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

Iffat Maab, **Junichi Yamagishi**

山岸研究室 Yamagishi Lab

NII Inter-University Research Institute Corporation Research Organization of Information and Systems National Institute of Informatics

Motivation

- Scientific literature is expanding rapidly, and the fast pace of NLP research makes it increasingly difficult for researchers to stay up to date.
- Scientific claim validation and knowledge tracking are becoming more challenging.
- Automated Fact-checking (AFC) is the process of using computational techniques to assess the veracity of claims by retrieving relevant evidence and generating verdicts with supporting justifications.
- Limited work focuses on fact-checking scientific claims, especially in the NLP domain.
- Can LLMs handle the creative and nuanced task of extracting core research questions and scientific claims in the fast-evolving field of NLP? How about SCINLP?

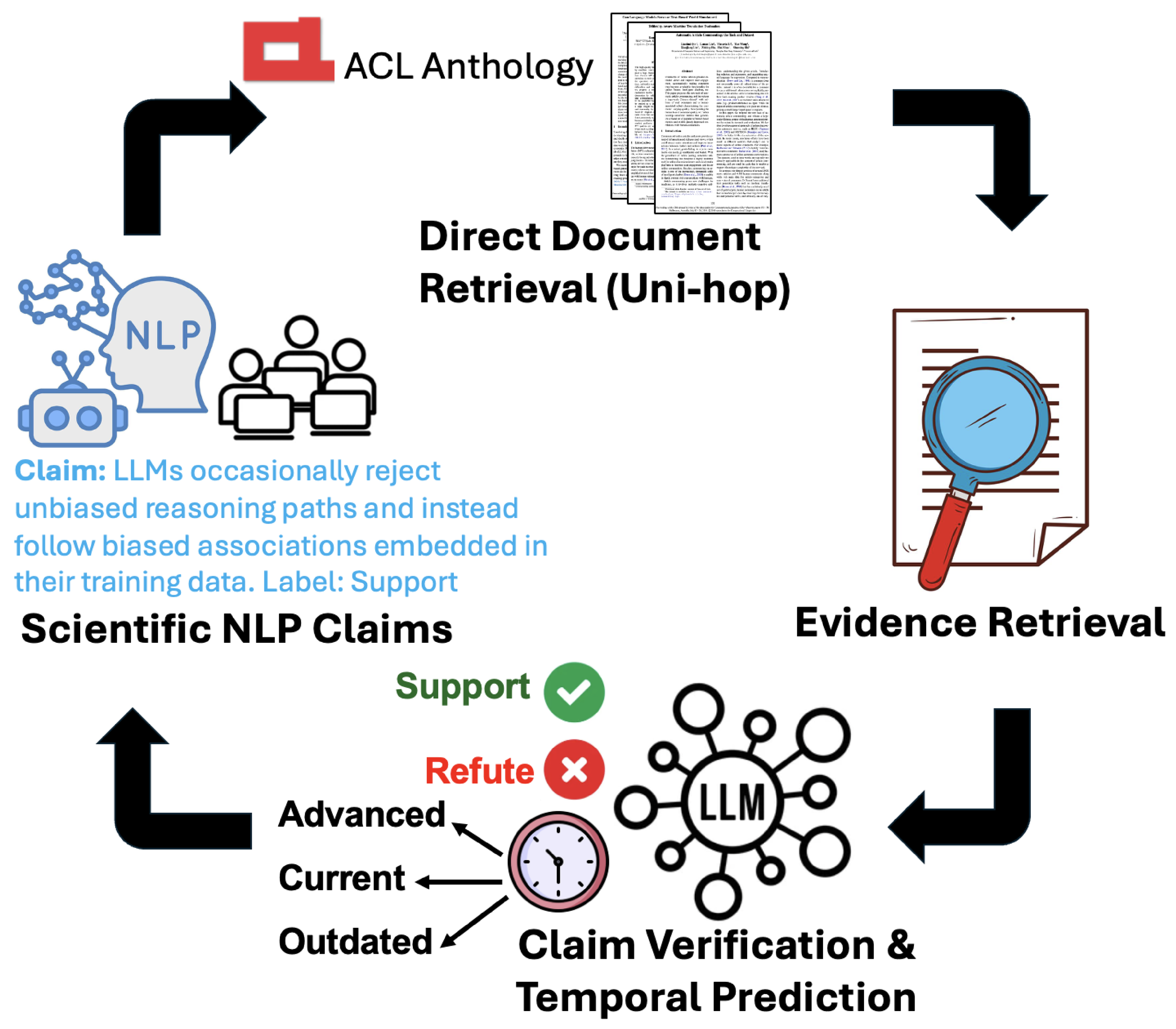

Figure: Our scientific AFC framework for SciNLP.

Exisitng Scientific Fact-Checking Datasets

DATASET	# of Claims	Evidence Source	Domain
SciFact	1,409	Research papers	Biomedical
Climate-Fev	1,535	Wikipedia	Climate
HealthVer	1,855	Research papers	Health
COVID-Fact	4,086	Research, news	COVID-19
SciFact-Open	279	Research papers	Biomedical
Med-Fact	150,000	Documents	Medical
Gsci-Fact	32,200	Documemts	Natural Sci.
SciNLP(Ours)	**7,033**	**Research papers (ACL)**	**AI (NLP)**

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

No fact-checking data exists for the AI or NLP-based domain.

Research Question

- Can AFC be applied to complex, real-world research claims derived from the NLP?
- Which components of scientific article rationales serve as an effective knowledge source for NLP-based fact-checking? Does claim decontextualization improve performance?

SciNLP Corpus Construction

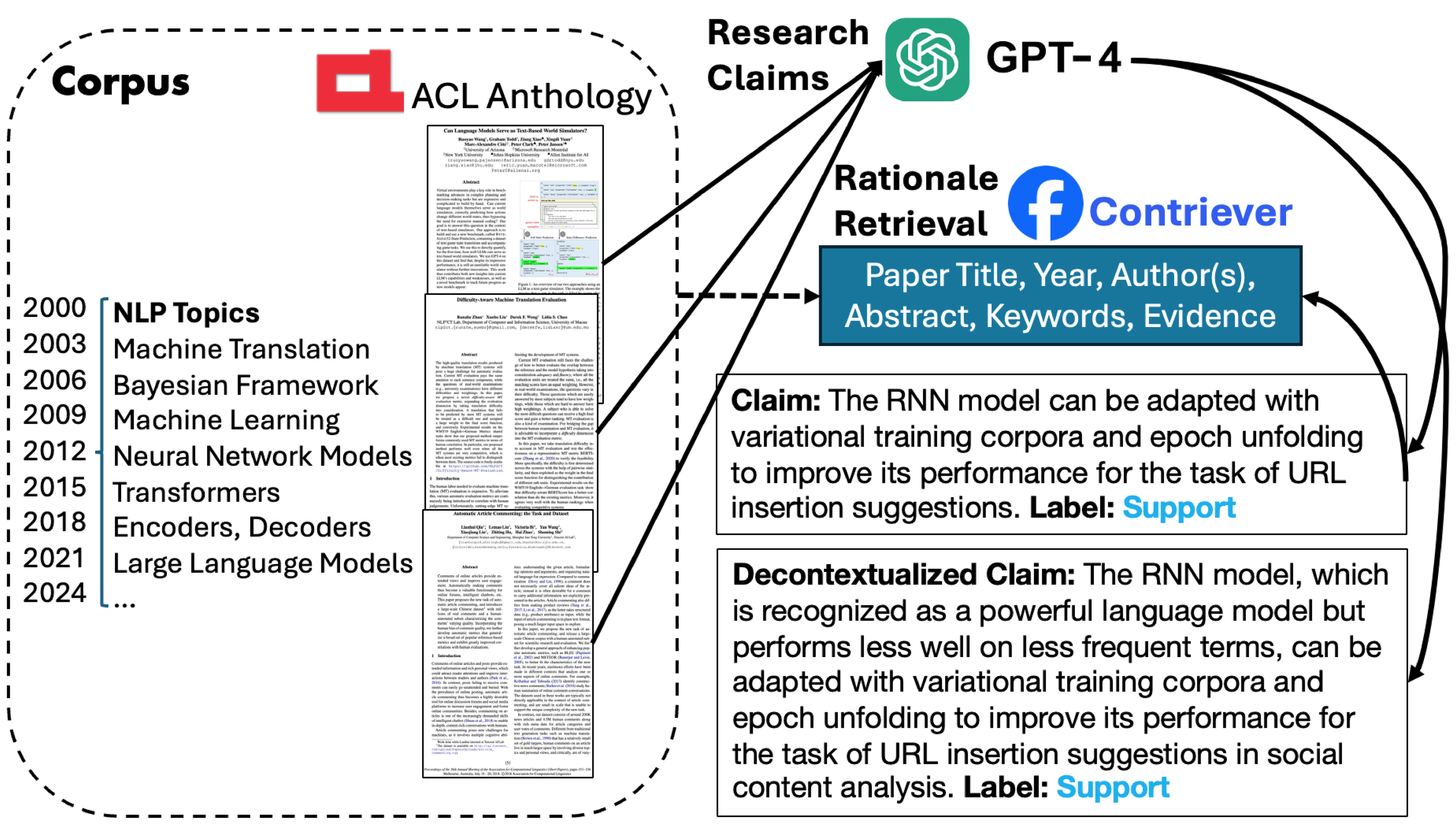

Figure: Research claims are identified for each document, and a claim is generated. Evidence retrieval and supplementary rationale information based on the source document is collected using Facebook/Contriever model.

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

- The claim should be fully understandable in isolation.
- Refuting claims must logically contradict an existing claim without relying on additional context.
- 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.

Acknowledgement: This work is supported by JST CREST Grant (JPMJCR20D3), Japan, and computational resources by TSUBAME 4.0 supercomputer at the Institute of Science, Tokyo.

Task Formulation & Results

- **Step-by-Step Analysis:** Perform detailed analysis over the claim and different pieces of evidence.
- **Verdict Prediction:** Assign a veracity label (Support/ Refute).

Oracle setting assumes ideal knowledge of the source article, while in **general setting** source article is unknown.

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

Table: Veracity prediction results on SciNLP for oracle versus general settings.

Temporal Prediction

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.

Conclusion

- This work aims to advance scientific AFC in NLP, a task of growing importance for researchers and digital media.
- We introduce SCINLP, the first NLP-based resource for systematically evaluating the veracity of scientific claims.
- Since human annotation is costly and obtaining new scientific fact-checking claims is difficult, our approach uses automatic scientific knowledge extraction to scale AFC to the rapidly expanding scientific literature.

maab@nii.ac.jp