Towards Scientific Contradiction Mining: Topic Discovery and Claim Extraction from AI Literature

Andrej Natev
an49507@student.uni-lj.si
Faculty of Computer and Information Science, University of Ljubljana
Ljubljana, Slovenia

Elham Motamedi
elham.motamedi@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Inna Novalija
inna.koval@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Scientific contradiction mining requires identifying papers that discuss comparable problems and recovering the empirical findings that may support or contradict one another. This work evaluates two prerequisites for such a pipeline. First, we investigate whether consensus across three scientific document encoders can identify stable and coherent topic groups. Second, we examine whether restricting extraction to explicitly result-oriented sections is sufficient or whether complete full-text processing is necessary. Experiments are conducted on 4,227 papers from the arXiv Computer Science – Artificial Intelligence category (`cs.AI`) using SciNCL, SPECTER, SPECTER2, and BERTopic. The strongest clustering configuration achieved a three-way Jaccard coefficient of 0.885, while manual evaluation of seven high-confidence groups produced 97.5% topic coherence. In a focused full-text case study, 1,147 automatically extracted candidates were manually evaluated, yielding 510 valid result-bearing findings. Result-oriented sections yielded substantially higher extraction precision than other sections (72.0% vs 34.1%) and a higher proportion of findings containing quantitative evidence (55.8% vs 37.3%). However, 55.7% of all validated findings occurred outside result-oriented sections, indicating a trade-off between precision and quantitative density on the one hand and broader recovery of validated findings on the other. These evaluations also produced two manually annotated reference datasets for topic coherence and scientific finding extraction. These results support a staged contradiction-mining pipeline in which cross-encoder consensus narrows the literature and full-text processing supplies evidence for subsequent claim decontextualization and contradiction detection.

Keywords

scientific literature mining, document embeddings, topic modeling, scientific claim extraction

1 Introduction

The increasing volume of scientific literature makes it difficult to determine how findings reported across different studies relate to one another. Papers addressing similar problems may report supporting, conflicting, or apparently contradictory results because of differences in datasets, models, experimental conditions, or evaluation procedures. Automatically identifying such relationships could support literature review, evidence synthesis, and the discovery of unresolved research questions.

Scientific contradiction mining requires more than identifying documents that discuss similar topics. Relevant papers must first be grouped reliably, after which their empirical findings must be recovered and transformed into comparable scientific claims. Previous work on scientific claim verification emphasises atomicity (a claim should express a single proposition), faithfulness (it should preserve the meaning of the source text), and self-containment (it should remain interpretable without relying on surrounding context) as important properties of claim formulation, since claims with these properties are a prerequisite for reliable comparison and downstream contradiction detection [9, 12, 5]. Contradiction-detection studies further show that contextual differences may distinguish genuine contradictions from condition-dependent results [4, 3].

We therefore investigate two prerequisites for scientific contradiction mining:

- **RQ1:** Does cross-encoder consensus identify stable and coherent scientific topic groups?
- **RQ2:** To what extent does full-text processing recover result-bearing findings beyond explicitly result-oriented sections?

RQ1 is evaluated on a corpus of 4,227 papers from the arXiv Computer Science – Artificial Intelligence category (`cs.AI`), while RQ2 is examined through a focused full-text case study within one high-confidence consensus group. The complete intended pipeline is illustrated in Figure 1. The present work evaluates topic modelling and scientific finding extraction, while reliable atomic claim decontextualization and contradiction detection are left for future work.

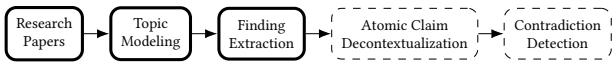


Figure 1: Intended scientific contradiction-mining pipeline. Solid stages are investigated in this work, while dashed stages are left for future work.

2 Related Work

Scientific document encoders such as SPECTER [1], SciNCL [6], and SPECTER2 [8] provide semantic representations for organising scientific literature, while BERTopic combines embeddings, dimensionality reduction, density-based clustering, and class-based TF–IDF to construct interpretable topic groups [2]. We use these representations to examine agreement across separately constructed encoder-specific topic spaces.

Scientific claim-verification systems generally assume that a claim is already available. SciFact introduced expert-written claims paired with supporting or refuting evidence [9], while later

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.119

work studied automatic claim generation and refinement [12, 5] and full-document or open-domain verification [11, 10]. Contradiction-focused studies further show that apparent conflicts may depend on contextual and experimental conditions [4, 3].

Our work addresses an earlier stage of this pipeline: rather than starting from predefined claims or candidate statement pairs, we first narrow a broad scientific corpus through cross-encoder topic consensus and then recover result-bearing findings from complete paper text for subsequent decontextualization and contradiction analysis.

3 Methods

This section describes the data and methodology used to address the two research questions. We first present the construction of the scientific paper corpus, followed by the cross-encoder topic modelling and consensus procedure linked to RQ1 and the full-text scientific finding extraction and validation procedure linked to RQ2.

3.1 Corpus

We used the structured UnarXive collection [7]. From an initial computer-science subset of 226,808 papers, we selected publications whose category was exactly `cs.AI`, retained English documents with cleaned abstracts of at least 200 characters, and removed duplicate paper–abstract combinations. The final corpus contained 4,227 papers. Titles and abstracts were used for topic discovery, while structured full-text sections were retained for subsequent finding extraction.

3.2 Topic Modelling and Cross-Encoder Consensus

Each paper was separately represented using SciNCL [6], SPECTER [1], and SPECTER2 [8]. Preliminary experiments also considered several general-purpose and scientific representations, including MiniLM, MPNet, DistilBERT/DistilRoBERTa, and SciBERT; SciNCL, SPECTER, and SPECTER2 were retained because qualitative inspection during preliminary experiments indicated clearer and more interpretable topic structures for the subsequent consensus analysis. BERTopic was then applied separately to each representation using UMAP and HDBSCAN.

The HDBSCAN minimum cluster size specifies the smallest group that can be retained as a cluster, whereas minimum samples controls the conservativeness of the density criterion and therefore the tendency to classify documents as noise. Several parameter combinations were explored during preliminary experiments. Two configurations producing useful topic structures at different resolutions were retained: coarse clustering with a minimum cluster size of 15 and minimum samples of 3, and granular clustering with values of 10 and 2, respectively. These configurations were selected empirically during exploratory analysis rather than optimized against the subsequent manual coherence labels.

Because numerical topic identifiers are specific to each encoder, consensus groups were defined through shared paper membership rather than topic numbers. Paper-level topic assignments from SciNCL, SPECTER, and SPECTER2 were merged, and papers classified as outliers by any encoder were excluded. Each observed combination of three encoder-specific topic assignments was treated as a candidate consensus group. Cross-encoder agreement was quantified using the strict three-model intersection, three-way and pairwise Jaccard coefficients, and model retention, defined as the proportion of each encoder-specific topic preserved in the strict three-model intersection.

We evaluated both the original HDBSCAN assignments (*native*) and assignments obtained after BERTopic outlier reassignment (*reduced*), producing four configurations: coarse/native, coarse/reduced, granular/native, and granular/reduced.

Consensus groups were required to contain at least 10 papers in their strict intersection. Eligible groups were ranked by three-way Jaccard coefficient, followed by minimum model retention and intersection size. Seven high-confidence granular/native groups additionally satisfied a three-way Jaccard coefficient of at least 0.50 and minimum model retention of at least 0.50. Papers from the union of the corresponding encoder topics were manually rated as clearly relevant (2), partially relevant (1), or unrelated (0). This produced a manually annotated reference dataset of 280 paper–topic judgments for evaluating topic coherence. Topic coherence was calculated as the proportion of paper–topic judgments rated clearly relevant (2). For comparison with individual encoder-specific topics, the same annotations were restricted to the papers belonging to each encoder-specific topic and to the strict three-encoder intersection. Coherence was calculated separately for SciNCL, SPECTER, SPECTER2, and the strict consensus, and macro-averaged across the seven evaluated groups.

3.3 Scientific Finding Extraction and Validation

For RQ2, we selected the 21-paper strict intersection of the backdoor and adversarial-attack consensus group as a focused case study. The group exhibited strong cross-encoder agreement (J_3 = 0.778) and an experimental focus suitable for evaluating result-bearing finding extraction. Structured full text was available for 20 papers, all of which were processed by a sequential full-text reader based on the 4-bit quantized Qwen2.5-7B-Instruct model. Each paper was divided into text chunks that were processed in document order. From each chunk, the model extracted candidate result-bearing findings together with their source section and supporting textual evidence.

We consider experimental outcomes, comparisons, observed effects, robustness results, ablations, sensitivity analyses, transferability, generalisation, and failure cases to be result-bearing findings. Method descriptions, experimental setup, dataset descriptions, metric definitions, research questions, background information, and future work are excluded.

All 1,147 extracted candidates were manually classified as either valid result-bearing findings or non-findings. This produced a second manually annotated reference dataset for evaluating scientific finding extraction.

For section-level analysis, headings were classified using a keyword-based rule. A section was considered *result-oriented* when its heading indicated empirical evaluation, including conventional labels such as Results, Experiments, and Evaluation and specialised terms referring to performance, comparison, ablation, effectiveness, robustness, transferability, and empirical analysis. All remaining headings were classified as other full-text sections. Extraction restricted to result-oriented sections was treated as a section-selection baseline against which complete full-text processing was compared.

We additionally identified explicit quantitative evidence using a conservative pattern-based rule covering percentages, scores

and decimal values, rates, counts, and measured values with units. Bare numeric identifiers, such as numbers occurring only in model or dataset names, were excluded.

4 Results

This section presents the empirical results for both research questions. We first evaluate the stability and semantic coherence of the cross-encoder topic groups for RQ1, and then examine extraction precision and the distribution of validated and quantitative findings across result-oriented and other full-text sections for RQ2.

Table 1: Cross-encoder consensus across clustering configurations. Best J_3 denotes the highest three-way Jaccard coefficient, and $R_{\min}$ is the minimum model retention of that same top-ranked consensus group. Mean top-10 J_3 is the mean three-way Jaccard coefficient across the ten highest-ranked eligible groups.

Configuration	Groups	Best J_3	$R_{\min}$	Mean top-10 J_3
Coarse/native	53	0.846	0.880	0.530
Coarse/reduced	83	0.636	0.695	0.351
Granular/native	38	**0.885**	**0.896**	**0.613**
Granular/reduced	68	0.601	0.664	0.391

4.1 Topic Stability and Coherence

Table 1 summarises cross-encoder agreement for the four clustering configurations. The granular/native configuration produced the strongest high-confidence consensus, achieving a best three-way Jaccard coefficient of 0.885, minimum model retention of 0.896, and a mean top-10 Jaccard of 0.613.

Outlier reassignment increased document coverage but consistently reduced cross-encoder agreement. For coarse clustering, mean top-10 three-way Jaccard decreased from 0.530 to 0.351 after reassignment, while for granular clustering it decreased from 0.613 to 0.391. This indicates a trade-off between assigning more documents to topics and retaining high-confidence agreement across different encoder representations.

Seven granular/native groups satisfied the high-confidence criteria. They represented argumentation, backdoor and adversarial attacks, temporal reasoning, graph learning, process mining, AI safety, and multi-agent reinforcement learning.

Human evaluation produced a reference dataset of 280 paper–topic judgments across the unions of the seven corresponding encoder topics. Of these, 273 were rated clearly relevant and seven unrelated, corresponding to an overall union-level coherence of **97.5%**. No papers received the intermediate score. Union-level coherence ranged from 92.3% to 100% across individual groups.

Using the same annotations for the individual-encoder comparison, macro-average coherence was 98.1% for SciNCL, 98.6% for SPECTER, and 98.0% for SPECTER2, compared with 99.0% for the strict three-encoder consensus. The mean individual-encoder coherence was 98.2%, so consensus provided a modest improvement of approximately 0.7 percentage points.

Several topic groups were also robust to changes in clustering configuration. The 21-paper backdoor/adversarial consensus core was identical under coarse/native and granular/native clustering ($J = 1.0$). The process-mining group achieved $J = 0.964$ between coarse/native and granular/native and $J = 1.0$ between granular/native and granular/reduced. The highest-ranked argumentation group achieved $J = 0.957$ between coarse/native and granular/native.

Table 2: High-confidence granular/native consensus groups. $|I|$ denotes the three-encoder intersection size, J_3 the three-way Jaccard coefficient, $R_{\min}$ minimum model retention, and Coh. manually evaluated coherence over the union of the corresponding encoder topics.

Topic	$\|I\|$	J_3	$R_{\min}$	Coh.
Argumentation	69	0.885	0.896	96.2%
Backdoor attacks	21	0.778	0.840	96.3%
Temporal reasoning	18	0.692	0.750	100%
Graph learning	17	0.654	0.773	92.3%
Process mining	27	0.628	0.711	97.7%
AI safety	14	0.538	0.636	100%
Multi-agent RL	29	0.537	0.644	100%

The backdoor/adversarial group consisted of encoder topics containing 23, 24, and 25 papers, with 21 papers in their strict intersection and 27 papers in their union, corresponding to a three-way Jaccard coefficient of **0.778**. Model retention was 0.913 for SciNCL, 0.875 for SPECTER, and 0.840 for SPECTER2. Together, these results indicate that agreement across separately generated encoder-specific topic structures can identify stable and highly coherent scientific neighbourhoods.

4.2 Full-Text Finding Extraction

The focused RQ2 case study processed 471 full-text chunks and generated 1,147 candidate result-bearing findings. All candidates were manually evaluated, yielding 510 valid findings and 637 non-findings, corresponding to an overall precision of **44.5%**.

Table 3: Extraction results by section type. Quantitative evidence percentages are calculated over valid findings within each section type.

Section type	Candidates	Valid	Non-valid	Quant. evidence
Result-oriented	314	226 (72.0%)	88 (28.0%)	55.8%
Other full text	833	284 (34.1%)	549 (65.9%)	37.3%
All	1,147	510 (44.5%)	637 (55.5%)	45.5%

Of the 1,147 extracted candidates, 314 originated from result-oriented sections and 833 from other full-text sections. Result-oriented sections contained 226 valid findings and 88 non-findings, corresponding to a substantially higher precision of **72.0%**. In comparison, 284 of the 833 candidates extracted from other full-text sections were valid, corresponding to a precision of **34.1%**.

Despite their lower precision, other full-text sections contributed 284 of the 510 validated findings, or **55.7%** of all valid findings. Result-oriented sections accounted for the remaining 226 findings (**44.3%**). Thus, restricting extraction to explicitly result-oriented sections would improve precision but would exclude more than half of the validated findings identified through full-text processing.

Quantitative-evidence analysis further showed that **55.8%** of valid findings from result-oriented sections contained explicit quantitative evidence. Importantly, **37.3%** of valid findings from other full-text sections also contained quantitative evidence. Across all 510 valid findings, 45.5% contained an explicit reported

quantity. This indicates that findings occurring outside result-oriented sections are not limited to qualitative discussion and frequently contain measurable empirical evidence relevant to downstream scientific comparison.

Taken together, these results reveal a clear precision–breadth trade-off: result-oriented sections provide a substantially denser source of valid and quantitatively supported findings, while processing the complete document recovers a larger overall set of validated findings.

5 Discussion and Conclusion

The results support a staged approach to scientific contradiction mining. For RQ1, consensus across three scientific document encoders produced highly coherent topic groups, several of which remained stable under alternative clustering configurations. Native HDBSCAN assignments consistently achieved stronger cross-encoder agreement than outlier-reassigned topics, indicating a trade-off between document coverage and topic confidence. Manual evaluation additionally produced a 280-judgment reference dataset with an overall topic coherence of 97.5%.

The RQ2 case study reveals a clear trade-off between precision, quantitative evidence, and breadth of finding recovery. Result-oriented sections were substantially more precise than other full-text sections (72.0% vs 34.1%) and contained a higher proportion of quantitatively supported findings (55.8% vs 37.3%). However, they accounted for only 44.3% of all validated findings, meaning that restricting extraction to these sections would exclude 55.7% of the findings recovered through full-text processing. Within this case study, result-oriented sections therefore provide a higher-precision and more quantitatively dense source of evidence, whereas full-text processing recovers a larger set of validated findings.

The overall extraction precision of 44.5% nevertheless shows that identifying result-bearing scientific findings from unrestricted full text remains challenging. Additional filtering and refinement are therefore required before extracted findings can be used reliably for downstream contradiction detection. The present evaluation measures precision among generated candidates but does not estimate extraction recall. Because the source papers were not exhaustively annotated for all result-bearing findings separately from the reader-generated candidates, findings that were never proposed as candidates may remain unobserved. Accordingly, the reported section-level differences describe the distribution of validated extracted findings rather than exhaustive recall over all findings present in the source papers.

A further limitation is that the manual evaluations were produced by a single annotator, so inter-annotator agreement could not be estimated. Future evaluation should therefore involve multiple independent annotators to quantify annotation reliability. Another challenge is transforming findings into atomic and self-contained claims while preserving entities, experimental conditions, comparisons, and quantitative values. Preliminary experiments showed that automatic decontextualization can alter semantic roles or omit important conditions; contradiction classification was therefore not evaluated in the present study.

The two manually annotated datasets created to address RQ1 and RQ2 are publicly available on Zenodo[1] and provide reference sets for evaluating future improvements to topic selection and scientific finding extraction. Overall, the results suggest that cross-encoder consensus can narrow a broad scientific corpus to coherent neighbourhoods, while the RQ2 case study indicates that full-text processing can recover empirical evidence that would be missed by approaches restricted to explicitly result-oriented sections.

Future work will extend the pipeline beyond the two stages evaluated here. First, finding extraction can be improved using the manually annotated RQ2 dataset to reduce false positives while preserving the additional validated findings recovered through full-text processing. Second, extracted findings should be transformed into atomic, self-contained claims through decontextualization methods that explicitly preserve experimental conditions and quantitative evidence. Third, comparable claims can be retrieved within and across the consensus topic groups before applying condition-aware contradiction classification capable of distinguishing genuine disagreement from results obtained under different settings. Finally, the full-text case study should be replicated across additional consensus groups and scientific domains, and evaluation should include multiple independent annotators to assess annotation agreement and the generalisability of the proposed pipeline.

[1] doi:10.5281/zenodo.22968108

Acknowledgements

This research was funded through the projects ELIAS (HE 101120237), DataPACT (HE 101189771), ARIS Gravitacija (GC-0001).

References

[1] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: document-level representation learning using citation-informed transformers. In *Proceedings of ACL 2020*. Association for Computational Linguistics, 2270–2282. doi:10.18653/v1/2020.acl-main.207.

[2] Maarten Grootendorst. 2022. BERTopic: neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

[3] Ashwin Kirubakaran and Henry Gagnier. 2026. BioConflict: a benchmark for evaluating large language models in biomedical contradiction detection and consensus synthesis. In *Proceedings of BioNLP 2026*. Association for Computational Linguistics, 552–558. doi:10.18653/v1/2026.bionlp-1.44.

[4] Dave Makhervaks, Plia Gillis, and Kira Radinsky. 2023. Clinical contradiction detection. In *Proceedings of EMNLP 2023*. Association for Computational Linguistics, 1248–1263. doi:10.18653/v1/2023.emnlp-main.80.

[5] Raúl Ortega and Jose Manuel Gomez-Perez. 2025. SciClaims: an end-to-end generative system for biomedical claim analysis. In *Proceedings of EMNLP 2025: System Demonstrations*. Association for Computational Linguistics, 141–154. doi:10.18653/v1/2025.emnlp-demos.11.

[6] Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of EMNLP 2022*. Association for Computational Linguistics, 11670–11688. doi:10.18653/v1/2022.emnlp-main.802.

[7] Tarek Saier, Johan Krause, and Michael Färber. 2023. unarXive 2022: all arxiv publications pre-processed for NLP, including structured full-text and citation network. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL 2023)*. IEEE, 66–70. doi:10.1109/JCDL57899.2023.00020.

[8] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: a multi-format benchmark for scientific document representations. In *Proceedings of EMNLP 2023*. Association for Computational Linguistics, 5548–5566. doi:10.18653/v1/2023.emnlp-main.338.

[9] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: verifying scientific claims. In *Proceedings of EMNLP 2020*. Association for Computational Linguistics, 7534–7550. doi:10.18653/v1/2020.emnlp-main.609.

[10] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, 4719–4734. doi:10.18653/v1/2022.findings-emnlp.347.

[11] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, 61–76. doi:10.18653/v1/2022.findings-naacl.6.

[12] Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of ACL 2022 (Volume 1: Long Papers)*. Association for Computational Linguistics, 2448–2460. doi:10.18653/v1/2022.acl-long.175.