Echoes of Citations: Automated Extraction of Claims from Full Scientific Papers

Neşet Özkan Tan [1], Niket Tandon[4], Oyvind Tafjord[3], Michael Witbrock [1], Peter Clark [2], Mark Gahegan [1]

[1]The University Of Auckland
[2]Allen Institute for Artificial Intelligence
[3]Google DeepMind
[4]Microsoft Research
neset.tan@auckland.ac.nz

Abstract

Automated extraction of core scientific claims, concise statements of a paper's primary contributions, is critical for navigating the growing scientific literature. We present a scalable framework that leverages citances, sentences from other papers citing the target work, as natural supervision, removing the need for costly manual labelling. Our method filters citances with a claim-focused rubric and aligns them with candidate claims to train two pipelines: an unsupervised extractor and a weakly supervised model. Experiments show our approach outperforms existing baselines, achieving up to 18% higher precision and 22% greater coverage. We further analyse claim distributions across paper sections and introduce a taxonomy of claim types, providing new insights into the rhetorical structure of scientific discourse.

Introduction

The exponential growth of scientific literature makes it increasingly difficult for researchers to stay informed and synthesize emerging knowledge (Knoth et al. 2023). At the heart of this challenge lies the need to identify a paper's core claims, concise statements that capture its main contributions. Automating this process can help researchers keep pace with new findings, assist reviewers in detecting supporting or conflicting evidence, and enable the construction of knowledge graphs that reveal relationships across studies. A scalable and accurate claim extraction system would thus accelerate knowledge discovery and foster interdisciplinary collaboration.

Existing approaches to claim extraction remain limited in scope or rely heavily on manual supervision. Most focus on structured abstracts (Wadden et al. 2020; Wei 2023; Tan et al. 2024), which are common in biomedical research but rare in disciplines such as computer science and mathematics. Restricting extraction to abstracts overlooks many key claims distributed throughout the full text. Moreover, supervised methods (Wadden et al. 2022) depend on manually annotated data, costly to produce and difficult to scale across domains. These limitations underscore the need for approaches that can extract claims from the full text with minimal human supervision.

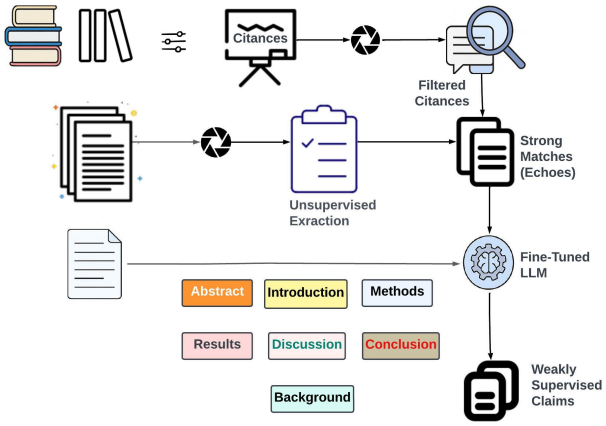

Figure 1: We introduce unsupervised and weakly supervised pipelines for extracting core claims from the full text of scientific papers. Our method leverages claim-focused citances as proxy supervision signals, fine-tuning a weakly supervised model on strong matches between unsupervised claims and citances. The model generalizes to any scientific text, including uncited or newly published papers, by weighting both structural sections and claim categories.

In this work, we leverage *claim-focused citances*, citations that explicitly highlight a paper's contributions, as weak supervision signals. By aligning these citances with candidate claims extracted from full-text research papers, we develop a scalable framework for identifying core scientific claims without manual labels. We introduce two complementary pipelines: an unsupervised method requiring no supervision, and a weakly supervised model trained on citance-aligned data. Together, these approaches balance precision and coverage, combining the comprehensiveness of full-text analysis with the precision of community-driven signals.

Scope and novelty: Our approach builds on prior work in citation-context and mention extraction, but extends it in three key ways: (i) rubric-based filtering of claim-focused citances, (ii) alignment of claims and citances using LLM-calibrated degree-of-match scoring, and (iii) a scalable train-

ing and evaluation protocol requiring no manual annotation. Rather than aiming to capture all possible contributions, we target a practical operating point that maximizes accuracy and generalizability, while acknowledging and mitigating the evaluative biases inherent in citance-anchored signals.

Contributions: Our main contributions are as follows:

- **Full-text, low-supervision claim extraction.** We integrate an unsupervised extractor with a weakly supervised model trained on rubric-filtered, citance-aligned data, achieving improved precision and coverage over abstract-only and sentence-level baselines without manual annotation.
- **Citance-anchored evaluation with human calibration.** We use claim-focused citances as scalable proxies for evaluation and calibrate degree-of-match thresholds with human judgments, clarifying precision–coverage trade-offs.
- **Dataset and reproducible protocol.** We release a dataset of 1,200+ computer science papers with claim-focused citances and a fully replicable pipeline for filtering, alignment, and evaluation.
- **Analyses and insights.** We examine where and what types of claims are extracted across sections and themes, identifying common error modes (e.g., novel terminology, numeric specificity, implicit background) and proposing concrete remedies.

Framework

Dataset

Using the Semantic Scholar API, we assessed the full-text availability of computer science papers, resulting in a dataset of 1,224 papers published between 2020 and 2023, each cited at least 20 times. This citation threshold was chosen to ensure access to rich citance signals for training and evaluation while tolerating some degree of noise. The papers span both Computer Science and Interdisciplinary Studies: 602 are exclusively categorized under *Computer Science*, while the remaining papers combine Computer Science with 19 other fields, including Law, Medicine, Engineering, Mathematics, and Linguistics.

We randomly split the dataset into 734 training samples and 490 testing samples, keeping the test set sufficiently large to enable robust evaluation and minimize distribution shift biases. Although the training set focuses on well-cited papers (20 or more citations) to leverage abundant citance supervision, our framework does not rely on citation density. Both the unsupervised and weakly supervised pipelines are designed to generalise beyond citation signals and remain effective for newly published or uncited papers by relying on full-text structure and content rather than citation counts.

Filtering Citances as a Proxy for Gold Claims

In our dataset, we observed an average of 29.46 citances per paper. However, many of these citances are vague, underspecified, or offer little insight into the cited work. Examples of such low-information citances are provided in Table 1.

Examples of Low-Information Citances
"A prior work did not cover the significant aspects found in the current paper."
(Lacks details on which aspects were not covered.)
"A prior work is mentioned in Table 5 as previous methods."
(Refers to a table without contextual information.)
"A prior work that can be used with the approach mentioned above."
(Vague reference to an approach without elaboration.)
"A prior work that reported results are included in the comparison."
(Does not specify the results or their significance.)

Table 1: Examples of vague or low-information citances that offer limited contribution to claim identification.

To reduce noise and isolate key contributions, we applied a rubric adapted from (Wei 2023) to filter citances. This rubric targets citances that explicitly contain:

- statements that declare something is better;
- statements that propose something new;
- statements that describe a new finding or a new cause–effect relationship.

By applying this rubric, we filtered out 'noisy' citances and focused on those that clearly reflect the primary contributions and findings of the cited works. These criteria also align with our operational definition of a scientific claim, as outlined in (Wei 2023).

Claim Extraction

To extract core claims from scientific texts, we designed two complementary pipelines (see Figure 1): (1) an **unsupervised approach** that relies solely on the paper's full-text structure and content without requiring manual labels, and (2) a **weakly supervised approach** that incorporates high-quality signals derived from claim-focused citances.

These pipelines jointly address the dual challenge of improving both coverage and precision while minimizing the dependence on costly human annotations. Importantly, they are *model-agnostic* and can be integrated with any large language model (LLM).

We conducted experiments with multiple state-of-the-art LLMs, including Gemini and Llama 3 70B. Among all tested models, GPT-4 Turbo consistently achieved the best trade-off between performance and computational cost. For reproducibility, we set the temperature to 0. Detailed results for all models are provided in the paper's supplementary material.

Unsupervised Pipeline The unsupervised pipeline extracts claims directly from the full text without relying on citations or manual annotations. The process begins with text preprocessing, including section segmentation and normalization (e.g., lowercasing, whitespace handling, and removal of extraneous formatting) to improve input quality.

Algorithm 1: Weakly Supervised Claim Extraction Flow

Input: Full text of research paper, claim-focused citances
Preprocess text and citances
Step 1: Align citances and claims
for each citance s **do**
  for each candidate claim c in paper **do**
    Compute match score $DM(c, s)$ using the rubric
    if $DM(c, s) \geq 7$ **then**
      Mark pair (c, s) as a strong match
    end if
  end for
end for
Step 2: Select high-quality training examples
for each paper with ≥ 6 strong matches **do**
  Include aligned pairs in training data
end for
Step 3: Fine-tune LLM
Train the model using selected pairs as weak supervision
Output: Fine-tuned model for robust claim extraction

The cleaned text is then fed to the LLM along with instructions derived from our claim rubric (Wei 2023), which guides the model to focus on statements that express improvements or superiority, propose novel methods or ideas, or describe new findings or cause-and-effect relationships.

For each paper, the LLM identifies candidate claims, associates them with their contextual sections, and assigns each to one of four broad themes: *Novelty*, *Performance*, *Applicability*, or *Background*. While this categorization abstracts over the diversity of claims in scientific writing, it offers a practical structure for downstream analysis. The framework is easily extensible to accommodate alternative or field-specific taxonomies.

Weakly Supervised Pipeline The weakly supervised pipeline extends the unsupervised approach by incorporating a crucial alignment step with claim-focused citances, as illustrated in Algorithm 1. The central idea is to treat citances as weak labels: they represent how the research community highlights a paper's key contributions, providing naturally occurring supervision at scale. To exploit this signal, we apply the same claim rubric to both candidate claims and citances, computing a semantic match score $DM(c, s)$ for each claim (c) and citance (s).

Fine-Tuning

The semantically strong matched claims are then used to fine-tune the language model, encouraging it to prioritize claims that align with community-validated evidence while remaining robust to noisy supervision. Specifically, the model is fine-tuned on pairs of (full paper text, core claims), where the core claims are derived from the six strongly matched claim–citance pairs per paper. This setup aims to guide the model toward generating core claims directly from the full text of a paper (see Algorithm 1).

Evaluation

We evaluated claim extraction quality by measuring the semantic alignment between extracted claims and their corresponding claim-focused citances using two methods: (1) LLM-based evaluation, where a large language model(GPT-4, (OpenAI 2023)) scored the semantic alignment on a 0–10 scale, and (2) an embedding-based approach (SBERT, (Reimers and Gurevych 2019)), where we computed cosine similarity between claim and citance embeddings.

To verify the reliability of the LLM-based evaluation and mitigate potential biases, we conducted a manual validation study. Two independent human annotators assessed a randomly sampled set of 100 claim–citance pairs, following sample sizes consistent with prior evaluator calibration studies. They judged whether the LLM's scores reflected true semantic alignment (See human annotation instruction in the supplementary materials). Human validation showed that 92% of the LLM's scores were within a ± 1 difference from human ratings. Inter-annotator agreement, measured using Cohen's κ, was $\kappa = 0.85$, indicating substantial agreement and confirming that the LLM scores are highly reliable.

For comparison, we also evaluated an embedding-based cosine similarity metric. Cosine similarity values (originally ranging from 0 to 1) were linearly normalized to the same 0–10 scale as the LLM scores to ensure consistency. Although this method achieved 78% consistency with human annotations ($\kappa = 0.79$), it underperformed relative to the LLM-based evaluation.

For both evaluation methods, we classified a pair as a match if its score was ≥ 7. This threshold was empirically chosen to balance precision and coverage while filtering out noisy pairs. Based on the manual human evaluation, which aligned more closely with the LLM scores than with the cosine similarity score, we used the LLM scores to calculate two key metrics: coverage and precision, as follows.

Definitions

Let us define the following sets and function:

- Let $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ be the set of extracted claims from a paper.
- Let $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ be the set of citances (sentences in other papers that cite the paper in question).
- For each pair (c_i, s_j), the LLM assigns a **degree of match** score $DM(c_i, s_j)$, which ranges from 0 to 10. This can be formally represented as:

$$DM : \mathcal{C} \times \mathcal{S} \to [0, 10].$$

A claim-citance pair (c_i, s_j) is deemed a **match** if it satisfies the condition:

$$DM(c_i, s_j) \geq DM_{\text{th}},$$

where the degree of match threshold is defined as DM_{th}.

Coverage is defined as:

$$\frac{|\{s_j \in \mathcal{S}_{\text{filtered}} \mid s_j \text{ matches some } c_i \in \mathcal{C}\}|}{|\mathcal{S}_{\text{filtered}}|}$$

where $\mathcal{S}_{\text{filtered}}$ is the set of filtered citances based on the rubric defined in (Wei 2023).

Method	Precision	Coverage
Abstract-only		
SciBERT Claim Classifier (Abstract-only)	0.51	0.24
Claim Distiller (Abstract-only)	0.55	0.26
Unsupervised Pipeline (Abstract-only)	0.70	0.58
Weakly Supervised Pipeline (Abstract-only)	**0.74**	**0.62**
Full-text		
SciBERT Claim Classifier (Full-text)	0.42	0.49
Claim Distiller (Full-text)	0.23	0.68
Unsupervised Pipeline (Full-text)	0.56	0.78
Weakly Supervised Pipeline (Full-text)	**0.63**	**0.79**

Table 2: Comparison of Average Precision and Coverage Across Claim Extraction Methods

Similarly, **Precision** is defined as:

$$\frac{|\{c_i \in \mathcal{C} \mid c_i \text{ matches some } s_j \in \mathcal{S}_{\text{filtered}}\}|}{|\mathcal{C}|}.$$

After applying the evaluation procedure across all papers in the dataset, we computed the following average metrics:

- **Average Coverage**:

$$\overline{\text{Coverage}} = \frac{1}{P}\sum_{k=1}^{P} \text{Coverage}_k$$

  where P is the total number of papers evaluated, and Coverage_k represents the coverage for each paper k.

- **Average Precision**:

$$\overline{\text{Precision}} = \frac{1}{P}\sum_{k=1}^{P} \text{Precision}_k$$

  where P is the total number of papers evaluated, and Precision_k represents the precision for each paper k.

- A higher average coverage indicates that the extracted claims effectively encompass more of the content that other papers have cited, suggesting comprehensive claim extraction.

- A higher average precision suggests that the extracted claims are highly relevant to the content cited by other papers, indicating a high level of accuracy in extraction.

Results

Comparison with Existing Methods

We compared our unsupervised and weakly-supervised claim extraction pipelines with two baselines:

ClaimDistiller. ClaimDistiller (Wei 2023) is a claim-worthiness classifier trained on manually annotated abstracts. For full-text evaluation, we applied ClaimDistiller sentence-by-sentence. However, its performance is limited by domain shift since it was trained only on abstracts.

SciBERT Classifier. We also trained a SciBERT-based binary classifier (Beltagy, Lo, and Cohan 2019) using our labeled dataset. Fine-tuning SciBERT allowed better handling of scientific language, but performance remained below our pipelines.

As shown in Table 2, both our unsupervised and weakly-supervised pipelines outperform the baselines in average precision and coverage for both abstract and full-text settings. The weakly-supervised approach yields the highest precision, demonstrating the benefit of leveraging weak supervision for claim identification.

Section-Based and Thematic Analyses

We further analyzed claim extraction across paper sections and claim themes.

Section-based Analysis. Sections were mapped to seven standardized categories (Abstract, Introduction, Background, Methods, Results, Discussion, and Conclusion) using LLM-guided mapping. Table 3 shows that the weakly-supervised pipeline consistently outperforms the unsupervised pipeline in precision and coverage across all sections, except for Introduction where coverage remains unchanged.

Interestingly, the unsupervised pipeline extracts the most claims from Methods, likely due to abundant procedural text. In contrast, the weakly-supervised pipeline focuses more on Abstracts, aligning with their role in summarizing key contributions. Figure 2 visualizes these distributions.

Thematic Analysis. We categorized extracted claims into four themes: *Novelty*, *Performance*, *Applicability*, and *Background*. This taxonomy is inspired by and aligns with prior work on scientific discourse and rhetorical zoning (Teufel, Moens, and Ananiadou 2000; Meyer et al. 2012; Liakata et al. 2012). Table 4 summarizes the average precision and coverage for each theme. The weakly supervised pipeline consistently improves both average precision and coverage across all themes, with particularly large gains in Applicability (+10% average precision) and Background (+10% average precision). Moreover, the model extracts proportionally more *Novelty* claims, which are crucial for identifying original research contributions (Figure 3).

Section	Unsupervised		Weakly Supervised	
	Precision	Coverage	Precision	Coverage
Abstract	0.70	0.58	**0.74**$^{+4}$	**0.62**$^{+4}$
Introduction	0.69	0.57	**0.73**$^{+4}$	0.57
Discussion	0.56	0.41	**0.64**$^{+8}$	**0.45**$^{+4}$
Background	0.53	0.36	**0.68**$^{+15}$	**0.56**$^{+20}$
Results	0.45	0.30	**0.53**$^{+8}$	**0.32**$^{+2}$
Methods	0.50	0.39	**0.55**$^{+5}$	**0.42**$^{+3}$
Conclusion	0.40	0.21	**0.60**$^{+20}$	**0.37**$^{+16}$

Table 3: Performance Comparison of Unsupervised and Weakly-Supervised Models by Section. This table highlights the average precision and coverage improvements across different sections of papers.

Theme	**Unsupervised**		**Weakly Supervised**	
	Prec.	**Cov.**	**Prec.**	**Cov.**
Novelty	0.70	0.64	**0.72**$^{+2}$	**0.65**$^{+1}$
Performance	0.47	0.36	**0.53**$^{+6}$	**0.37**$^{+1}$
Applicability	0.54	0.38	**0.64**$^{+10}$	**0.40**$^{+2}$
Background	0.47	0.27	**0.57**$^{+10}$	**0.32**$^{+5}$

Table 4: Comparison of Precision and Coverage metrics for Unsupervised and Weakly Supervised models across claim themes.

Both pipelines extract claims across all sections and themes, with the weakly supervised model outperforming the unsupervised one in average precision and coverage, especially for abstracts and claims categorised as novel.

Out-of-Domain Validation in Uncited/Low-Citation Papers

Method	**Precision**	**Coverage**
Claim Distiller (Full-text)	0.21	0.58
SciBERT Claim Classifier (Full-text)	0.38	0.42
Unsupervised Pipeline (Full-text)	0.51	0.69
Weakly Supervised Pipeline (Full-text)	**0.58**	**0.72**

Table 5: Out-of-domain validation on 20 computational biology papers with few or no citations. Results follow the same LLM-based evaluation protocol as the main results. The weakly supervised pipeline maintains the best precision and coverage, consistent with in-domain findings.

To evaluate the generalizability of our approach beyond computer science, we applied the same extraction and evaluation pipeline to 20 computational biology papers with few or no citations. Gold core claims were created by two expert annotators, with any inter-annotator disagreements resolved through discussion (see extraction protocols in the supplementary materials).

As shown in Table 5, the weakly supervised pipeline achieves the highest precision (0.58) and coverage (0.72), closely mirroring its in-domain performance. This suggests strong cross-domain robustness despite disciplinary and stylistic differences between fields.

Ablation

Sensitivity to Degree-of-Match Threshold

We analyze how varying the degree-of-match threshold DM_{th} affects evaluation metrics for both the *Unsupervised* and *Weakly Supervised* pipelines under *Abstract-only* and *Full-text* settings. Lower thresholds (e.g., 6) increase coverage but reduce precision, whereas higher thresholds (e.g., 8) improve precision at the expense of coverage. Threshold 7 strikes a balance, as highlighted in Table 6.

Discriminative Fine-Tuning We first evaluated a discriminative fine-tuning strategy by training the LLM as a binary claim classifier using positive/negative samples generated from our dataset. Applying this model to claims extracted by the unsupervised pipeline increased full-text precision from **0.56 to 0.66**, but reduced coverage from **0.78 to 0.60** due to aggressive filtering.

Setting	Threshold	LLM-Prec.	LLM-Cov.	Cosine-Prec.	Cosine-Cov.
Abstract-only					
Unsupervised	6	0.64	0.83	0.48	0.76
	7	0.70	0.58	0.53	0.59
	8	0.74	0.48	0.58	0.45
Weakly Supervised	6	0.68	0.85	0.50	0.78
	7	0.74	0.62	0.56	0.63
	8	0.78	0.52	0.61	0.48
Full-text					
Unsupervised	6	0.55	0.82	0.44	0.75
	7	0.56	0.78	0.47	0.68
	8	0.60	0.68	0.52	0.55
Weakly Supervised	6	0.60	0.84	0.46	0.76
	7	0.63	0.79	0.49	0.70
	8	0.67	0.70	0.54	0.56

Table 6: Effect of threshold variation (DM_{th}) on average precision and coverage for Unsupervised and Weakly Supervised pipelines, evaluated with both LLM-based and cosine similarity metrics.

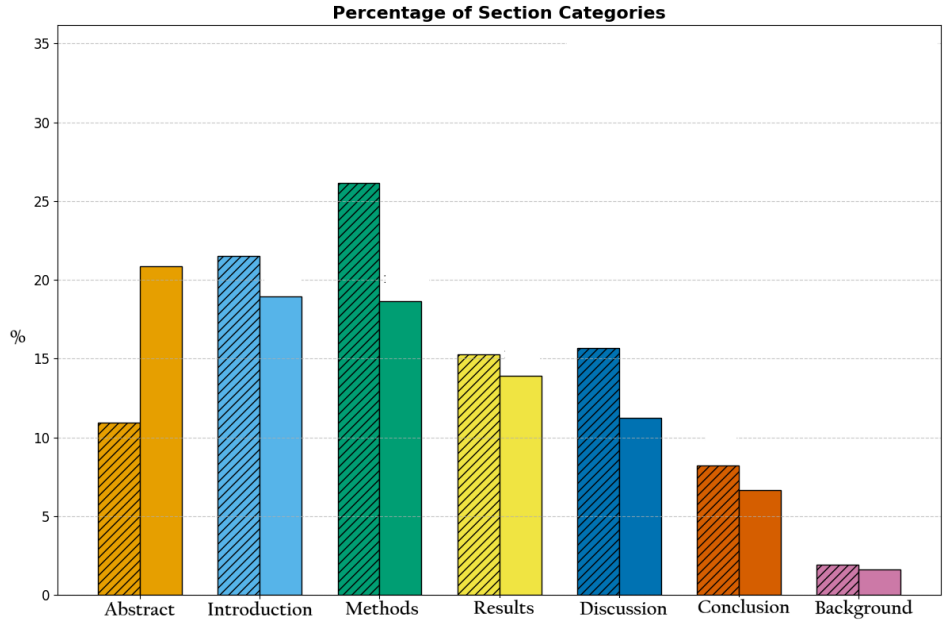

Figure 2: Section-based differences in claim extraction. The weakly-supervised model (dashed lines) extracts more claims from Abstracts, where it also achieves the highest precision and coverage (see Table 3).

Takeaway: Fine-tuning improves precision at the expense of coverage, reflecting a conservative claim selection.

Two-Phase Extraction Next, we tested a two-phase unsupervised extraction. In phase one, the LLM extracted claims from individual subsections to exploit shorter context windows; in phase two, it clustered and deduplicated claims. Contrary to expectations, this approach did not improve precision or coverage compared to the one-phase pipeline.
Takeaway: Despite better handling of context, the added complexity offered no measurable benefit, so we retained the simpler one-phase design.

Analysis of Failure Cases in Claim–Citance Matching

To diagnose weaknesses, we analyzed 100 claims that lacked strong matches with citances. Three main patterns emerged:

- **Novel model names (60%)** — Claims introducing abbreviated model names (e.g., *MIML-RE*, *SHARK2*) lacked sufficient context for citance matching.
- **Quantitative claims (30%)** — Highly specific numerical comparisons were often absent in citances, leading to mismatches.
- **Implicit background claims (10%)** — Well-known claims among experts were not explicitly restated in citances, hindering alignment.

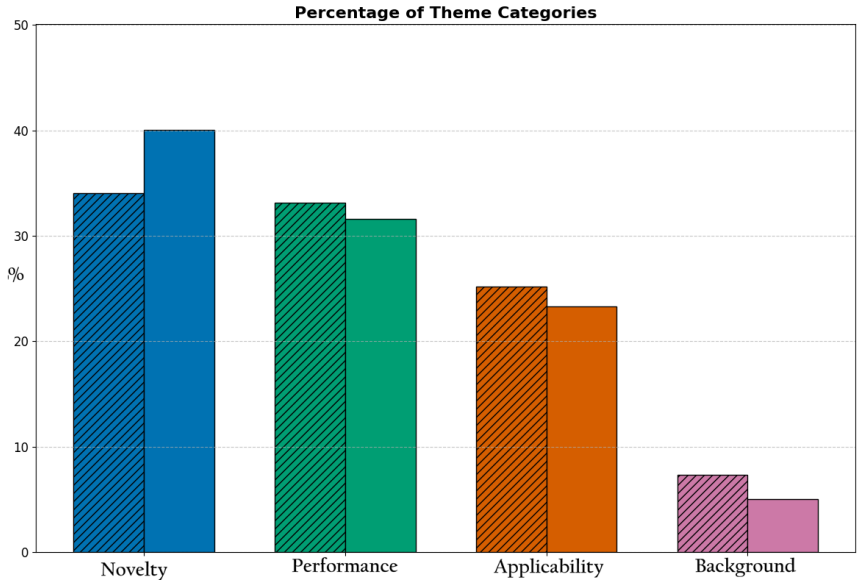

Figure 3: Thematic differences in claim extraction between unsupervised and weakly-supervised models on the test set. The weakly-supervised model (dashed lines) increases extraction of Novelty claims, which achieves the highest precision and coverage (see Table 4), indicating its effectiveness at prioritizing novel contributions.

Takeaway: Edge cases expose limitations in citance quality, particularly with novel terminology and numerical details. Addressing these gaps could further improve matching performance in future work.

Related Work

Automated claim extraction and evaluation spans LLM-based assessment, claim classification, scientific discourse, and bibliometrics. Liang et al. (2023) show partial alignment between LLM feedback and human reviewers, suggesting utility for assessing scientific content and motivating the use of LLMs as weak supervisors or evaluators.

Most claim-extraction work targets abstracts or narrow domains. Sternfeld et al. (2024) model claims as subject–predicate–object triplets to flag inconsistencies in biomedical abstracts; Wei (2023) use contrastive learning to classify claim sentences; and Achakulvisut et al. (2019) combine rule-based and deep methods on abstract-centric datasets. Open-domain systems like EnClaim (Saha, Sinha, and Dasgupta 2024) and ClaimBuster (Hassan et al. 2017) emphasize political/general claims, while Wright and Augenstein (2021) study cite-worthiness in scientific text, which relates to prioritizing salient statements.

Scientific claim verification is related but distinct. Alvarez, Bennett, and Wang (2024) propose a scalable, open-source framework for zero-shot claim generation and verification using full articles and citances; we regard it as a strong related baseline. Our goal is automated identification and summarization of core, paper-level claims in the target article, emphasizing scalable extraction and organization across sections rather than support/refutation, though our outputs can feed verification pipelines.

Discourse work categorizes claims by document structure (Kiepura et al. 2024), and relational modeling of support/refutation (Li et al. 2022; Tan et al. 2023; Özkan Tan et al. 2023) motivates graph-based organization of extracted claims. Controlled-language approaches like AIDA (Jansen and Kuhn 2017) normalize claims for downstream linking; we instead operate over full texts without controlled formulations or manual annotations, leveraging citances as weak supervision.

In bibliometrics, mention extraction and analysis (Petrovich et al. 2024) complements our citance-based salience modeling by characterizing how, where, and why papers are cited. Automated retrieval of publication–citation records (Ruths and Al Zamal 2010) underpins scalable harvesting of citation networks and citances, which our pipeline builds upon. Cross-lingual citation practices (Saier, Färber, and Tsereteli 2022) inform our treatment of generalizability to multilingual and interdisciplinary corpora and highlight challenges for extending citance-driven extraction beyond English.

Work on identifying impactful research via text and figures (Stamenovic, Schick, and Luo 2017) is complementary to automated claim evaluation and suggests linking extracted claims to impact signals and figure-derived evidence. Robustness is critical when citances reference flawed work: Heibi and Peroni (2022) analyze incoming citations to retracted articles, and Meng, Varol, and Barabási (2024) outline broader limits and biases of citation/citance-based signals that we consider in our Limitations/Impact section.

Evidence-synthesis literature often extracts structured elements (e.g., PICO—Population, Intervention, Comparison, and Outcome) rather than compact, paper-level claims (Schmidt et al. 2021; Jonnalagadda, Goyal, and Huffman 2015). Methods for automatic evidence retrieval in system-

atic reviews (Choong et al. 2014) and clinical information extraction (Ford et al. 2016) offer transferable modeling and evaluation practices relevant to integrating claim extraction into review workflows.

Domain-specific full-text extraction spans clinical trial characteristics (ExaCT) (Kiritchenko et al. 2010), HIV treatment insights (Biziukova et al. 2020), and additive manufacturing (Feldhoff et al. 2025). Recently, LLMs have been applied to extract numerical Randomized Controlled Trial (RCT) results (Yun et al. 2024), underscoring challenges in precise quantitative claims. Our contribution complements these efforts by targeting compact, paper-level core claims across full texts, using citances as weak supervision and as evaluation signals. Our contribution complements these efforts by targeting compact, paper-level core claims across full texts, using citances as weak supervision and as evaluation signals.

In summary, beyond abstract-only detection, bibliometric mention extraction, and verification, we offer a full-text, low-supervision framework that leverages citances for scalable training and evaluation; it integrates with verification and impact-assessment pipelines, addresses cross-lingual and interdisciplinary settings, and is supported by section- and theme-level analyses and a curated computer science dataset.

Conclusion

We introduce unsupervised and weakly supervised pipelines for full-text claim extraction that leverage citances to minimize reliance on costly manual annotations. Unlike prior approaches, our method combines rubric-based citance filtering, LLM-calibrated claim alignment, and a fully replicable, scalable training and evaluation protocol.

Our contributions are threefold: (i) a low-supervision extraction framework that outperforms sentence- and abstract-only baselines in precision and coverage, (ii) a citance-anchored evaluation methodology calibrated with human judgments, and (iii) a publicly available dataset of over 1,200 computer science papers with claim-focused citances and a reproducible pipeline for filtering, alignment, and evaluation.

Experimental results show that both pipelines surpass existing methods, with the weakly supervised approach achieving the highest precision and coverage. Section-wise analysis shows strong performance for Abstracts and Methods, while thematic analysis highlights high gains for Novelty and Applicability claims. Out-of-domain validation on low-citation computational biology papers confirms robust generalization.

Overall, integrating LLMs with citance-guided signals can significantly enhance automated claim extraction, offering practical tools for understanding and analyzing scientific literature at scale.

Limitations

This work centers on the computer science domain, selected for its accessibility, scale, and diversity of research styles. While the focus ensures methodological consistency, extending the approach to other scientific fields presents an opportunity for broader validation. The use of well-cited papers in training supports data quality and clarity, and the models are readily applicable to newer or less-cited works. Empirical thresholds for identifying claim–citance matches were calibrated for stability, as confirmed through sensitivity analyses. Although large-scale full-text processing requires notable computational resources, continued advances in model efficiency and infrastructure will further enhance scalability.

References

Achakulvisut, T.; Bhagavatula, C.; Acuna, D.; and Kording, K. 2019. Claim Extraction in Biomedical Publications using Deep Discourse Model and Transfer Learning. *arXiv [cs.CL]*.

Alvarez, C.; Bennett, M.; and Wang, L. L. 2024. Zero-shot Scientific Claim Verification Using LLMs and Citation Text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, 269–276.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv [cs.CL]*.

Biziukova, N.; Tarasova, O.; Ivanov, S.; and Poroikov, V. 2020. Automated extraction of information from texts of scientific publications: Insights into HIV treatment strategies. *Front. Genet.*, 11: 618862.

Choong, M. K.; Galgani, F.; Dunn, A. G.; and Tsafnat, G. 2014. Automatic evidence retrieval for systematic reviews. *J. Med. Internet Res.*, 16(10): e223.

Feldhoff, K.; Wiemer, H.; Träger, P.; Kühne, R.; Zimmermann, M.; and Ihlenfeldt, S. 2025. Automatic information extraction from scientific publications based on the use case of additive manufacturing. *Appl. Sci. (Basel)*, 15(17): 9331.

Ford, E.; Carroll, J. A.; Smith, H. E.; Scott, D.; and Cassell, J. A. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J. Am. Med. Inform. Assoc.*, 23(5): 1007–1015.

Hassan, N.; Zhang, G.; Arslan, F.; Caraballo, J.; Jimenez, D.; Gawsane, S.; Hasan, S.; Joseph, M.; Kulkarni, A.; Nayak, A. K.; Sable, V.; Li, C.; and Tremayne, M. 2017. ClaimBuster: The First-ever End-to-end Fact-checking System. *Proc. VLDB Endow.*, 10: 1945–1948.

Heibi, I.; and Peroni, S. 2022. A protocol to gather, characterize and analyze incoming citations of retracted articles. *PLoS One*, 17(7): e0270872.

Jansen, T.; and Kuhn, T. 2017. Extracting Core Claims from Scientific Articles. In *Proceedings (verify venue: workshop/-conference)*. Foundational AIDA-based core claim extraction. Please verify venue/URL/DOI.

Jonnalagadda, S. R.; Goyal, P.; and Huffman, M. D. 2015. Automating data extraction in systematic reviews: a systematic review. *Syst. Rev.*, 4(1): 78.

Kiepura, A.; Gao, Y.; Lam, J.; Gu, N.; and Hahnloser, R. H. R. 2024. SciPara: A new dataset for investigating paragraph discourse structure in scientific papers.

Kiritchenko, S.; de Bruijn, B.; Carini, S.; Martin, J.; and Sim, I. 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med. Inform. Decis. Mak.*, 10(1): 56.

Knoth, P.; Herrmannova, D.; Cancellieri, M.; Anastasiou, L.; Pontika, N.; Pearce, S.; Gyawali, B.; and Pride, D. 2023. CORE: A Global Aggregation Service for Open Access Papers. *Sci Data*, 10(1): 366.

Li, M.; Gangi Reddy, R.; Wang, Z.; Chiang, Y.-S.; Lai, T.; Yu, P.; Zhang, Z.; and Ji, H. 2022. COVID-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 135–144. Stroudsburg, PA, USA: Association for Computational Linguistics.

Liakata, M.; et al. 2012. A multi-dimensional annotation scheme for classifying claims and evidence in biomedical full-text articles. In *BioNLP 2012*, 50–59.

Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D.; Yin, Y.; McFarland, D.; and Zou, J. 2023. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *arXiv [cs.LG]*.

Meng, X.; Varol, O.; and Barabási, A.-L. 2024. Hidden citations obscure true impact in science. *PNAS Nexus*, 3(5): gae155.

Meyer, B.; et al. 2012. Discourse structure and scientific knowledge: rhetorical moves and their discourse functions. *Discourse Studies*, 14(2): 145–164.

OpenAI. 2023. GPT-4 Technical Report. https://openai.com/research/gpt-4.

Petrovich, E.; Verhaegh, S.; Bös, G.; Cristalli, C.; Dewulf, F.; van Gemert, T.; and IJdens, N. 2024. Bibliometrics beyond citations: introducing mention extraction and analysis. *Scientometrics*, 129(9): 5731–5768.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv [cs.CL]*.

Ruths, D.; and Al Zamal, F. 2010. A method for the automated, reliable retrieval of publication-citation records. *PLoS One*, 5(8): e12133.

Saha, D.; Sinha, M.; and Dasgupta, T. 2024. EnClaim: A Style Augmented Transformer Architecture for Environmental Claim Detection. In *Natural Language Processing meets Climate Change @ ACL 2024*.

Saier, T.; Färber, M.; and Tsereteli, T. 2022. Cross-lingual citations in English papers: a large-scale analysis of prevalence, usage, and impact. *Int. J. Digit. Libr.*, 23(2): 179–195.

Schmidt, L.; Finnerty Mutlu, A. N.; Elmore, R.; Olorisade, B. K.; Thomas, J.; and Higgins, J. P. T. 2021. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. *F1000Res.*, 10: 401.

Stamenovic, M.; Schick, S.; and Luo, J. 2017. Machine identification of high impact research through text and image analysis. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 98–104. IEEE.

Sternfeld, A.; Kucharavy, A.; David, D. P.; Mermoud, A.; and Jang, J. 2024. LLM-resilient bibliometrics: Factual consistency through entity triplet extraction. In *EEKE-AII2024*, 85–93.

Tan, N.; Nguyen, T.; Bensemann, J.; Peng, A.; Bao, Q.; Chen, Y.; Gahegan, M.; and Witbrock, M. 2023. Multi2Claim: Generating Scientific Claims from Multi-Choice Questions for Scientific Fact-Checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2652–2664. Dubrovnik, Croatia: Association for Computational Linguistics.

Tan, N. O.; Tandon, N.; Wadden, D.; Tafjord, O.; Gahegan, M.; and Witbrock, M. 2024. Faithful reasoning over scientific claims. *Proceedings of the AAAI Symposium Series*, 3(1): 263–272.

Teufel, S.; Moens, M.-F.; and Ananiadou, S. 2000. Argumentative zoning: Information extraction from scientific text. In *Proceedings of the 38th annual meeting on association for computational linguistics*, 110–117. Association for Computational Linguistics.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550. Online: Association for Computational Linguistics.

Wadden, D.; Lo, K.; Kuehl, B.; Cohan, A.; Beltagy, I.; Wang, L. L.; and Hajishirzi, H. 2022. SciFact-Open: Towards open-domain scientific claim verification. *arXiv [cs.CL]*.

Wei, X. 2023. ClaimDistiller: Scientific claim extraction with supervised contrastive learning. 65–77.

Wright, D.; and Augenstein, I. 2021. CiteWorth: Cite-worthiness detection for improved scientific document understanding. *arXiv [cs.CL]*.

Yun, H. S.; Pogrebitskiy, D.; Marshall, I. J.; and Wallace, B. C. 2024. Automatically Extracting Numerical Results from Randomized Controlled Trials with Large Language Models. In *Machine Learning for Healthcare Conference*. PMLR.

Özkan Tan, N.; Yuxuan Peng, A.; Bensemann, J.; Bao, Q.; Hartill, T.; Gahegan, M.; and Witbrock, M. 2023. Input-length-shortening and text generation via attention values. *arXiv e-prints*, arXiv:2303.07585.