Pushing the Frontiers of Scientific Fact-Checking: The SCINLP Dataset

Iffat Maab[1], Junichi Yamagishi[1]

[1]Yamagishi Lab, National Institute of Informatics (NII), Tokyo

EACL 2026

Introduction to Automated Fact-Checking (AFC)

AFC is the process of using computational techniques to assess the veracity of claims by retrieving relevant evidence and generating verdicts with supporting justifications.

Guo et al. (2022) - A Survey on Automated Fact-Checking

Motivation

- Can AFC (Automated Fact-Checking) be applied to complex, real-world research claims derived from the NLP?

- The scientific literature is growing rapidly (Bornmann et al., 2021).
- Rapid growth in NLP research makes it **increasingly difficult** for researchers to stay up to date.
- Manual fact-checking is slow and resource-intensive.
- **Scientific claim validation** and knowledge tracking are becoming more challenging.
- LLMs are being explored to analyze how scientific research evolves.
- **Limited work** focuses on fact-checking **scientific claims**, especially in the NLP domain.
- How about a novel scientific fact-checking dataset
  - **SCINLP**: a novel scientific fact-checking dataset tailored specifically to NLP research

Can LLMs support automated claim extraction and verification?

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. (2021) Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications, 8(1):1– 15.

Existing Scientific Fact-checking Datasets

- No fact-checking data exists for the AI/Computer science or NLP domain.

DATASET	# of Claims	Claim Origin	Evidence Source	Domain
SciFact	1,409	Researchers	Research papers	Biomedical
PubHealth	11,832	Fact-checkers	Fact-check sites	Public health
Climate-Fev	1,535	News articles	Wikipedia	Climate
HealthVer	1,855	Search queries	Research papers	Health
COVID-Fact	4,086	Reddit posts	Research, news	COVID-19
CoVERT	300	Twitter posts	Research, news	Biomedical
SciFact-Open	279	Researchers	Research papers	Biomedical
Med-Fact	150,000	Mul. choice QA	Documents	Medical
Gsci-Fact	32,200	Mul. choice QA	Documemts	Natural Sci.
FACTors	118,112	Fact-check orgs.	Multiple	Various
SciNLP(Ours)	**7,033**	**ACL anthology**	**Research papers**	**AI (NLP)**

Table: Summary of existing scientific fact-checking datasets and our proposed SCINLP

SCINLP

Using extensive prompts and taking advantage of the context window of GPT-4, extract claims with criteria:

1) The claim should be fully understandable in isolation.
2) Refuting claims must logically contradict an existing claim without relying on additional context.
3) Each claim should be of substantive interest (check-worthy), with the labeling justified by insights drawn from architectural choices, evaluation results, specific datasets, baselines, etc.

Note: LLM-based veracity prediction and temporal classification to capture paradigm shifts in NLP.

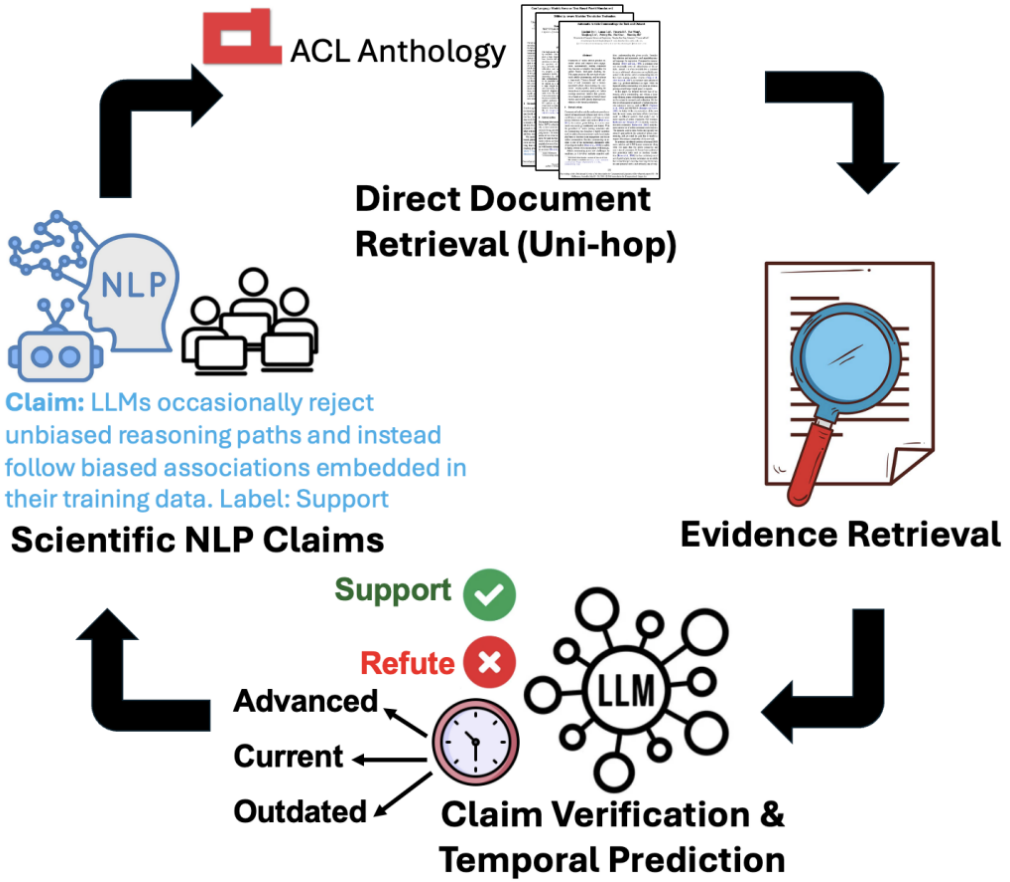

Figure: Our scientific AFC framework, supported by evidence from SCINLP from a curated collection of reputable NLP papers published between 2000 and 2024.

Highlights of Proposed Approach

- Which components of scientific article rationales serve as an effective knowledge source for NLP-based fact-checking?
- Does claim decontextualization improve performance?
- Systematically verifying the veracity of complex scientific research questions across varying rationale contexts.
- In our framework, we use multiple LLMs and diverse rationale contexts from SCINLP to examine scientific claims and research focus, complemented by **feasibility judgments for scientific reasoning in NLP.**

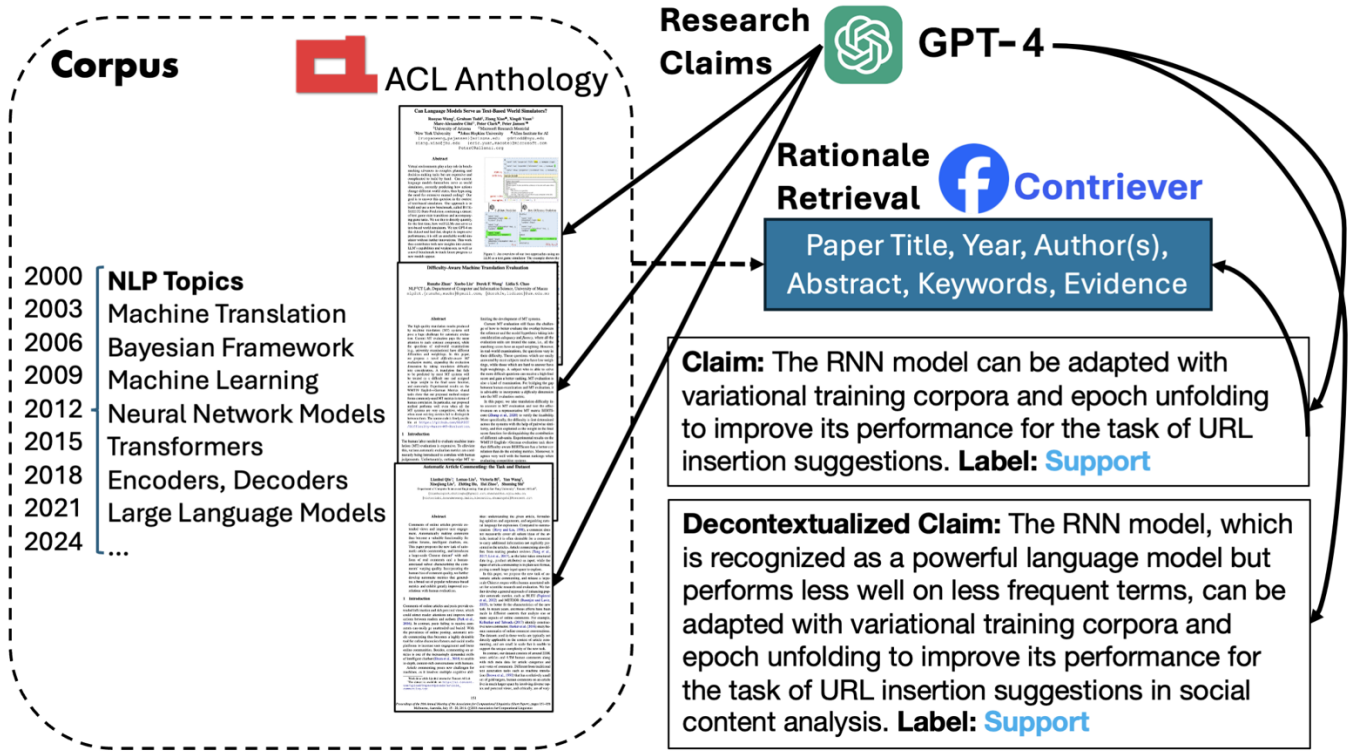

Figure: Given the claim, evidence retrieval (contextual information) and supplementary rationale information based on the source document is collected using Facebook/Contriever.

Task Formulation & Results

- Define fact-checking as a multi-component task with three objectives:
  - Step-by-Step Analysis: Perform detailed analysis over the claim and different pieces of evidence.
  - Verdict Prediction: Assign a veracity label (Support/ Refute) based on analysis.
  - Justification Generation: Provide a concise explanation that justifies the verdict.
  - Oracle setting assumes ideal knowledge of the source article, while in general setting source article is unknown.

Model	Input Setting (What is given to the model)	Oracle Setting Acc.	Oracle Setting F1	General Setting Acc.	General Setting F1
Qwen-2.5-72B					
c_{only}	Claim only	0.6074	0.5563	-	-
$c + t$	Claim + Title	0.8344	0.8116	**0.8021**	**0.7824**
$c + \mathcal{A}$	Claim + Abstract	0.8538	0.8412	0.7468	0.7399
$c + e$	Claim + Evidence	0.7006	0.6974	0.7528	0.7481
$c + t + \mathcal{A} + \mathcal{K}$	Claim + Title + Abstract + Keywords	**0.8726**	**0.8589**	0.7740	0.7636
$c + t + e + \mathcal{A} + \mathcal{K}$	Claim + Title + Evidence + Abstract + Keywords	0.8487	0.8384	0.7653	0.7596
Llama-3.3-70B					
c_{only}	Claim only	0.7387	0.6480	-	-
$c + t$	Claim + Title	0.8275	0.7961	**0.7931**	**0.7663**
$c + \mathcal{A}$	Claim + Abstract	0.8007	0.7940	0.6995	0.6972
$c + e$	Claim + Evidence	0.7644	0.7387	0.7729	0.7646
$c + t + \mathcal{A} + \mathcal{K}$	Claim + Title + Abstract + Keywords	0.8309	0.8216	0.7201	0.7154
$c + t + e + \mathcal{A} + \mathcal{K}$	Claim + Title + Evidence + Abstract + Keywords	**0.8564**	**0.8452**	0.7562	0.7506

Temporal prediction of SCINLP

Year	Claim	Time Label	Temporal Justification (Qwen-2.5-72B)
2024	*The fine-tuned Flan-T5 model generates biographies based on the given personal attributes.*	Advanced	As of the latest research and developments in NLP, the use of Flan-T5 for text generation tasks, including biography generation, is considered a cutting-edge approach. This model is among the state-of-the-art in generating human-like text based on given inputs, making the claim advanced.
2021	*MOLEMAN performs nearest neighbor inference on an index of 700 million mentions.*	Still holds true	While MOLEMAN was innovative at the time of publication for its use of a large index and contextualized mention-encoder, more recent advancements in NLP have introduced transformer-based models that offer improved performance in entity linking tasks.
2010	*The KA* algorithm makes use of outside edge items which represent the many possible derivations of a sentence.*	Outdated	While the KA* algorithm was innovative at the time of its publication, newer approaches in NLP, particularly those involving neural networks and deep learning, have since emerged and are now widely adopted for tasks such as parsing. Notable advancements include the use of transformer models for syntactic parsing, which have shown significant improvements in accuracy and speed.

Table: Sample claims across different years with corresponding temporal assessment, and temporal justifications produced by Qwen-2.5-72B model.

Temporal prediction of SCINLP

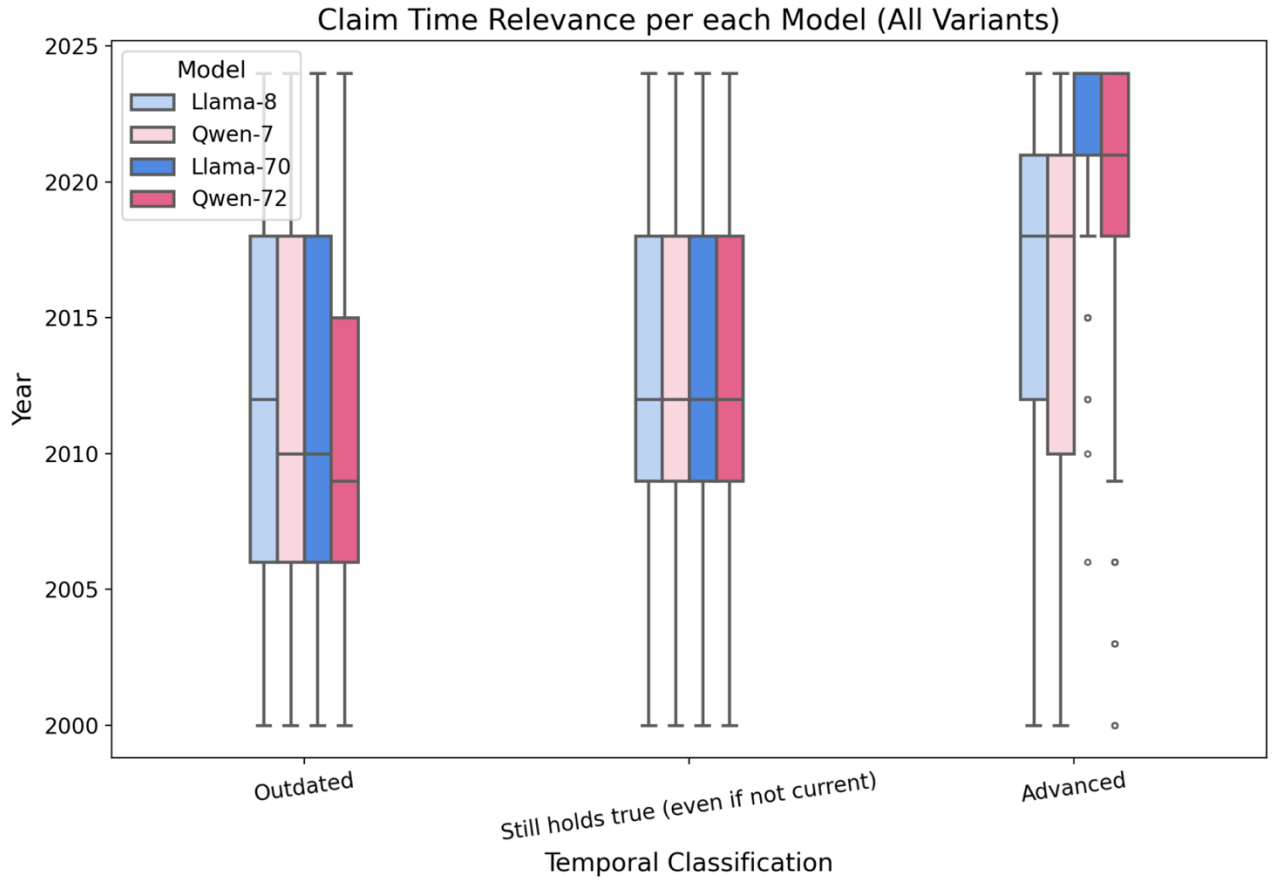

Figure: Temporal classifications (Outdated, Still holds true, Advanced) of SCINLP across Llama and Qwen model variants, showing year-wise relevance.

Conclusion

- This work introduced the first NLP-focused benchmark for systematic evaluation of scientific claim veracity.
- We use LLMs to leverage automatic scientific knowledge extraction to scale AFC.
- Our framework also supports temporal analysis and paradigm shifts across NLP research.
- We analyze model robustness, reasoning behavior, and limitations in scientific claim verification.
- We hope our work provides a foundation for automated verification of scientific claims in fast-evolving research domains.

Acknowledgements

The authors wish to express gratitude to the funding organization, as this work is supported by the JST CREST Grant (JPMJCR20D3), Japan, and the TSUBAME 4.0 supercomputer at the Institute of Science, Tokyo, whose computational resources are gratefully acknowledged.

Thank you very much!

Appendix

Claim Decontextualization

CLAIM	CLAIM DECONTEXTUALIZATION	L	EVIDENCE
Language models may have 'degraded' due to the diachronic changes in language over the past few years.	**Decontexualized Claim:** Pre-trained language models may have 'degraded' due to the diachronic changes in natural languages, particularly within the context of social media, over the past few years. **Subject:** Diachronic degradation of pre-trained language models in the context of social media. **Disambiguation Criteria:** Language Evolution and Model Performance Context.	S	Why predictive models trained on an older sample of language, may fail to work on contemporary language. How can these differences be used to track the changes in the affect around a particular topic? We first establish the diachronic validity of language-based models through predictive evaluations.
The pseudo-error sentences and domain adaptation technique are not effective in resolving the problem of collecting paired sentences.	**Decontexualized Claim:** The pseudo-error sentences and domain adaptation technique, used for grammar error correction for Japanese particles, are not effective in resolving the problem of collecting large error corpora. **Subject:** The use of pseudo-error sentences and domain adaptation technique in grammar error correction for Japanese particles. **Disambiguation Criteria:** Technique or Approach	R	Namely, when particles appear in the correct sentence, they are replaced by incorrect ones in a probabilistic manner by applying the phrase table (which stores the error patterns) in the opposite direction. The link features are important for the error correction task because the system has to judge output correctness. Example of Phrase Table (partial) approach, the conversion approach can correct multiple errors of all types in a sentence.

Table 2: Examples from the SciNLP dataset with scientific claims, their decontextualized forms, labels (**L**), and supporting evidence from FB/Contriever. Labels: S = *Support*, R = *Refute* (see Table 7) for more examples.