al., 2015). Few studies use citation context to assess scientific impact, and those that do focus on classifying citation intent individually rather than aggregating them in summaries (Jurgens et al., 2018; Valenzuela et al., 2015). Our approach uses fine-grained citation context analysis to capture and aggregate specific citation reasons across impact-revealing citations. Other work links impact to novelty of the paper's content (Rüdiger et al., 2021; Shi and Evans, 2023; Arts et al., 2021), but overlooks how papers are used. Most also equate impact with praise (Zhang et al., 2024a; Valenzuela et al., 2015; Zhu et al., 2015), whereas we distinguish confirmatory from correction citations, offering a fuller view. As detailed in Appendix A, we are the first to generate time-aware, multifaceted impact summaries from fine-grained citation analysis.

Citation intent prediction. Prior work on citation intent classification uses pre-defined categories (e.g., method use, background) with early methods relying on supervised learning and linguistic features (Teufel et al., 2006; Jurgens et al., 2018; Tuarob et al., 2020). Later approaches applied multi-task learning (Cohan et al., 2019), while (Lauscher et al., 2022) introduced a multi-label method using

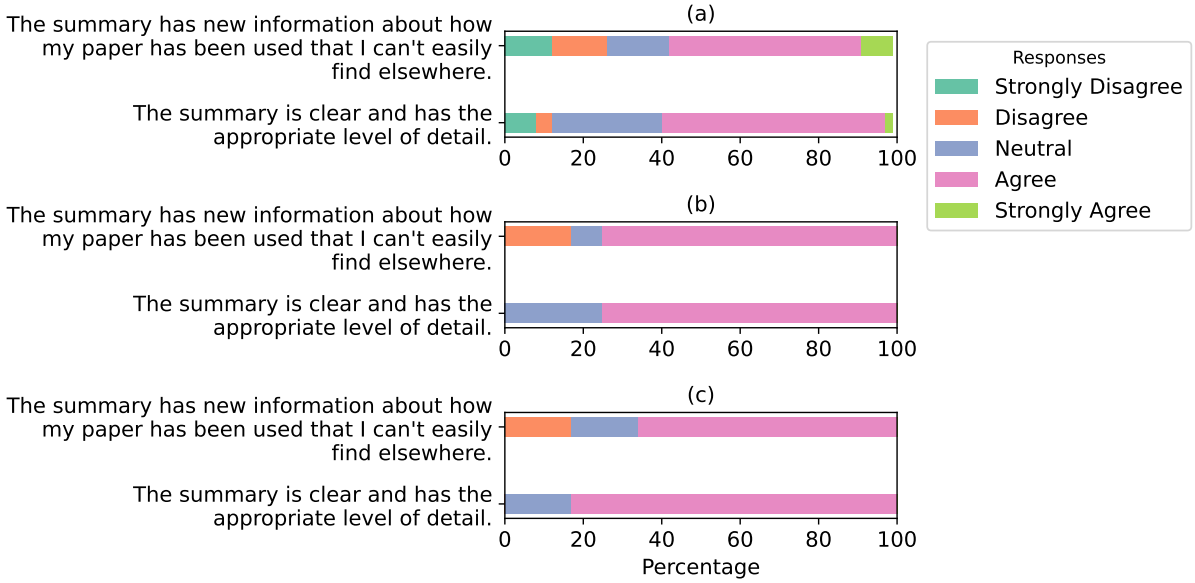

Figure 3: We show the overall results in (a), and results for two subsets of papers, namely papers with the most impact-revealing citations in (b), and papers with the highest number of citations in (c). Both subsets show impact summaries that are rated higher, likely because higher counts of impact-revealing citations and citations in general allow for richer contexts, which provide the model with clearer signals about the paper's impact, leading to more accurate and insightful summaries.

fine-tuned SciBERT (Beltagy et al., 2019). These rely on coarse categories; in contrast, we use in-context learning to generate fine-grained, free-form intent descriptions (Table 11), offering more precise insights, an approach recently shown to aid related work generation (Sahinuç et al., 2024).

Multi-document summarization in the era of LLMs. Recent work in query-focused multi-document summarization uses LLMs to generate context-aware summaries (Roy and Kundu, 2024), with approaches like graph-based QA over private corpora (Edge et al., 2024) and retrieval-augmented summarization (Zhang et al., 2024b). In the scientific domain, prior efforts focus on summarizing papers for related work (Sahinuç et al., 2024; Lu et al., 2020; Chen et al., 2021; Wu et al., 2021) or reviews (Kasanishi et al., 2023). In contrast, we generate free-text summaries of scientific impact using a novel time-aware, multi-document approach.

LLM Evaluators. Retrieval-Augmented Generation (RAG) systems condition generation on external information retrieved from user queries (Yu et al., 2024; Gao et al., 2023), but evaluating them is challenging due to multiple stages like chunking and retrieval. Tools like RagChecker (Ru et al., 2024) use entailment-based metrics to assess faithfulness, while RAGAs (ES et al., 2024) offers reference-free evaluation of relevance and faithfulness. Inspired by these, we assess our impact summaries' faithfulness and coverage by validating them against the citation context corpus. While inspired by RAG methods, our approach introduces a distinct task focused on impact and temporal dynamics. By targeting impact-revealing citations and modeling shifts over time, we support more nuanced, citation-centric summarization. The LLM-as-a-judge paradigm, used for tasks like QA and writing evaluation (Zheng et al., 2023; Shao et al., 2024; Asai et al., 2024a), enables scalable assessment of complex qualities. Given the novelty of scientific impact summarization, we adopt G-Eval (Liu et al., 2023b), a CoT-based framework, to assess informativeness.

6 Conclusion

This work introduces a novel approach for generating time-aware impact summaries of research papers by analyzing evolving fine-grained confirmatory and correction citation intents. Our approach goes beyond citation counts, offering a more nuanced understanding of a paper's impact over time. Expert evaluation highlights the value and potential for refinement of our summaries. We release our data and code to support research in this new task.

7 Limitations

Language. Our work focuses on English papers, as it is the dominant language in most research fields. Extending this approach to a cross-lingual setting is a promising future direction.

Human evaluation. Constructing a large-scale user study is challenging because assessors need deep knowledge of the papers' impact. We focused on experts evaluating their own work, but this limited our pool to 9 experts, as highly cited papers are typically authored by busy professors. We chose these experienced authors with multiple impactful papers to maximize the number of impact summaries evaluated per person.

Trustworthiness evaluation. Without restrictions, our approach shows unsatisfying coverage (0.34) and citation year compliance (0.59), compared to other metrics such as faithfulness. Regarding the year compliance, citation phrases often include multiple references accompanied by publication years, which can confuse the model and lead to incorrect citation year associations. In the future, we plan to remove or distinguish unrelated, potentially distracting numbers from the citation context (e.g., using well-crafted heuristics or prompt-based methods). For coverage, the LLM may focus on certain prominent themes while underrepresenting others, especially with complex topics. We briefly reported the coverage numbers for the most frequent themes, which showed a great improvement. In the future, we plan to explore this further, for instance, by ranking the importance of the themes and factor this into the coverage computation at different ranks. However, for this study, we decided to test full coverage to make sure the model captures a wide range of themes, without focusing on any specific ones, to check its overall understanding of the paper's impact. Finally, although checking potentially hundreds of citation contexts/intents at once raises concerns, the strong LLM-human agreement reported in Appendix G.4 suggests such issues are unlikely.

Experimenting with more LLMs. In this paper, we prioritize introducing and exploring the new task in depth, which is why we only experimented with the long-context model GPT-4o. We leave testing a wider range of models for future work. Although using the same LLM for both generation and evaluation can introduce bias, it will also promote consistency in task interpretation, and since our evaluation is not preference-based but scoring-based, bias is unlikely to affect results.

Broader spectrum of scientific impact. While our study operationalizes scientific impact primarily through confirmation and correction, we acknowledge that real-world impact spans a broader spectrum (e.g., parallel development). Our framework is designed to be extensible, and future work can enrich the training examples with additional categories to capture a more nuanced landscape of scholarly influence, though we note that such categories may also introduce additional challenges and interpretive ambiguity.

Data reliability and quality controls. We rely on the S2AG resource from Semantic Scholar, whose scale, coverage, active curation, and provision of citation contexts and structured metadata make it a uniquely reliable foundation for studying citations. While the data is generally high quality, our pipeline adds safeguards such as removing duplicate contexts, discarding citations lacking sufficient information, and enforcing temporal consistency to ensure robustness. We acknowledge that very large-scale datasets may have completeness limitations, but such issues are common across bibliographic resources, and further citation-quality filtering is a natural future enhancement rather than a prerequisite for the validity of our current findings.

Ethics Statement

The data used in this study is publicly accessible and provided under open licenses, ensuring long-term reproducibility and building upon our work. For the expert study, we do not disclose the names of participating professors or publish individual votes and verbatim feedback. Instead, we report aggregated votes and rephrased feedback for suggestions of future improvements. While this study explores a promising new task, we caution that inaccuracies in the generated summaries could mislead readers about a paper's actual impact. Therefore, we do not recommend using our method directly and without human verification for critical decisions, such as academic hiring or grant allocations.

Acknowledgments

This work has also been co-funded by the LOEWE Distinguished Chair "Ubiquitous Knowledge Processing", LOEWE initiative, Hesse, Germany (Grant Number: LOEWE/4a//519/05/00.002(0002)/81 and by the European Union (ERC, InterText, 101054961).

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Sam Arts, Jianan Hou, and Juan Carlos Gomez. 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research policy*, 50(2):104144.

Sam Arts, Nicola Melluso, and Reinhilde Veugelers. 2025. Beyond citations: Measuring novel scientific ideas and their impact in publication text. *Review of Economics and Statistics*, pages 1–33.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Daniel S. Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024a. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *CoRR*, abs/2411.14199.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024b. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Bilal Aslam, Wei Wang, Muhammad Imran Arshad, Mohsin Khurshid, Saima Muzammil, Muhammad Hidayat Rasool, Muhammad Atif Nisar, Ruman Farooq Alvi, Muhammad Aamir Aslam, Muhammad Usman Qamar, et al. 2018. Antibiotic resistance: a rundown of a global crisis. *Infection and drug resistance*, pages 1645–1658.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Guillermo Bernal, María I Jiménez-Chafey, and Melanie M Domenech Rodríguez. 2009. Cultural adaptation of treatments: A resource for considering culture in evidence-based practice. *Professional Psychology: Research and Practice*, 40(4):361.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 549–556. AAAI Press.

Arthur R. Bos and Sandrine Nitza. 2019. Interdisciplinary comparison of scientific impact of publications using the citation-ratio. *Data Sci. J.*, 18:19.

Christian Catalini, Nicola Lacetera, and Alexander Oettl. 2015. The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences*, 112(45):13823–13826.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6068–6077. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3586–3596. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 407–413. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander M. Rush. 2019. Commonsense knowledge mining from pre-

trained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1173–1178. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130.

Shahul ES, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*, pages 150–158. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6556–6576. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. 2017. Minie: Minimizing facts in open information extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2630–2640. Association for Computational Linguistics.

Xuemei Gu and Mario Krenn. 2024. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *CoRR*, abs/2402.08640.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. Opennre: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 169–174. Association for Computational Linguistics.

Jiangen He and Chaomei Chen. 2018. Temporal representations of citations for understanding the changing roles of scientific publications. *Frontiers Res. Metrics Anal.*, 3:27.

B Ian Hutchins, Xin Yuan, James M Anderson, and George M Santangelo. 2016. Relative citation ratio (rcr): a new metric that uses citation rates to measure influence at the article level. *PLoS biology*, 14(9):e1002541.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

David Jurgens, Srijan Kumar, Raine Hoover, Daniel A. McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguistics*, 6:391–406.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. Scireviewgen: A large-scale dataset for automatic literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6695–6715. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: evaluating faithfulness and content selection in book-length summarization. *CoRR*, abs/2404.01261.

Daphne Koller and Mehran Sahami. 1997. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997*, pages 170–178. Morgan Kaufmann.

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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, pages 1875–1889. Association for Computational Linguistics.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu-Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.

Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 469–477.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8068–8074. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1980–1989. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Antonio Mastropaolo, Simone Scalabrino, Nathan Cooper, David Nader-Palacio, Denys Poshyvanyk, Rocco Oliveto, and Gabriele Bavota. 2021. Studying the usage of text-to-text transfer transformer to support code-related tasks. In *43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021, Madrid, Spain, 22-30 May 2021*, pages 336–347. IEEE.

Chao Min, Qingyu Chen, Erjia Yan, Yi Bu, and Jianjun Sun. 2021. Citation cascade and the evolution of topic relevance. *J. Assoc. Inf. Sci. Technol.*, 72(1):110–127.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *CoRR*, abs/1902.06673.

Zheng Pang, Renee Raudonis, Bernard R Glick, Tong-Jun Lin, and Zhenyu Cheng. 2019. Antibiotic resistance in pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies. *Biotechnology advances*, 37(1):177–192.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331.

Prasenjeet Roy and Suman Kundu. 2024. Review on query-focused multi-document summarization (QMDS) with comparative analysis. *ACM Comput. Surv.*, 56(1):5:1–5:38.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Matthias Sebastian Rüdiger, David Antons, and Torsten-Oliver Salge. 2021. The explanatory power of citations: a new approach to unpacking impact in science. *Scientometrics*, 126(12):9779–9809.

Furkan Sahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024. Systematic task exploration with llms: A study in citation text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4832–4855. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6252–6278. Association for Computational Linguistics.

Feng Shi and James Evans. 2023. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1):1641.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Grzegorz Siudem, Barbara Żogała-Siudem, Anna Cena, and Marek Gagolewski. 2020. Three dimensions of scientific impact. *Proceedings of the National Academy of Sciences*, 117(25):13896–13900.

Robyn Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP, Collaboratively Constructed Language Resources*, Theory and Applications of Natural Language Processing, pages 161–176. Springer.

Derald Wing Sue. 2001. Multidimensional facets of cultural competence. *The counseling psychologist*, 29(6):790–821.

Stanley Sue. 1998. In search of cultural competence in psychotherapy and counseling. *American psychologist*, 53(4):440.

Zeyu Sun, Qihao Zhu, Yingfei Xiong, Yican Sun, Lili Mou, and Lu Zhang. 2020. Treegen: A tree-based transformer architecture for code generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8984–8991. AAAI Press.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 103–110. ACL.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Suppawong Tuarob, Sung Woo Kang, Poom Wettayakorn, Chanatip Pornprasit, Tanakitti Sachati, Saeed-Ul Hassan, and Peter Haddawy. 2020. Automatic classification of algorithm citation functions in scientific literature. *IEEE Trans. Knowl. Data Eng.*, 32(10):1881–1896.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January, 2015*, volume WS-15-13 of *AAAI Technical Report*. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif M. Mohammad. 2023. We are who we cite: Bridges of influence between natural language processing and other academic fields. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12896–12913. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9440–9450. Association for Computational Linguistics.

Jia-Yan Wu, Alexander Te-Wei Shieh, Shih-Ju Hsu, and Yun-Nung Chen. 2021. Towards generating citation sentences for multiple references with intent control. *CoRR*, abs/2112.01332.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *CoRR*, abs/2405.07437.

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

Weijia Zhang, Jia-Hong Huang, Svitlana Vakulenko, Yumo Xu, Thilina Rajapakse, and Evangelos Kanoulas. 2024b. Beyond relevant documents: A knowledge-intensive approach for query-focused summarization using large language models. In *Pattern Recognition - 27th International Conference, ICPR 2024, Kolkata, India, December 1-5, 2024, Proceedings, Part XIX*, volume 15319 of *Lecture Notes in Computer Science*, pages 89–104. Springer.

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

Xiaodan Zhu, Peter D. Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *J. Assoc. Inf. Sci. Technol.*, 66(2):408–427.

A Existing Work on Scientific Impact Analysis

In Table 6, we compare our work with existing research on scientific impact analysis. More on this in Section 5.

B Prompt for Identifying Impact-revealing Citation Intents

The prompt used is shown in Figure 4.

C Prompt for Generating Impact Summaries and Output Schema

To ensure that the generated output contain all the components required to construct the impact summaries, (namely the impact periods, their dominating intent, details about the period, and citations used as evidence), we make use of OpenAI's structured output feature[11]. Moreover, this facilitates a more systematic automated evaluation for our prompt variants. The prompt design is illustrated in Figure 5, and the corresponding structured output schema is presented in Figure 6.

D Prompts for Evaluating Impact Summaries

Trustworthiness metrics: The prompt we use for our faithfulness evaluation is shown in Figure 7, and for our two-step coverage evaluation in Figure 8. Informativeness metrics: The evaluation steps are shown in Figure 9.

E Details about Zero-shot vs. ICL for Fine-grained Intent Generation

E.1 Training examples

The list of manually annotated examples used in this experiment are in Table 7.

E.2 More on setup and results

We use GPT-4o mini with temperature 0. As for the source of papers and citation context, we use the Semantic Scholar Academic Graph (S2AG)[12].

Figure 10 shows the quantitative results per paper field, and Table 9 shows qualitative examples. We do not observe any large variations indicating that our method is not limited to a specific field. However, the highest recall (a perfect score of 1.0) was achieved in the psychology field. This may suggest that citation contexts in psychology papers tend to express intentions more explicitly, particularly those that reveal impact, e.g. "*Some of those assumptions have been controversial, as researchers disagree about whether the kinds of behaviors measured by particular implicit tests should be considered indicators of attitudes or something else*". This seems to be in line with our observations made during error analysis. To understand where our ICL prompt failed, we plot the confusion matrices in Figure 11 and inspect the 10% impact-revealing citation phrases that were misclassified as "other". We notice that the classifier considers *minor* "resource use" intent as non-impact revealing, e.g., "*We retrieved 400 results for each keyword, resulting in a total of 145,682 stories downloaded in the JavaScript Object Notation (JSON) format [48].*" On the same note, it appears that when "inspiration" is only hinted at in the citation context, the classifier is not able to pick up the signal, e.g., "*The search problem here corresponds closely to a binary dynamic constraint satisfaction problem [Mittal & Falkenhainer , 1990]*". Inter-annotator agreement on the annotated results is in Table 8. They range from fair to substantial. In this evaluation, annotators also indicate their confidence level in their decision on the impact-revealing classification as either "low", "medium", or "high". If there is a disagreement, the annotation with the higher confidence level takes precedence. If the confidence levels are tied, the tie is resolved randomly.

F Effect of Different Number of Shots (K)

We conduct an experiment to test our ICL prompt with different Ks. We manually polished the 200 instances from the previous experiment and randomly split the set into train 40%, dev 30%, test 30%. For the number of shots, we test from K=1 to 80. For every dev and test instance, we run the prompt 5 times and report the average, using GPT4o-mini, shuffling the order of the shots. Results in Figure 12 show that increasing the number of shots has an effect on all metrics, outperforming both the zero-shot, and the always-impact-revealing baselines. Performance improves sharply up to K = 50, after which gains are minimal. At K = 50, metrics plateau with precision at 0.90, recall at 0.94, F1 at 0.92, and accuracy at 0.92.

[11] https://platform.openai.com/docs/guides/structured-outputs

[12] https://www.semanticscholar.org/product/api

Work	Impact as	Facets	Time	Method	Intents	Fields
(Valenzuela et al., 2015)	Influential-citation counts	P.	✗	Supervised Machine Learning	coarse	Computer science
(Zhu et al., 2015)	Influential-citation counts	P.	✗	Supervised Machine Learning	✗	10 fields
(Hutchins et al., 2016)	Relative citation ratio	n/a	✓	Statistical measures	✗	Medicine
(He and Chen, 2018)	Quantified change rate of citation context	n/a	✓	unsupervised temporal embedding analysis	✗	Biomedical
(Jurgens et al., 2018)	Citation counts	n/a	✓	Supervised Machine Learning	coarse	Computer Science
(Bos and Nitza, 2019)	Citation ratio	n/a	✗	Statistical measures	✗	13 fields
(Siudem et al., 2020)	Bibliometric indexes	n/a	✗	Statistical measures	✗	Computer science
(Rüdiger et al., 2021)	Word lists	n/a	✗	Text clustering	✗	Information systems
(Arts et al., 2021)	Concept combinations	P.	✗	Text processing	✗	U.S. patents
(Min et al., 2021)	Citation cascades	n/a	✓	Network analysis	✗	Physics
(Wahle et al., 2023)	Citation Field Diversity Index	n/a	✓	Statistical measures	✗	23 fields, focus on Computer science
(Shi and Evans, 2023)	Concept combinations	P.	✓	Hypergraph model	✗	Medicine, Physics
(Zhang et al., 2024a)	Influential-citation counts	P.	✗	LLM	✗	Computer science
(Gu and Krenn, 2024)	Citation counts	n/a	✓	FNN	✗	Physics
(Arts et al., 2025)	Paper counts	P.	✓	Statistical measures	✗	Nobel Prize-winning papers
This work	Textual Summary	P., C.	✓	LLM	fine-grained	Computer science, Psychology, Medicine

Table 6: Comparison with existing work on scientific impact analysis. Our work is the first to express scientific impact through time-aware textual summaries derived from fine-grained citation context analysis, which covers both praise (confirmation) and critique (correction).

Identifying impact-revealing intents

A **citation context** in a scientific paper refers to the specific part of a paper p' where another paper p is mentioned, including the surrounding sentences or paragraphs that explains how and why p is relevant to p'.
A citation context with an `impact-revealing` intent is a type of citation in scientific writing that highlights the significance or influence of a previously published work, often emphasizing its contribution or importance to the current research or the broader field, e.g., its role in inspiring, motivating, supporting, filling gaps, critically analyzing, or contributing methods, tools, data, extensions, or benchmarks for the current research.
`Other` types of intents include a reference to prior work in a scientific paper that provides background or context without emphasizing the impact, significance, or influence of the cited work. It acknowledges the source in a routine or supporting role rather than showcasing its importance to the research.
Given a citation context, describe, in a few words, the intention behind this citation phrase. Then, decide on the category of this intention. In particular, whether the intention behind this citation phrase is impact-revealing or not (i.e., incidental or that there isn't enough information to realize the real intention behind it). For the intention category, only return one of the following two labels `impact-revealing` or `other`.
Below are examples:
`$examples$`

Figure 4: Prompt for generating and classifying fine-grained intents.

Generating an impact summary about a research paper

The **scientific impact summary of a research paper** describes the impact a given paper had on other papers, including both praise and critique. To understand the impact of a paper, one needs to understand how exactly it has been utilized and discussed by other papers. This is normally referred to as citation intents. One also needs to understand the evolution of the impact and citation intents over time.
Given an input paper's title, its publication year, and its citation context, describe the impact of that paper.
The citation context includes five components: <citation ID, citation title, citation year, citation context, citation intent>.

Given the input paper with id `$paperId$` titled `$title$` published in `$year$`, and the following list of papers citing it:
`$citation_context$`
Generate an impact summary about the input paper.

Figure 5: Prompt for generating an impact summary about a research paper.

Citation context	Intent	Class
In order to reduce the memory requirements, we apply a minimization process [1].	use of minimization methodology	impact-revealing
Motor adaptation is the process of re-shaping acquired motor skills through the reduction of errors (Hardwick and Celnik, 2014; Krakauer, 2009)	defines the term motor adaptation	other
Moreover, none of the above studies explored whether temporally primary PLEs are associated with an increased risk of subsequent insomnia [1,2].. In this study, we explored the changes in prevalence of insomnia and PLEs before and during the pandemic.	identifying and addressing knowledge gap in literature	impact-revealing
Chiu and Nichols (2016) introduced convolutional neural networks for NER	background about NER methods	other
We employ the single-link method to compute the similarity be-tween two clusters, which has been applied widely in prior research (Bagga and Baldwin (1998); Mann and Yarowsky (2003))	use of cluster similarity methods	impact-revealing
Quirk and Poon (2017) and Peng et al. (2017) build two distantly supervised datasets without human annotation, which may make the evaluation less reliable. In this paper, we present DocRED, a large-scale human-annotated document-level RE dataset..	criticizing existing datasets and proposing a better one	impact-revealing
In adults with severe malaria, increased Ang-2 plasma levels were associated with a decrease in NO bioavailability, higher lactate plasma concentrations, and patient mortality [101].	reporting on existing studies about malaria	other
, similarity measures [3]) to select an answer.	not enough information	other
Inspired by the reference [4], we proposed a novel algorithm called Soft-DDQN and applied it to the robot PAP skill learning problem	drawing inspiration from prior work to propose a new algorithm	impact-revealing
For instance, while Yoon et al. (2018) show that expanding the network capacity thorough width is helpful, they have not studied the impact of increasing capacity when the depth increases, nor why increasing the width is helpful. Overall, in this work, we are interested in understanding the impacts of network structure (e.g., width and depth)..	highlighting gaps in existing research on network capacity	impact-revealing

Table 7: Training examples (manually created from real contexts) for fine-grained intent generation and classification.

Task	Cohen's kappa
Impact-revealing classifier	0.68
Fine-grained intent: correctness	0.20
Fine-grained intent: conciseness	0.34

Table 8: Inter-annotator agreement on fine-grained intent generation results (Section 3.1).

```
# Schema for structured output
{"name": "impact_statement",
"schema": {"type": "object",
            "properties": {
            "input_paper_info": {
                "type": "object",
                "items": {
                "input_paper_id": "id of input paper",
                "input_paper_title": "title of input paper",
                "input_paper_year": "year of input paper"}},
            "impact_periods":{
              "type": "array",
              "items": {
                  "impact_period": "start year - end year",
                  "aspect_of_period": "the dominating citation intent(s) of that
                      period",
                  "impact_description": "a paragraph to describe the impact of that
                      period",
                  "evidence": "citing papers from that period to back up the
                      described impact aspect"}}
            }
        }
}
```

Figure 6: Output schema for impact summary generation.

Citation context	**Baseline**	**Ours**
Policy optimization is performed using an implementation of TRPO from rllab [62] with a step size of 0.	The intention behind this citation phrase is to acknowledge the use of a specific tool or method (TRPO from rllab) in the research without emphasizing its impact or significance `other`	reporting method used for policy optimization `impact-revealing`
Alternative methods that instead rely on constraining model parameters, so-called regularization approaches [18,37], have in turn been shown to perform poorly on medical data.	This citation phrase acknowledges prior work related to regularization approaches and highlights that these methods perform poorly on medical data. `other`	highlighting the limitations of regularization approaches on medical data `impact-revealing`
Additionally, suffering can occur due to illness, to the way in which illness is experienced, or, even, in absence of illness (Cassell, 1982; Frank, 2001).	The citation phrase acknowledges prior work to provide context and background for the concept of suffering in relation to illness and its experience. It doesn't emphasize the impact or significance of the cited works. `other`	providing context on the nature of suffering `other`

Table 9: Sample generated intents, baseline= ZS, Ours=ICL (10 examples).

Faithfulness evaluation prompt

Task: Verify the faithfulness of an impact description regarding the paper "{{PAPER_NAME}}". It is faithful if it can be supported by one or more of the paper's citations from the provided list. Each citation is formatted as <title>:citation_text, where the title indicates the title of the citing paper.

Impact description to Verify:
<impact-description>
{{DESCRIPTION}}
</impact-description>

Citation List:
<citations>
{{SOURCES}}
</citations>

Steps to Complete the Task:

1. Understand the impact description and its specified time period.
2. Review each citation in the provided list.
3. Determine if the impact description can be supported by any single citation or a combination of citations.
4. If the impact description is supported, identify the relevant citations that support it.
5. If the impact description cannot be supported or is contradicted by the citations, determine it as unfaithful.

Response Format:

- **Analysis:** <analysis> [Provide your analysis here] </analysis>
- **Answer:** <answer> [yes/no] </answer>
- **Proof:** <proof> [List the exact text of the citations that support the impact description, or "none" if unfaithful] </proof>

Additional Guidelines:

- The answer should be "yes" or "no" only.
- In the proof section, include only the exact text of relevant citations without explanations.
- List all necessary citations if multiple are needed to support the impact description.
- If the impact description is unfaithful, state "none" in the proof section.
- Avoid additional commentary outside the specified format.

Figure 7: Prompt for verifying the time-aware faithfulness of scientific impact summaries.

Prompt for coverage evaluation

Step 1: Given this list of phrases `$listOfPhrases$`, cluster highly similar phrases and give a label (expressive theme) for every cluster.

Step 2: Given this list of themes in the format of a python list: `$listOfThemes$`, determine how many of these themes were implicitly or explicitly mentioned in this summary `$summary$`, and list them.

Figure 8: Prompt for measuring coverage of an impact summary.

```
from deepeval.test_case import LLMTestCase, LLMTestCaseParams
from deepeval.metrics import GEval

insight_metric = GEval(
    name="Insightfulness",
    evaluation_steps=[
        "Determine whether the impact summary describes how the paper has been
            directly used by or influenced by other works.",
        "Assess how well the impact summary articulates the paper's influence with
            informative details.",
        "You should heavily penalize the impact summary for lack of insight."
            ]
)
trend_metric = GEval(
    name="Trend awareness",
    evaluation_steps=[
        "Determine whether the impact summary mentions how the impact of the paper
            has changed over time, ensuring each impact period is clearly identified
             with descriptive titles.",
        "You should heavily penalize the impact summary if the titles of consecutive
             impact periods are not diverse."
            ]
)
specif_metric = GEval(
    name="Specificity",
    evaluation_steps=[
        "Determine whether the impact summary mentions specific techniques,
            frameworks, or studies influenced by the paper, or if it remains broad
            and lacking supporting details.",
        "You should heavily penalize the impact summary if it only restates the
            title and abstract or provides vague, generic statements without
            concrete examples of the influence of the paper."
            ]
)
```

Figure 9: Code snippet for chain of thoughts (CoTs) evaluation steps for 3 metrics of the impact summary evaluation.

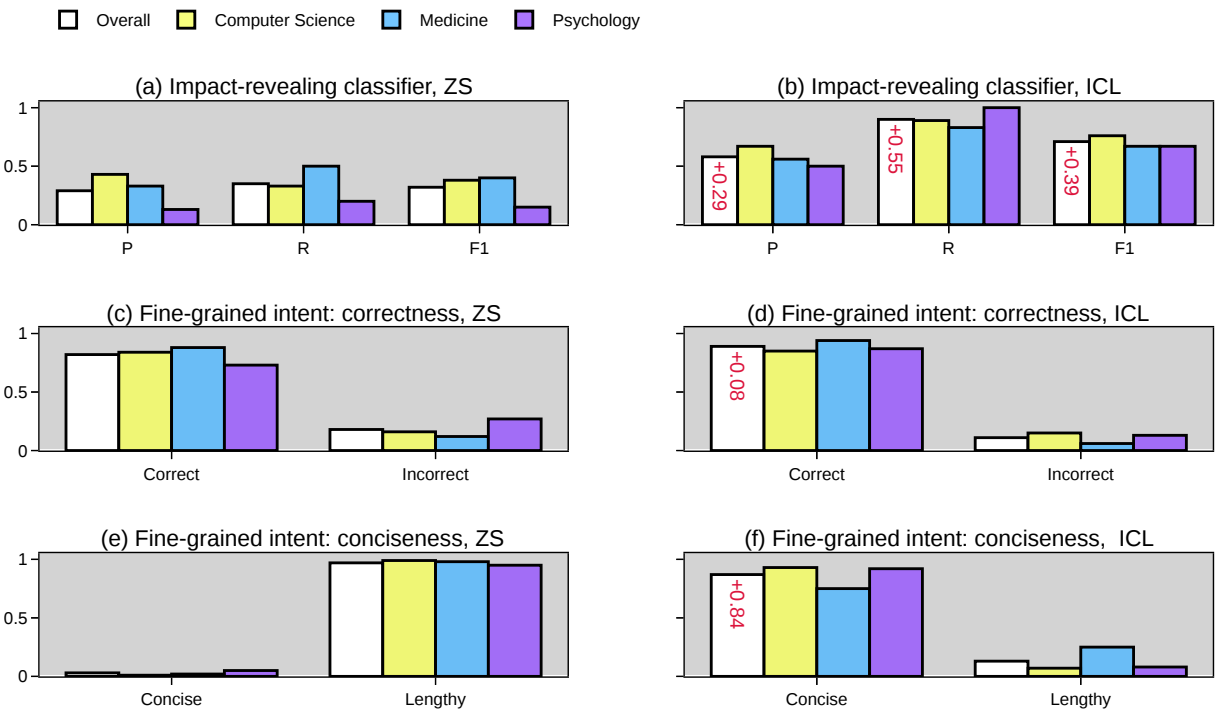

Figure 10: Results for fine-grained intent generations (ZS vs. ICL with 10 examples). Adding examples show significant improvements on metrics like precision, recall, F1.

F.1 A new dataset for identifying impact-revealing citations

Since our task on classifying citation context into "impact-revealing" (for both confirmation/praise and correction/critique) or "other" is new, there is a lack of groundtruth data. To construct such a resource, we build upon on an existing human annotated dataset, namely PST-Bench (Zhang et al., 2024a). An instance in this dataset consists of a pair of research papers annotated for whether one of them influenced (impacted) the other. Influence here is defined only through a positive lens, and is further restricted to the following two intents: "inspiration" and "motivation". To create our dataset, first, we randomly select 1k influential (impact-revealing) papers, as labeled by the PST-Bench dataset annotators, and look up their aggregated citation contexts[13]. Since the meaning of influential, according to PST-Bench, only focuses on two intents "inspiration" and "motivation", both limited to praise, we augment the dataset with other types for both confirmatory and correction citation intents. To do so, we manually craft a handful of textual patterns such as "has led to open questions/challenges/unresolved issues" to capture intents indicating research limitations, then ask GPT4-o to provide a longer list with variations of these patterns. We end up with a total of 33 patterns (examples shown in Figure 13). Following this step, we crawl $\sim$ 200k aggregated citation contexts and search them using the list of impact-revealing phrases, then randomly sample 1k instances from the matching cases. To augment the data with non-impact-revealing ("other") examples, we sample 2k non-influential instances from the PST-Bench dataset. These are citations that are listed under the category "other citations", as in non-influential ones. To ensure the deemed non-influential examples are not in fact impact-revealing (ones which belong to intents other than "motivation" and "inspiration"), we verify that there is no match with all the impact-revealing phrases. Our newly constructed dataset, which we release with this work, consists of 4k citation contexts augmented with either "impact-revealing" or "other" intent class (balanced). After manually inspecting a random sample of 100 instances, we found that 90% were correctly identified as impact-revealing; the remaining 10% included rare cases where phrases like motivated by" appeared to signal impact (e.g., Voters are motivated by partisan social identities... (Greene 2004)") but did not reflect the actual citation intent.

[13]The cited paper might be mentioned several times in the citing paper.

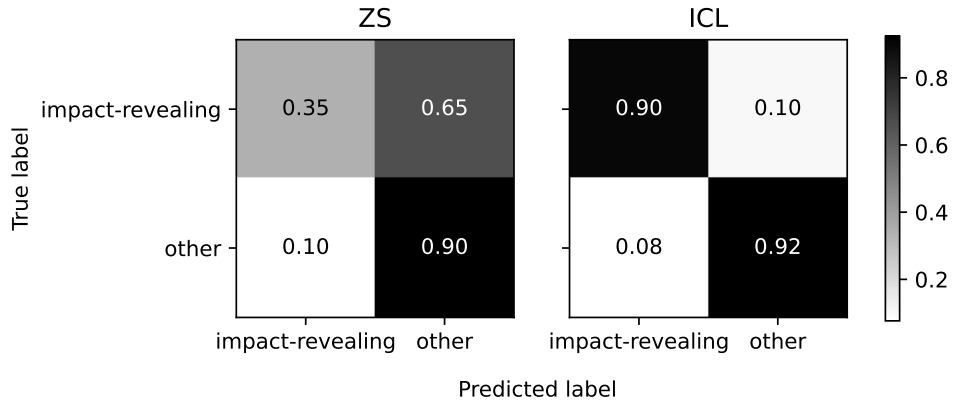

Figure 11: Confusion matrix for the impact-revealing classification task. More insights into misclassified instances in Section E.2

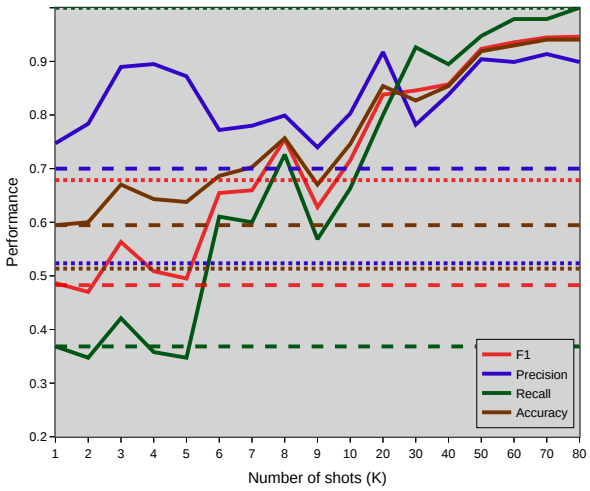

Figure 12: Effect of different number of shots (K) on detecting impact-revealing citations. Zero-shot: horizontal dashed lines; *always impact-revealing*: dotted lines. This shows that adding a reasonable number of examples (around 50) is sufficient to yield a 60% improvement in recall.

F.2 Details about comparisons with intent classifiers

Mappings of the baselines' intents to ours are shown in Table 10. Two of the external methods for intent classification, namely the work of (Cohan et al., 2019) and (Lauscher et al., 2022) allow for multi-label classification. This means: a citation context can receive both "impact-revealing" and "other" labels (e.g., with intents "Result" and "Background"). However, in these cases, a ranking of the predictions can be utilized. For (Lauscher et al., 2022), we use the class with the highest prediction probability. For (Cohan et al., 2019), we count the frequency of every intent class describing the concatenated citation context, and most frequent class represented in the context. We break ties with a random pick.

Sample results are shown in Table 11.

G Details about the Ablation Study

G.1 Examples

Table 17 shows examples of verified (faithful) impact descriptions. Table 18 shows examples for coverage evaluation. Note that we instruct the LLM to return the actual list of topics covered instead of only the number of themes covered, in order to determine its performance compared to a human evaluator (see Table 19). Table 20 shows examples from the citation year compliance evaluation. In Table 21, we show qualitative examples in informativeness.

G.2 Long tail impact coverage

We conduct a focused, quantitative analysis to evaluate the model's ability to capture long-tail yet meaningful aspects of a paper's impact. Specifically, we manually select 10 intent themes associated with 10 different papers that fall outside the top-3 most frequent themes (and are therefore not covered in the Coverage@3 metric) but nonetheless reflect important impact dimensions. We then analyze 50 generated summaries, 10 per variant, and measure the extent to which these summaries capture the selected long-tail themes. We refer to this metric as long tail impact coverage, or LTI for short. As shown in Table 12, LTI ranges from 30% for the "all citation context, no intents" variant to 50% when citation intents are included. These findings suggest that the model is capable of recognizing and reflecting impact-revealing citations

```python
# Python list containing impact-revealing phrases (both praise and critique)
impact_revealing_phrases = [
    r"\b(builds? (upon|on)|extended by|extends the work of)\b",
    r"\b(serve|serves|served) as (a foundation|a basis|the groundwork) for\b",
    r"\b(has|have|had) (paved the way|opened avenues|led to advancements|sparked
        further research)\b",
    r"\b(is|was|has been) (criticized|challenged|questioned) for\b",
    r"\b(suffer(s|ed)? from|is|was|has been) (limited|constrained|hindered|flawed)
        by\b",
    r"\b(has|have|had) (left|created|highlighted|led to) (gaps|challenges|open
        questions|unresolved issues)\b",
   ...
]
```

Figure 13: Samples impact-revealing textual patterns (33 in total). These were used to extend the PST-Bench dataset to include more diverse impact-revealing citations (especially critique/correction).

Intent Classifier	**Classes**
Structural Scaffolds (Cohan et al., 2019)	Background `other` Method `impact-revealing` Result `impact-revealing`
Meaningful Citations (Valenzuela et al., 2015)	non-meaningful `other` meaningful `impact-revealing`
Multi-cite (Lauscher et al., 2022)	Background `other` Motivation `impact-revealing` Future Work `other` Similar/Difference `impact-revealing` Uses `impact-revealing` Extension `impact-revealing`

Table 10: Mapping existing coarse intent classes of existing intent classifiers to our "impact-revealing" or "other" classes.

Structural Scaffolds	Meaningful Citations	Multi-cite	Ours
Citation context: "*Inspired by the theory of hierarchical abstract machines (Parr and Russell 1998), we cast the task of profile reviser as a hierarchical Markov Decision Process (MDP).*"			
Intent predicted by:			
`impact-revealing` (method)	`other` (non-meaningful)	`impact-revealing` (uses)	`impact-revealing` (drawing on theoretical foundations for task formulation)
Citation context: "*..on social network datasets, it is quite intuitive trying to extract information from text data to do ideology-detection, only a few paid attention to links [9, 13]... Even though some realized the importance of links [9, 13], they failed to provide an embedding.*"			
Intent predicted by:			
`other` (background)	`other` (non-meaningful)	`impact-revealing` (motivation)	`impact-revealing` (highlighting the gap in ideology detection methods)
Citation context: "*Recurrent networks are dedicated sequence models that maintain a vector of hidden activations that are propagated through time (Elman, 1990; Werbos, 1990; Graves, 2012).*"			
Intent predicted by:			
`other` (background)	`other` (non-meaningful)	`other` (background)	`other` (providing context on recurrent networks)

Table 11: Intent classes predicted by existing work on intent classification. The intents between parentheses are the actual labels predicted by the method, later mapped to either "impact-revealing" or "other" based on the mappings shown in Table 10. For our method, the output is both the intent and its class.

Prompt variant		LTI Coverage
Citations	Intents	
None	✗	0.1
All	✗	0.3
All	✓	0.5
Impact-rev.	✗	0.4
Impact-rev.	✓	0.4

Table 12: **LTI**: long tail impact coverage

even when it appears less frequently in the input text, highlighting its sensitivity to semantically important but infrequent signals.

G.3 Results per field

Table 22 presents the results for papers of different research fields. We observe that faithfulness and citation year compliance exhibit similar patterns in specific fields as in the overall results, with variants with impact-revealing citations achieving better results. The only exception is the computer science, with a minor difference on year compliance of 1% in favor of variants that receive all citations. We notice that providing citation intents has a negligible effect on faithfulness and decreases the year compliance. This is in line with the overall results. For informativeness, we notice that the same variant wins on all metrics with the exception of insightfulness in the psychology field, where the variant with all citation context and no intents wins by +1%.

G.4 Human-LLM correlation

In Table 19, we report the correlation between LLM and human raters on our evaluation metrics and find that it ranges from moderate to strong, with statistical significance (reported too).

G.5 G-Eval reasoning

We show a few examples of reasons for both high and low scores for each of our metrics in Table 23.

G.6 Visualized impact statements

Examples are shown in Figures 14 (computer science paper), 15 (psychology paper), 16 (psychology paper), and 17 (medicine paper). These impact summaries align closely with our definition. They capture the evolution of citations intents, from initial adoption and theoretical contributions to critiques, methodological refinements, policy influences, and integration into modern applications. For example, the impact summary of "*Hierarchically Classifying Documents Using Very Few Words*" (ICML'97) illustrates how the paper initially shaped the research paradigm in document classification across fields like bio-informatics and legal document analysis. Later, it faced critiques related to scalability and multi-class accuracy, leading to refining the existing method. Another example is the impact summary of "*Sex Bias in Neuroscience and Biomedical Research*" (Neuroscience & Biobehavioral Reviews'11), which highlights the early discussions on gender disparities that this paper sparked. Those discussions later led to policy changes, promoting balanced gender representation in preclinical studies, influencing new research methods across various disciplines.

G.7 Error analysis of our best variant

Although our intent classifier outperforms established baselines (Section 3.2), misclassifications at the intent prediction stage can still compromise the quality of the generated impact summaries. For example, when an incidental citation context is mistakenly classified as impact-revealing, the resulting summary may include statements that do not reflect the true scholarly impact. For instance, in the following part of an impact summary, "impact period: 2003-2010 **Initial Contextual Reference**: During this early period post-publication, the paper was often cited to provide comprehensive **background information** on the prevalence.. and risks associated with Alzheimer's and Parkinson's diseases...", we can see that being mentioned for background information and context is listed under one of the impact periods. This is not in line with our definition of impact summaries, which only restrict impact to direct use of work. We inspect the reason for this mistake and track it back to a misclassification in the previous step, where intents such as "providing context on the prevalence and significance of Parkinson's disease" were classified as impact-revealing.

Conversely, when an impact-revealing citation is misclassified as incidental, the summary is deprived of key evidence of impact. For instance, we observed that a statement in the impact summary describing the use of a graph neural network methodology for drug discovery disappears when the related citation contexts, those referencing this methodological application, are removed from the input.

These cases illustrate the sensitivity of the generation step to errors in intent classification, underscoring the importance of continued research

in this area. We hope our work lays the groundwork for future studies aimed at improving intent classification as a critical component of generating accurate and informative impact summaries.

G.8 Generating impact summaries using other LLMs

We selected GPT-4o as our primary language model for this task due to its state-of-the-art performance in long-context reasoning and summarization. While we recognize the potential biases that may arise from relying solely on a single model family, the use of a consistent LLM and evaluation framework enables us to more clearly isolate the effects of input design. To evaluate the robustness of our method across model families, we also generate impact summaries using Qwen-2.5-72B (Yang et al., 2025) and Gemini-2.5-flash (Comanici et al., 2025), in addition to GPT-4o.

Quantitative results are presented in Table 13. Interestingly, the baseline variant using Qwen achieves the highest insightfulness score. However, this comes at the cost of faithfulness, with the baseline producing 18% fewer faithful claims than our best-performing variant. In contrast, Gemini demonstrates strong performance in citation year compliance, particularly with the impact-revealing with intents variant. Gemini tends to include numerous citations to substantiate its claims during impact periods and accurately aligns each citation with its corresponding time period. This variant also achieves the highest faithfulness (0.96, tied with all citations with intent variant), coverage (0.58), and specificity (0.83).

We present example summaries generated by both Qwen and Gemini in Tables 14 and 15.

H Details on Model Usage

See Table 16.

I More on Practical Applications

I.1 Equal citation count, different citing reasons

We select 10 diverse research topics (e.g., LLMs for code generation, cultural sensitivity in clinical psychology), and analyze 3 papers within each. We ensure that the topic-specific papers have close citation counts. In this use case, we demonstrate how *fine-grained* citation intent analysis provides a deeper understanding of research impact, moving beyond raw citation counts. Results are in Tables 24 and 25. We observe that some papers' main impact is inspiring or motivating new work, some are used for their method or data, others are cited to point out remaining challenges and propose improvements, etc. For example, for papers on open-information extraction with citation counts of $\sim$ 200, (Cui et al., 2018) is frequently cited to motivate new work or highlight existing limitations, while (Gashteovski et al., 2017) and (Han et al., 2019) are cited for their method. In commonsense knowledge mining, only one of the 3 papers is often cited for its dataset (Hwang et al., 2021), and even though the other 2 papers are cited to highlight limitations, the limitations are different, namely challenges with handling constraints and reasoning for one (Davison et al., 2019) , but data noise and limited coverage for the other (Speer and Havasi, 2013). These examples show that even when citation counts are similar, citation intents reveal the specific citing reasons, shaping the perception of a paper's true impact and assisting in the impact summary generation task.

I.2 Author-level impact summaries

The focus of this paper is to generate impact summaries for individual papers. However, we would like to briefly show how these paper-level summaries can be used to create author-level summaries. Given the top-cited[14] papers of a given author, we generate impact summaries using our best variant, and then ask LLM to aggregate these summaries and infer the author's overall impact. The prompt used for generating author-level summaries in Figure 18. We show two samples of such summaries for two senior NLP researchers in Figures 19 and 20. This application case hold great potential for future work, and could be extended to research labs, universities, venues, or countries. Methodologically, it would also be particularly interesting to explore different aggregating strategies to generate these summaries, more specifically whether these kind of impact summaries have better quality by aggregating existing paper-level summaries, or by directly use the citation context about an author's papers regardless of which paper they came from.

[14]Top-10 by citation counts.

Prompt variant		Trustworthiness			Informativeness		
Citations	**Intents**	**Faith.**	**Cov.**	**Cyc.**	**Insi.**	**Trend.**	**Spec.**
Qwen							
None	✗	0.70	0.42	n/a	**0.73**	**0.84**	0.76
All	✗	0.86	0.44	**0.46**	0.67	0.81	**0.78**
All	✓	0.86	0.44	0.44	0.66	0.80	**0.78**
Impact-rev.	✗	0.86	0.44	0.43	0.70	0.81	0.77
Impact-rev.	✓	**0.88**	**0.45**	0.42	0.68	0.81	0.76
Gemini							
None	✗	0.91	0.44	n/a	0.68	**0.84**	0.82
All	✗	0.95	0.52	0.85	0.70	0.83	**0.83**
All	✓	**0.96**	0.57	0.91	**0.72**	**0.84**	0.79
Impact-rev.	✗	0.93	0.54	0.90	0.70	**0.84**	0.80
Impact-rev.	✓	**0.96**	**0.58**	**0.93**	0.68	0.83	**0.83**

Table 13: Ablation results for Qwen and Gemini - Faithfulness: **Faith.**, Coverage: **Cov.**, Citation Year Compliance: **Cyc.**, Specificity: **Spec.**, Insightfulness: **Insi.**, Trend Awareness: **Trend.** These results show that adding impact-revealing contexts and their intents has an improvement on almost all metrics.

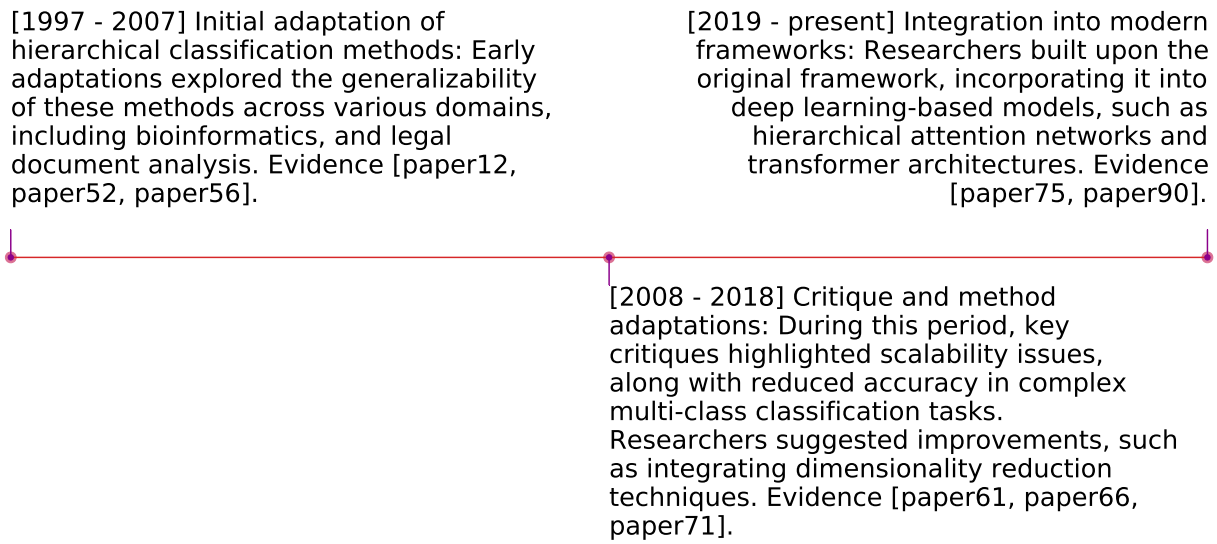

Figure 14: Impact of "Hierarchically Classifying Documents Using Very Few Words", published at ICML'97. (Full example from Figure 1)

Paper: Hierarchically Classifying Documents Using Very Few Words, ICML, 1997.	
Qwen	**Gemini**
Impact period: 1998-2004 *Initial Adoption and Methodological Foundations:* The paper laid the groundwork for hierarchical text classification by introducing a method that effectively classifies documents using very few words.. The initial adoption of this methodology was seen in several studies that built upon or compared their methods against the Koller and Sahami model. For example, the work by TaxE (2004) highlighted the challenge of comparing different hierarchical classification approaches due to the lack of a common dataset, underscoring the need for standardized evaluation metrics. Additionally, early works in 1998-2000 focused on extending and validating the use of hierarchical structures in text classification. **Impact period: 2005-2010** *Extension to Various Domains and Techniques:* During this period, the methodologies proposed in the paper were extended and applied to various domains, including medical vocabularies and web taxonomies. The work by Koller and Sahami was influential in demonstrating the effectiveness of hierarchical classification, which led to the development of new techniques such as the use of word embeddings and unsupervised learning. Research during this period also explored the integration of hierarchical structures with other machine learning algorithms like SVMs and neural networks. For instance, a 2007 study by Chakrabarti et al. drew parallels with Koller and Sahami's approach to attribute selection, while another paper in 2011 critiqued existing feature selection methodologies, suggesting improvements for hierarchical classification. **Impact period: 2011-2020** *Advanced Techniques and Real-World Applications:* In this period, the paper's influence continued to grow as researchers developed more advanced techniques for hierarchical text classification. The work by Koller and Sahami served as a foundational reference for numerous studies that sought to improve the robustness and efficiency of classification models. Specifically, research in 2018 explored the use of word embeddings for hierarchical text classification, building on the initial idea of leveraging hierarchical structures. Additionally, real-world applications of hierarchical classification were demonstrated in areas such as protein function prediction and image classification. The impact of the paper was also seen in the development of testbeds like TaxE, which aimed to standardize the evaluation of hierarchical classifiers.	**Impact period: 1998-2005** *Pioneering Hierarchical Text Classification and Feature Selection:* The paper rapidly established itself as a seminal work in text classification. The paper was widely recognized for introducing and validating the use of class hierarchies to improve classification efficiency and accuracy over traditional flat models (e.g., [15, 16]). Researchers often cited it as the first proper study on hierarchical text categorization (e.g., [12]) and a primary reference for feature selection strategies in high-dimensional text data. However, early discussions also began to identify potential limitations, such as susceptibility to overfitting, the inability to recover from errors made high in the hierarchy, and the constraints of existing datasets for comprehensive evaluation (e.g., [12, 14]). **Impact period: 2006-2015** *Methodological Evolution & Critical Analysis:* This period solidified the paper's status as a foundational reference for hierarchical classification, particularly its top-down approach, often referred to as the pachinko machine model. Researchers continued to build upon and compare against its principles, acknowledging its contributions to computational efficiency through techniques like subtree pruning (e.g., [33, 86]). However, there was a growing critical analysis of its inherent limitations, most notably the blocking issue or error propagation. The paper served as a baseline against which new algorithms were compared (e.g., [28, 92]), and its limitations, such as difficulties scaling to very large web taxonomies, were often highlighted to motivate new research (e.g., [8, 29]). The paper's influence also extended to other domains like protein function prediction and image classification (e.g., [37, 56]). **Impact period: 2016-Present** *Enduring Relevance & Benchmark Status:* The paper maintains its standing as a seminal and foundational benchmark for hierarchical text classification. Its principles are still acknowledged, even as the field evolves with advanced techniques like deep learning (e.g., [38, 41]). Crucially, the paper's identified limitations, such as the propagation of errors in deep hierarchies, are frequently highlighted as motivations for current research to develop more robust and scalable solutions, particularly in areas like hierarchical multi-label classification and for handling large numbers of classes (e.g., [0, 10]).

Table 14: Examples of impact summaries generated by Qwen and Gemini - Part 1.

Paper: Comparability of telephone and face-to-face interviews in assessing axis I and II disorders, The American Journal of Psychiatry, 1997.	
Qwen	**Gemini**
Impact period: 1997-2005 *Validation of Telephone Assessments:* The study provided empirical evidence that telephone interviews could yield results comparable to face-to-face interviews, thereby expanding the methods available for clinical and epidemiological research. This validation was particularly important for large-scale studies and surveys where face-to-face interactions were logistically challenging or cost-prohibitive. The study's findings supported the reliability of tools such as the Personality Disorder Examination (PDE) and the Composite International Diagnostic Interview (CIDI) when administered via telephone, enhancing their utility in diverse settings. **Impact period: 2006-2012** *Equivalence and Reliability in Mental Health Research:* The influence of this paper extended into a broader validation of telephone assessments for various psychiatric disorders. Research during this period consistently cited the study's findings to support the equivalence and reliability of telephone interviews compared to face-to-face methods. This period saw the application of these validated methods in numerous mental health studies, including those on major depressive disorder, anxiety disorders, and substance abuse disorders. The equivalence of telephone assessments was particularly valuable in longitudinal studies and clinical trials, where maintaining consistent and high-quality data collection was essential. Tools such as the SCID (Structured Clinical Interview for DSM-IV) were commonly used in these applications. **Impact period: 2013-2023** *Practical Applications and Cost-Effectiveness:* The practical applications of telephone interviews in mental health research and clinical practice became more prominent. The validated equivalence of telephone assessments led to increased adoption in various settings, including telehealth services, crisis intervention, and routine mental health care. This period also emphasized the cost-effectiveness of telephone assessments, particularly in large-scale surveys and during global events like the COVID-19 pandemic. The practicality and efficiency of telephone interviews were highlighted, making them a preferred method for data collection, especially when face-to-face interactions were not feasible.	**Impact period: 1997-2007** *Establishing Comparability and Validity of Telephone Interviews:* The paper provided crucial evidence demonstrating that telephone interviews yield comparable results to traditional face-to-face assessments for various disorders, including anxiety and major depressive disorders. Researchers frequently cited the paper to support the validity of using telephone methods for large-scale data collection in clinical and epidemiological studies. The paper also began to inform discussions on the practical advantages of telephone interviews, such as cost-efficiency, broader geographical coverage, and potential for higher response rates, while also highlighting the importance of interviewer-subject rapport and identifying early nuances like potential underreporting of sensitive information (e.g., substance abuse) and the comparability for specific diagnostic tools like the CIDI or SCID. **Impact period: 2008-2017** *Broadened Acceptance and Practical Utility:* Over this period, the paper continued to be cited as authoritative evidence for the validity and reliability of telephone interviews, extending their application across a wider range of mental health conditions, including PTSD and eating disorders. Its influence shifted from merely establishing comparability to emphasizing the practical benefits of telephone methods, such as reduced participant burden, improved subject retention in longitudinal studies, and enhanced logistical efficiency for large-scale research. While broadly affirming the equivalence of telephone and face-to-face methods, citations during this time also reflected ongoing discussions about specific methodological considerations, such as the potential for differences in reporting stigmatizing behaviors or the role of nonverbal cues. **Impact period: 2018-2024** *Enduring Relevance and Standard Practice:* In recent years, the paper's impact has solidified its position as a foundational reference for the routine use of telephone interviews in clinical assessment. It is cited as evidence for the established reliability and validity of telephone-based diagnostic methods, supporting their application as a standard practice for diagnosing various mental health conditions and even for monitoring adverse events. The citations highlight the practicality and effectiveness of telephone interviews, especially in contexts requiring reduced burden or remote data collection, such as during the COVID-19 pandemic.

Table 15: Examples of impact summaries generated by Qwen and Gemini - Part 2.

Task/Component	Model Name	Details
intent classification and generation	gpt-4o-mini	temp=0, version=2024-07-18, Input: 128,000, Output: 16,384
impact summary generation	gpt-4o	temp=0, version=2024-11-20, Input: 128,000, Output: 16,384
impact summary evaluation	gpt-4o	temp=0, version=2024-11-20, Input: 128,000, Output: 16,384
textual pattern generation (for groundtruth data extension)	gpt-4o	temp=0, version=2024-11-20, Input: 128,000, Output: 16,384
practical applications (citation context analysis)	gpt-4o-mini	temp=0, version=2024-07-18, Input: 128,000, Output: 16,384
practical applications (author-level summary generation)	gpt-4o	temp=0, version=2024-11-20, Input: 128,000, Output: 16,384

Table 16: Details about the LLMs used in our experiments.

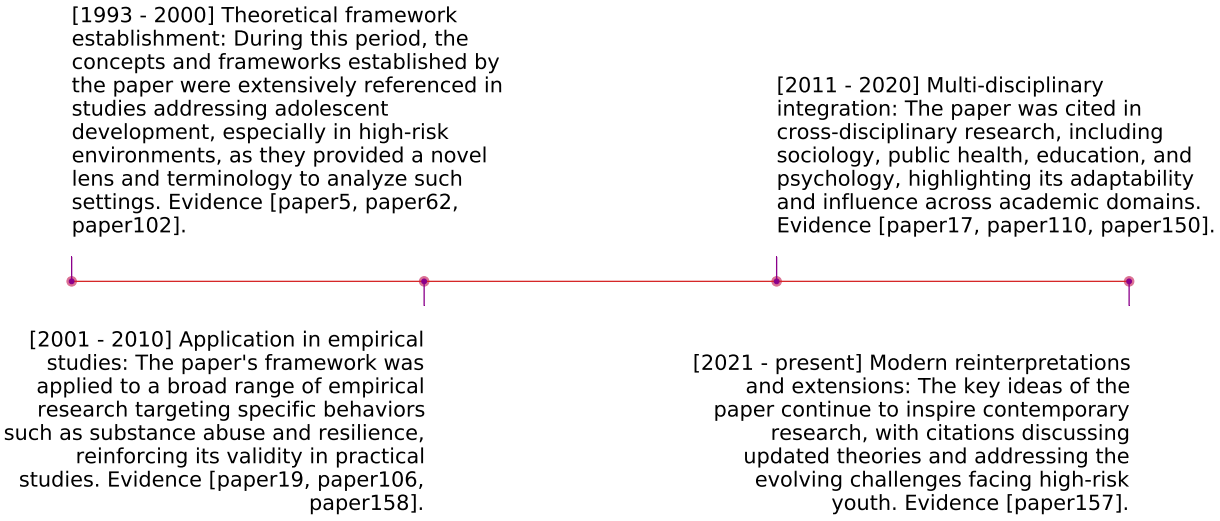

Figure 15: Impact of "Successful Adolescent Development among Youth in High-Risk Settings", published in American Psychologist'93.

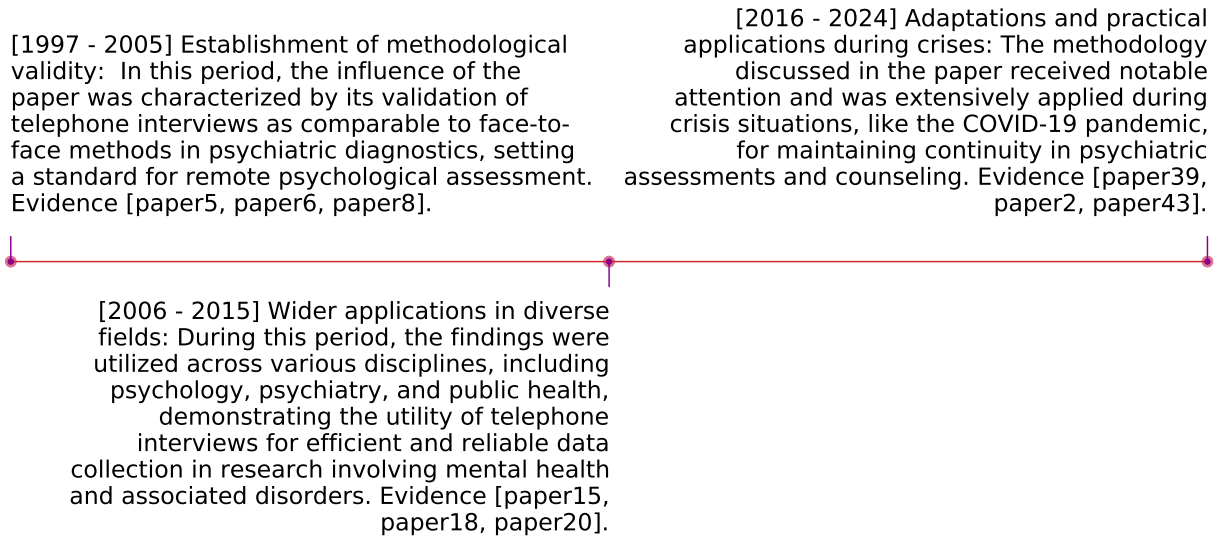

Figure 16: Impact of "Comparability of telephone and face-to-face interviews in assessing axis I and II disorders", published in The American Journal of Psychiatry'97.

Impact summary	Analysis and model's verdict	Proof (partial citations list for brevity)
Paper: *The importance of race and ethnic background in biomedical research and clinical practice* **[2011-2020] Expansion and debate** The discussion around the provided insights flourished, with broader interdisciplinary citations related to ethics, sociological interpretations, and medical trial designs highlighting its significance.	The impact description discusses the flourishing discussion around insights related to race and ethnic background in biomedical research, emphasizing interdisciplinary citations that highlight its significance in ethics, sociological interpretations, and medical trial designs. Several citations in the provided list address the importance of race and ethnicity in biomedical research and clinical practice, indicating that these discussions are indeed significant and supported by various studies. **Verdict**: *Verifiable*	**<Associations Between Vitamin D Receptor Polymorphisms and Susceptibility to Ulcerative Colitis and Crohn's Disease: A Meta-analysis>**: "The importance of ethnic background has been raised in bio-medical research and clinical practice.(35) Genetic variation that predisposes to IBD appears to vary between different ethnic groups." **<Conceptual approaches to the study of health disparities.>**: "Some researchers have argued that this correlation between genetically identified classifications and self reported race justifies the continued use of self-reported race and ethnicity as a proxy for genetic differences in epidemiologic studies (11, 71)."
Paper: *Myocardial infarction redefined–a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction.* **[2006 - 2015] Integration into clinical practice guidelines** During this period, the criteria discussed in the input paper were incorporated into widely used clinical guidelines, affirming its essential role in healthcare practice and education.	The impact description states that the criteria discussed in the paper were incorporated into widely used clinical guidelines, affirming its essential role in healthcare practice and education. The citations provided indicate that the Joint European Society of Cardiology/American College of Cardiology criteria for diagnosing myocardial infarction have been referenced and utilized in various studies and guidelines, confirming their integration into clinical practice. This supports the claim that the criteria have been widely adopted in healthcare settings. **Verdict**: *Verifiable*	**<Trace element status in Saudi patients with established atherosclerosis.>**: "A diagnosis of a myocardial infarction was made in accordance with Joint European Society of Cardiology/American College of Cardiology Committee criteria [17]." **<Improved treatment and prognosis after acute myocardial infarction in Estonia: cross-sectional study from a high risk country>**: "The criteria applied for AMI diagnosis on 2001 and 2007 study populations were based on the consensus document published by the European Society of Cardiology in 2000 [12]."

Table 17: Examples of verified (faithful) impact summaries impact descriptions.

Impact summary	Themes to be covered
During this early period following the publication, the method proposed by the paper, such as the Spy technique, received attention by its developers who utilized it in the domain of semi-supervised learning. Researchers particularly explored its applications in areas like text document classification and highlighting limitations in existing learning paradigms. The study's methodology spurred advancements in classifications approaches dealing with positive and unlabeled data. This phase not only saw adaptation of the methodology into various applied fields like web document classification but also critical analysis and theoretical exploration particularly in precision improvement. Researchers evolved the PU learning methods by optimizing the constraints and extending it into more extensive datasets and alternative models leading towards scaling and efficiency enhancements in text and online information domains. In this recent phase, the Spy technique from the discussed paper continues to be influential, inspiring modern machine learning techniques like adaptive PU-learning and text mining approaches. The bridging of neural methods with earlier methodologies showcases the adaptability and robustness over time.	Performance Comparison Novel Algorithm Development Limitations Acknowledgment Method Adaptation and Improvement Applications of PU Learning Classification Techniques Data Handling and Sampling
During these years, the paper served as a cornerstone for advancing the implementation of individualized approaches in clinical decision-making. Several studies cited the paper while exploring personalized medicine's integration. Focus on translational research and advances. This period saw the application of concepts introduced in the paper in fine-tuning medical treatments and drug therapies for individual patients. The references highlight adaptations in various domains from rheumatology to oncology. The paper's concepts shaped the research direction towards the future application of precision medicine utilizing technologies like AI and computational advances.	Advancements in Precision Medicine Role of Genetic Variants and Pharmacogenomics Technological Impact on Personalized Medicine Challenges and Gaps in Personalized Medicine Patient Variability and Individualized Treatment Applications of Personalized Medicine in Specific Fields Ethical and Regulatory Considerations
The input paper, published in 2020, provided a significant synthesis of knowledge about the public health importance of hookworm infections.	Effectiveness of Treatment for Anemia Public Health Significance of Hookworm Disease Limitations of Targeting School-Age Children Need for Studies on Military Infection Burdens ..140 more..

Table 18: Examples of automated coverage results (themes actually covered).

	Faithfulness	Coverage	Insightfulness	Trend Awareness	Specificity
Spearman	0.668*	0.481**	0.489*	0.535*	0.683*
Kendall-Tau	0.668*	0.443**	0.482*	0.533*	0.676*

Table 19: Human and LLM correlation and agreement on evaluating results (Section 3.3). p-value: *$\leq$0.001, **$\leq$0.05.

Impact summary	Complying citations	Noncomplying citations
[2021 - 2022] supporting economic and societal discussions of cardiovascular diseases Studies during this period heavily relied on the paper for understanding the economic burden and healthcare strategies towards cardiovascular diseases, often referencing the comprehensive statistics presented.	**Title**: "Extrinsically Conductive Nanomaterials for Cardiac Tissue Engineering Applications", **Year**: **2021**, **Intent**: "highlighting the significant economic impact of cardiovascular diseases"	**Title**: "Optimising the treatment of chronic ischemic heart disease by training general practitioners to deliver very brief advice on physical activity (OptiCor): protocol of the systematic development and evaluation of a complex intervention", **Year**: **2024**, **Intent**: "highlighting the significant economic impact of cardiovascular events on the German healthcare system"
[2021 - 2023] guiding therapeutic practices and further clinical recommendation development Multiple citations within this period reflect that the publication served as a cornerstone source for clinical trials and therapeutic evaluation studies aiming to manage PAD effectively.	**Title**: "Clinical considerations after endovascular therapy of peripheral artery disease", **Year**: **2021**, **Intent**": "highlighting the significance of managing PAD as a CAD equivalent"	**Title**: "Ankle Brachial Index: An Easy and First-Choice Screening Marker of Peripheral Artery Disease and Physical Function", **Year**: **2024**, **Intent**: "highlighting the significance of ABI in identifying PAD and associated risks"
[2004 - 2010] Initial adoption of stochastic gradient descent methods After its publication in 2004, the paper introduced stochastic gradient descent (SGD) as an efficient method for solving large-scale linear prediction problems, enabling significant computational advantages in machine learning applications.	**Title**: "Fast Stochastic Frank-Wolfe Algorithms for Nonlinear SVMs", **Year**: **2010**, **Intent**: "reporting efficiency of stochastic programming techniques"	**Title**: "A Robbins-Monro Sequence That Can Exploit Prior Information For Faster Convergence", **Year**: **2024**, **Intent**: "highlighting the widespread application and significance of a widely used algorithm"

Table 20: Examples of evidence citations from compliant and noncompliant years. Note the similarity between the summary impact description and the intent of noncompliant citations (highlighted), which might be the source of the error.

Paper	Baseline	Ours
Liu, Bing, et al. Partially Supervised Classification of Text Documents. 2002.	..**[2005-2012]: Development and application** *The approaches defined were adapted and extended in a number of practical applications to fine-tune frameworks for domain-specific text classification tasks ..*	..**[2008-2012]: Applications & theoretical exploration** *The methods were applied to problems beyond text classification, such as bioinformatics, personalized recommendations, and data mining. This phase not only saw adaptation of the methodology but also critical analysis and theoretical exploration particularly in precision improvement ..*
Martens, James. Deep learning via Hessian-free optimization. 2010.	..**[2015-2020]: Adoption in new areas** *The techniques described in the paper facilitated advancements in optimizing models in diverse fields like natural language processing and reinforcement learning ..*	..**[2016-2020]: Integrating into broader optimization frameworks and comparisons** *As the field evolved, the approaches from this work became integrated into broader optimization discussions and comparison studies. Researchers began leveraging data-driven adaptations and contextualizing the paper's contributions relative to emerging advancements like Adam and other methods, enriching overall optimization frameworks ..*
Lee, Honglak, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. 2009.	..**[2013-2017]: Facilitating Advances in Image Recognition** *Techniques and methodologies from the paper influenced deep learning architectures, becoming foundational in computer vision applications ..*	..**[2012-2017]: Advancements in convolutional neural network methodologies** *This period marks significant enhancements in deep convolutional techniques, where the paper provided foundational principles that inspired advancements in methodologies across image processing, feature extraction, and specific application areas such as bioinformatics and health diagnostics ..*
Myin-Germeys, Inez, et al. Experience sampling research in psychopathology: opening the black box of daily life. 2009.	..**[2015-2023]: Application in diverse psychopathological studies** *The methodologies introduced by the paper were applied extensively in studies of mood disorders and other psychopathological conditions to dissect complex behavioral and cognitive interactions ..*	..**[2015-2019]: Expanding applications** *The method presented catalyzed its application across diverse practices such as genetic interaction studies, psychological intervention assessments, and computational behavioral analysis. Researchers recognized its role in enhancing ecological validity and precision of dynamic mental state evaluations* **[2019-Present]: Technological Integration** *This period saw a surge in mobile and wearable technologies for real-time data collection, integrating ESM with machine learning and mobile health platforms to advance personalized medicine and therapeutic interventions. ..*

Table 21: Examples of impact summaries from the informativeness evaluation, baseline = no-knowledge variant, our method = impact-revealing variants.

Prompt variant		Trustworthiness			Informativeness		
Citations	**Intents**	**Faith.**	**Cov.**	**Cyc.**	**Insi.**	**Trend.**	**Spec.**
Medicine							
None	✗	0.70	0.23	n/a	0.75	0.93	0.82
All	✗	0.80	0.30	0.52	0.82	0.95	0.87
All	✓	0.82	0.30	0.47	0.83	0.94	0.88
Impact-rev.	✗	**0.85**	0.31	**0.57**	0.82	0.95	0.86
Impact-rev.	✓	**0.85**	**0.32**	0.56	**0.87**	**0.96**	**0.89**
Psychology							
None	✗	0.74	0.23	n/a	0.75	0.95	0.80
All	✗	0.86	0.32	0.54	**0.84**	**0.98**	**0.89**
All	✓	0.86	0.32	0.52	0.79	0.94	0.86
Impact-rev.	✗	**0.90**	**0.34**	**0.63**	0.82	0.97	**0.89**
Impact-rev.	✓	0.89	0.32	0.59	0.83	**0.98**	**0.89**
Computer science							
None	✗	0.87	0.30	n/a	0.60	0.94	0.63
All	✗	0.83	0.35	**0.59**	0.75	0.96	0.83
All	✓	0.84	0.36	0.46	0.77	0.98	0.82
Impact-rev.	✗	0.85	0.35	0.58	0.74	0.95	0.86
Impact-rev.	✓	**0.89**	**0.37**	0.52	**0.79**	**0.99**	**0.86**

Table 22: Ablation results per field - Faithfulness: **Faith.**, Coverage: **Cov.**, Citation Year Compliance: **Cyc.**, Specificity: **Spec.**, Insightfulness: **Insi.**, Trend Awareness: **Trend.**

Criteria	Score	Reason
Insightfulness	0.9	The output provides a clear description of the paper's influence, detailing its foundational, methodological, and expanded relevance. It highlights specific uses and integration into modern frameworks. Slightly more detailed examples could enhance insight further.
Insightfulness	0.4	While the output describes stages of adoption and application, it lacks detailed insight into specific ways the paper directly influenced or was used by other works.
Trend Awareness	1.0	The output clearly identifies how the paper's impact has changed over time with descriptive period titles. These titles are diverse and informative, fulfilling the evaluation criteria.
Trend Awareness	0.4	The output mentions distinct time periods with descriptions of the paper's impact, but the titles for the periods are not diverse, as both cover similar themes without clear differentiation.
Specificity	0.9	The output provides detailed periods showcasing the paper's influence on foundational discussions, specific applications, and technological progress, meeting most evaluation criteria. A minor deduction is due to the lack of explicit mention of individual techniques influenced by the paper.
Specificity	0.3	The output generically states the paper's influence and provides broad areas of impact, but it lacks concrete details, specific techniques, studies, or frameworks influenced by the paper.

Table 23: Examples of scores and reasons from the LLM evaluator.

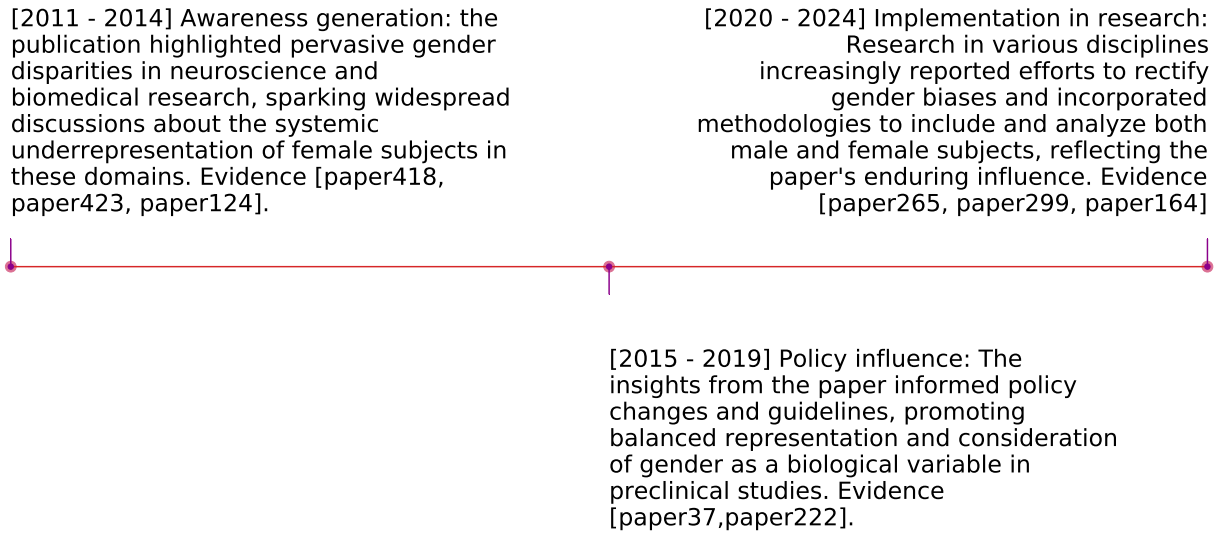

Figure 17: Impact of "Sex Bias in Neuroscience and Biomedical Research", published in Neuroscience & Biobehavioral Reviews'11.

Generating a scientific impact summary about a researcher

The scientific impact of a researcher can be inferred from the impact of their publications and how they have been used by others. Given a list of impact summaries about papers of a certain researcher, summarize their overall impact, and its evolution over time.

This is the list of semi-structured paper summaries about the researcher:
`$paper_impact_summaries$`
Generate an impact summary that describes the overall impact of their papers. Focus more on how their papers have been used than what the content of their papers was.

Figure 18: Prompt for generating author-level scientific impact summaries.

Topic: Open information extraction; citation count: ∼200		
Paper: (Cui et al., 2018) **Top citation intents:** motivating new work in neural OIE, acknowledging limitations in generating facts and schema strength	(Gashteovski et al., 2017) comparing proposed methods with sota approach, method use for OIE	(Han et al., 2019) method use and extension for relation extraction, highlighting performance issues in capturing various linguistic cues
Topic: Commonsense knowledge mining; citation count: ∼350		
(Davison et al., 2019) acknowledging limitations in handling constraints and reasoning, promoting the perspective of LLMs as knowledge base	(Speer and Havasi, 2013) highlighting limitations of data noise and limited relation coverage, emphasizing importance of commonsense knowledge graphs in AI applications	(Hwang et al., 2021) reporting performance comparison and improvements, use as data source for commonsense reasoning
Topic: Retrieval augmented generation; citation count: ∼[400-500]		
(Asai et al., 2024b) motivating the effectiveness of RAG in addressing hallucinations in LLMs, reporting variability in LLMs' performance with RAG methods	(Ram et al., 2023) motivating the need for well-chosen context in RAG, reporting limitations of retrievers in RAG	(Mallen et al., 2023) motivating research on LLMs for question answering, motivating the goal of improving LLMs' factuality
Topic: LLM as a judge; citation count: ∼[370-470]		
(Fu et al., 2024) motivating the growing interest in using LLMs for automatic evaluations, used on evaluating conversational agents and summary faithfulness	(Chan et al., 2024) highlighting the growing trend of using LLMs for evaluation, use of agent-based methods in evaluation processes	(Wang et al., 2024) motivating the need for further investigation into biases in Judge LLMs, highlighting concerns regarding fairness and bias in using LLMs for evaluation
Topic: Medical domain LLMs; citation count: ∼[1900-2100]		
(Gu et al., 2022) use of domain-specific pretrained models, comparing performance of various transformer models in biomedical tasks	(Singhal et al., 2022) highlighting a gap in existing research on interactive medical services, performance comparisons of LLMs in medical knowledge tasks	(Thirunavukarasu et al., 2023) highlighting privacy concerns in data collection, performance comparison between LLMs and fine-tuned small models
Topic: Cultural sensitivity in clinical psychology; citation count: ∼[1300-1500]		
(Bernal et al., 2009) highlighting the ongoing debate regarding cultural competency in treatment approaches, motivating the need for comparative analysis of culturally adapted evidence-based practices	(Sue, 2001) highlighting limitations in existing multicultural discourse research, motivating the application of cultural competency models in transgender care	(Sue, 1998) motivating work on cultural sensitivity in therapy, highlighting the gap in research on treatment efficacy for diverse ethnic populations

Table 24: Similar citation counts, different citation intents. Part I.

Topic: LMs for code generation; citation count: ∼[200-300]		
Paper: (Sun et al., 2020) **Top citation intents:** drawing inspiration from prior work, comparing various state-of-the-art methods for code generation	(Le et al., 2022) suggesting potential integration of various search strategies for code generation, motivating the trend of using deep learning for code generation	(Mastropaolo et al., 2021) highlighting limitations in existing methods, building on previous ideas for code summarization
Topic: Automated fact checking; citation count: ∼700		
(Ma et al., 2018) motivating the growing interest in ML and NLP for rumor and fake news detection, comparing the proposed model with existing methods for effectiveness verification	(Monti et al., 2019) reporting shortcomings in fake news detection methods, comparing performance improvements of different methods	(Bian et al., 2020) highlighting the limitations of existing datasets for the task, drawing inspiration from existing strategies for early detection and localization
Topic: Multi-modal deep learning; citation count: ∼2000		
(Kim et al., 2021) reporting performance comparisons among multimodal methods, highlighting advancements in multimodal pre-training and proposing new tasks	(Liu and Tuzel, 2016) reporting model performance compared to state-of-the-art, highlighting the success of GANs in image generation	(Li et al., 2021) highlighting the limitations of the current method in comparison to other visual LMs, drawing inspiration from existing multimodal models to propose a new method
Topic: Antibiotic resistance mechanisms; citation count: ∼[2300-3000]		
(Aslam et al., 2018) motivating the urgent need for new antibiotics due to rising resistance, highlighting the potential therapeutic effects of the method	(Pang et al., 2019) highlighting the challenges and variations in antibiotic effectiveness against biofilms, motivating the urgent need for novel treatment strategies	(Wang et al., 2024) motivating the need for ongoing antibiotic development and resistance study, highlighting the role of human behavior in antibiotic resistance

Table 25: Similar citation counts, different citation intents. Part II.

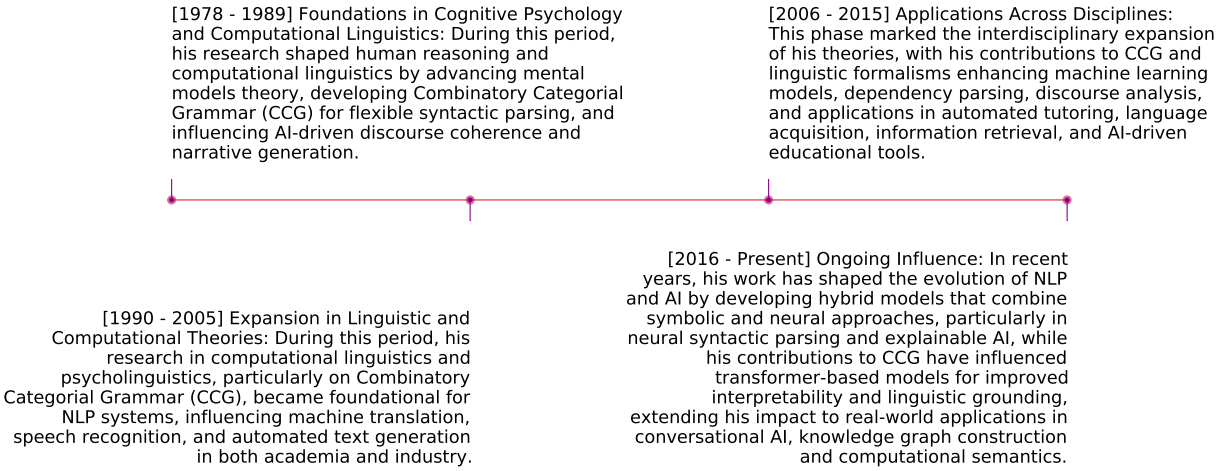

Figure 19: The impact summary of the work of a computer science researcher (focus on NLP, cognitive science), H-index=69.

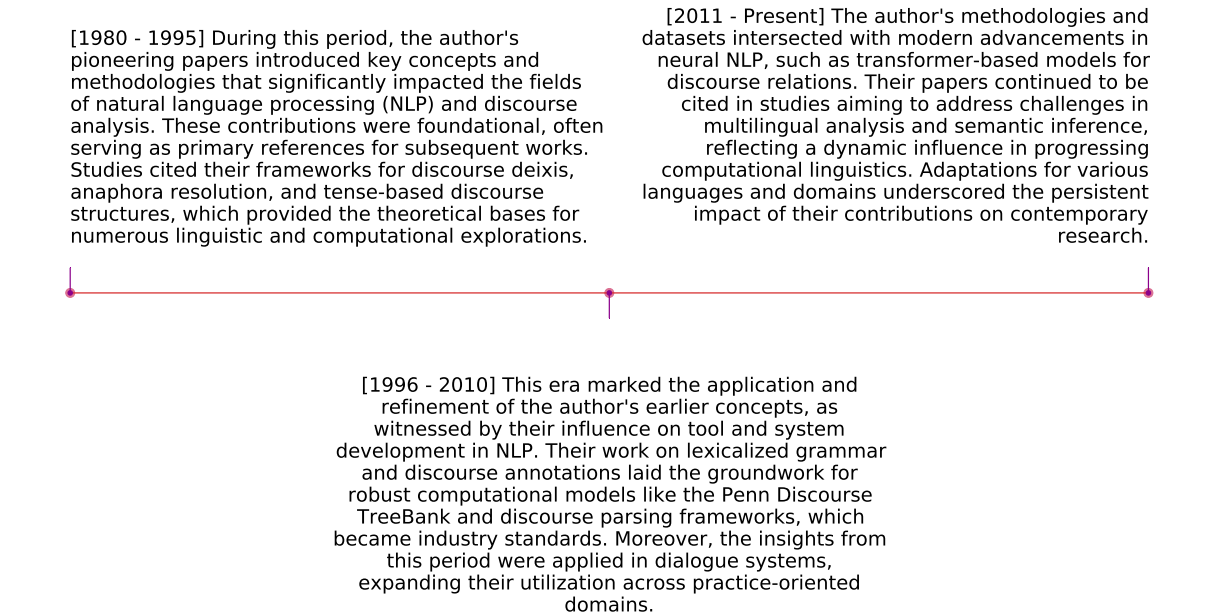

Figure 20: The impact summary of the work of a computer science researcher (focus on NLP, computational linguistics), H-index=61.