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

Iffat Maab[1], **Junichi Yamagishi**[1]
[1]National Institute of Informatics, Tokyo, Japan
{maab,jyamagis}@nii.ac.jp

Abstract

Large Language Models (LLMs) are increasingly being used to understand how scientific research evolves, drawing growing interest from the research community. However, limited work has explored the scientific fact-checking of research questions and claims from manuscripts, particularly within the NLP domain, an emerging direction for advancing scientific integrity and knowledge validation. In this work, we propose a novel scientific fact-checking dataset SciNLP tailored to the NLP domain. Our proposed framework on SciNLP systematically verifies the veracity of complex scientific research questions across varying rationale contexts, while also assessing their temporal positioning. SciNLP includes supporting and refuting research questions from a curated collection of influential and reputable NLP papers published between 2000 and 2024. In our framework, we use multiple LLMs and diverse rationale contexts from our dataset to examine scientific claims and research focus, complemented by feasibility judgments for deeper insight into scientific reasoning in NLP.

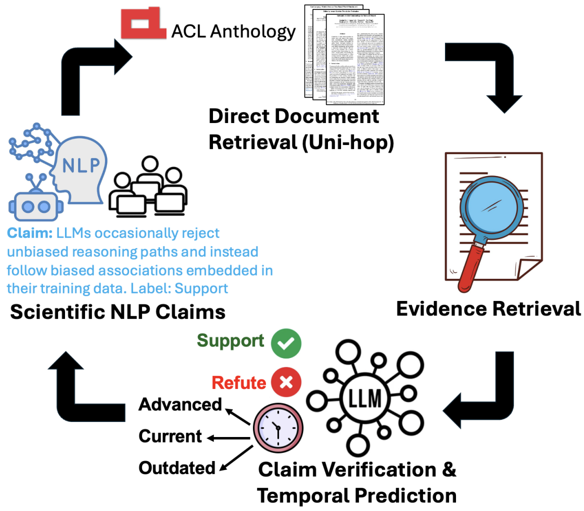

Figure 1: Our scientific AFC framework with evidence retrieved via a uni-hop approach from SciNLP and applies LLM-based veracity prediction and temporal classification to capture paradigm shifts in NLP.

1 Introduction

The scientific literature is growing rapidly (Bornmann et al., 2021; Fire and Guestrin, 2019). In particular, NLP has advanced from rule-based and statistical methods to models that capture compositional semantics and support the development of next-generation narrative-driven technologies (Jonker et al., 2024; Cambria and White, 2014). This rapid growth not only broadens the horizons of NLP research but also burdens researchers who must continuously review an ever-growing body of publications to remain current in their fields. This makes scientific claim validation and knowledge tracking increasingly challenging, highlighting the need for novel methods to process and assimilate scientific knowledge (Rogers and Augenstein, 2020; Kang et al., 2018).

Consequently, the rapid progress of large language models (LLMs) in areas such as knowledge representation and reasoning has created new opportunities for scientific applications. Recent work shows that LLMs can generate complete scientific papers (Yamada et al., 2025; Lu et al., 2024), assist research (Lehr et al., 2024), uncover patterns in extensive text corpora (Lam et al., 2024), retrieve literature (Press et al., 2024; Ajith et al., 2024), and solve complex mathematical problems and construct proofs (Trinh et al., 2024; Collins et al., 2024). Although these applications accelerate scientific discovery, an open question remains: Can LLMs handle the creative and nuanced task of extracting core research questions and scientific claims in the fast-evolving field of NLP?

Verifying scientific claims is far more challenging than political or factoid claims, which are plentiful online and easily verified by the crowd (Wadden et al., 2022). Prior work has explored the auto-

Findings of the Association for Computational Linguistics: EACL 2026, pages 1590–1617

matic generation of NLP research ideas (Si et al., 2024) and the autonomous creation of complete manuscripts (Yamada et al., 2025; Lu et al., 2024), but none has addressed fact-checking of NLP-based claims. Building on this line of research, we differ in that we extract real-world claims from NLP papers rather than generating synthetic ones. Our workflow is shown in Figure 1. Scientific claims require expert knowledge, and finding direct evidence on the web or through external sources is often challenging (Pan et al., 2023a). For example, consider the claim: *"The RNN model for CCG supertagging uses continuous vector representations for features."* This makes verifying complex claims more challenging; therefore, in this work, we evaluate pre-trained knowledge of LLMs together with rationales derived from the articles themselves to verify NLP-based claims. Specifically, we extract research questions, retrieve supporting evidence from relevant paragraphs using a unihop and multi-document retrieval strategy within the dataset, and predict their veracity. We present SCINLP, a real-world scientific fact-checking dataset sourced from papers published between 2000 and 2024 in the ACL Anthology[1], a highly recognized resource in the NLP community. We extract the core NLP research claims using a proprietary GPT-4 model. We focus on a central question: Are current LLMs capable of performing automatic fact-checking (AFC) for complex NLP-related claims? Since veracity prediction is a fundamental aspect of factuality, it serves as a critical benchmark for the broader goal of scientific AFC (Vladika and Matthes, 2023), yet building such datasets is inherently challenging due to the cost and time required to recruit expert annotators (Deng et al., 2025).

SCINLP comprises 7,033 claims with substantial representation from areas such as machine learning, machine translation, neural network models, transformer-based encoder-decoder architectures, and LLMs. Our framework goes beyond the evidence retrieved by evaluating additional rationale contexts—title, author (s), year, abstract, evidence, and keywords—to assess which best support veracity prediction of complex NLP claims. We test three open-source LLM families (8B, 70B) in six rationale distributions, resulting in 24 settings to systematically investigate rationale integration. Our method also incorporates LLMs' pre-trained knowledge to generate justifications, complementing paper evidence for a more comprehensive evaluation. Our findings also highlight the importance of decontextualization of claims (Gunjal and Durrett, 2024), which consistently improves performance across models and configurations. We evaluate 3,615 decontextualized claims, with and without rationale integration between models of different sizes.

To this end, we design a carefully controlled AFC framework that unifies claim verification components and evaluate it under various settings, such as claims-only that mimic real-world scenarios, to show how AFC systems perform in these conditions. We also introduce a novel task of analyzing temporal relevance of scientific claims by categorizing them as Advanced, Outdated, or Still Holds True, based on their publication years. This allows us to investigate whether LLMs can reflect paradigm shifts in NLP methodologies over time. Although ground-truth annotations for these temporal categories are not available, our analysis provides initial insights into how models handle diachronic changes in scientific claims. Unlike prior approaches that often depend on fine-tuning or separate stages for classification and explanation (Eldifrawi et al., 2024; Wang and Shu, 2023; Pan et al., 2023b), our framework uses LLMs in a zero-shot setup without model fine-tuning for claim analysis, verification, and justification as natural language explanations. Through extensive experimentation and reference-free evaluation, we provide nuanced insights into how rationale, model size, metadata, and decontextualization influence system performance. The two specific research questions are:

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

2 Related Work

In recent years, advances in AI have sparked considerable interest in automating scientific research, with several studies demonstrating significant potential to transform scientific innovation (Cornelio et al., 2023; Lu et al., 2024; Wang et al., 2023; Xu et al., 2021; Kitano, 2021). PeerQA (Baumgärtner et al., 2025) is a scientific QA dataset derived from ARR 2022 peer review questions, designed to sup-

[1] https://aclanthology.org/venues/acl/

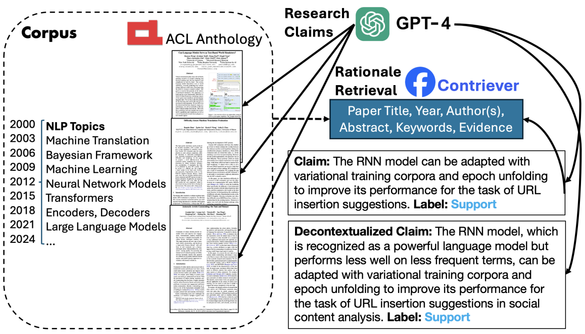

Figure 2: SCINLP corpus construction. Research claims are identified for each document, and a claim is generated by GPT-4. Evidence retrieval and supplementary rationale information based on the source document is collected using Facebook/Contriever model.

port evidence retrieval and answer generation in long-context scientific documents. Si et al. (2024) evaluated the ability of LLMs to autonomously generate research ideas by directly comparing LLM-generated and human-expert proposals. Recent advances in autonomous scientific discovery have produced agent-based systems that generate hypotheses, conduct experiments, and even write complete peer-reviewed articles (Yamada et al., 2025).

Natural Language Inference (NLI) is a well-established task of NLP to infer whether a given premise logically supports, contradicts, or is neutral to a particular hypothesis (Solanki et al., 2024). This task relates to AFC, where verification of the claim depends on available evidence. For scientific studies, prior studies have explored NLI in specialized domains, such as computational linguistic claims in SciNLI (Sadat and Caragea, 2022), medical claims in MedNLI (Romanov and Shivade, 2018), and clinical claims from breast cancer reports in NLI4CT (Jullien et al., 2023).

Scientific AFC has been extensively explored in the medical domain. Wright et al. (2022) proposed supervised and unsupervised methods to generate atomic biomedical scientific claims. In SCIFACT, claims are taken from citations of scientific papers (Wadden et al., 2020), while this approach parallels UKP Snopes, a political fact-checking dataset with claims obtained from fact-checking websites (Hanselowski et al., 2019). SciFact use only abstracts as evidence, whereas our SCINLP dataset incorporates full articles with richer contextual information. Table 1, adapted and expanded from (Vladika and Matthes, 2023), summarizes the main scientific AFC datasets, highlighting that

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 1: Summary of existing scientific fact-checking datasets, including SCIFACT (Wadden et al., 2020), PUBHEALTH (Kotonya and Toni, 2020), CLIMATE-FEVER (Diggelmann et al., 2020), HEALTHVER (Sarrouti et al., 2021), COVID-FACT (Saakyan et al., 2021), COVERT (Mohr et al., 2022), SCIFACT-OPEN (Wadden et al., 2022), MED-FACT (Tan et al., 2023), GSCI-FACT (Tan et al., 2023), FACTORS (Altuncu et al., 2025), and our proposed SCINLP.

while there are several resources - most of which are focused on the medical domain - ours is the only dataset dedicated to the domain of NLP.

3 SCINLP Dataset

For the purpose of creating a useful resource for scientific AFC in NLP, we constructed a structured dataset of research questions sourced from the ACL Anthology, a highly regarded repository of computational linguistics known for its rigorous peer review process. We use GPT-4 to extract research claims from each article according to the following criteria: (1) the claim should be fully understandable in isolation; (2) refuting claims must logically contradict an existing claim without relying on additional context or information, recognizing that support claims are generally easier to extract; and (3) each claim should be of substantive interest, with the labeling justified by insights drawn from architectural choices, evaluation results, specific datasets, comparisons with baselines, or other methodological considerations.

3.1 Corpus Construction

In SCINLP, a scientific claim is a statement that articulates a finding, hypothesis, technique, algorithm, data source, theory, software tool, empirical evaluation, or evaluation metric that can be verified using a single source as shown in Figure 2. SCINLP comprises documents sampled at three-year intervals between 2000 and 2024, ensuring a comprehensive temporal coverage of advances in the NLP field. Figure 4 shows the topical coverage, discussed in Appendix B. To ensure that our dataset comprises only high-quality documents, we restrict

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.

our analysis to top-tier ACL conference articles, a publicly available corpus with short papers. The ACL short papers are limited to five pages, ensuring that the full content of each article fits within the context window of GPT-4. This allows complete claim extraction without truncation or loss of contextual information. Papers whose content exceeded the model's context window were excluded to avoid generating claims from incomplete evidence. Following this filtering process, a total of 597 articles met the context-length constraint and were fully processed for claim generation. More information on the year and resources can be found in the Appendix A.

We use GPT-4's[2] 8,192-token context window—which also allows up to 8,192 output tokens—to process up to five standard pages of each ACL short paper. We retrieved claims from a total of 597 articles, with the per-article counts shown in Figure 5 in the Appendix. For papers in which the content was too dense to fit within this window, we excluded those articles from our dataset. For years such as 2000 and 2006, where ACL short papers were not available, we processed only the first five pages of each ACL paper to extract research claims. To generate claims that refute an existing statement or where an article refutes a research claim, we instruct GPT-4 to produce logically contradictory versions of the original claims. In our prompt, we specifically caution against introducing bias by avoiding explicit negation keywords such as "not" (Gururangan et al., 2018; Schuster et al., 2019). Later in section 6, we show that the "claim-only" verification exhibits low accuracy, which suggests that our approach to generating refuted claims did not result in significant artifacts. Details on data distribution are shown in Table 6, while on data ontology and prompt remain in Figure 5 and Figure 8 in the Appendix. A subset of the dataset will be made available to the non-profit research community upon request, and the code to construct and evaluate the dataset is publicly available at the link[3].

3.2 Decontextualization

We also decontextualize the SciNLP claims to ensure that each statement is self-contained and fully comprehensible in isolation, enabling its independent use while preserving its original meaning (see examples in Table 2 & Table 7). In our study, we use decontextualization to make our claims not as atomic facts, but rather as statements that incorporate discrete conceptual knowledge. This approach involves minimally augmenting claims with just enough additional information to enable an accurate contextual interpretation. Although prior studies have applied decontextualization primarily for evidence retrieval in fact-checking systems (Gunjal and Durrett, 2024; Zeng and Gao, 2024), our approach extends this process to ensure that the extracted claims themselves are standalone. The adaptation to decontextualized claims primarily involves incorporating domain-relevant terminology (e.g., technique, dataset, methodology, approach) that commonly appears in scientific articles. In our work, this step is particularly significant, as some claims in SciNLP are ambiguous or lack sufficient context regarding the underlying techniques or methodologies used in NLP-based manuscripts.

[2] https://platform.openai.com/docs/models/gpt-4

[3] https://github.com/nii-yamagishilab/scientific-fact-checking

By decontextualizing the claims, we address these ambiguities and enhance their clarity for downstream tasks.

Our prompts are adapted from (Gunjal and Durrett, 2024), with modifications to better suit the NLP domain by including more detailed and relevant instructions. To ensure that the claims are thoroughly decontextualized, we include both the abstract and the title of the corresponding paper. To facilitate effective decontextualization in the NLP domain, we provide the model with six few-shot examples, each accompanied by the corresponding abstract and title (see Table 12 for prompt, and Figures 10 and 11 for few-shot examples in the Appendix for details). Due to the substantial size of our dataset, the length of few-shot examples, and the computational cost of using the GPT-4 model, we limited decontextualization on approximately half of the available input claims, ensuring balanced representations from each year. In total, 3,615 decontextualized claims are processed, with yearly distributions summarized in Table 6 and examples illustrating the transformation of claims along with the corresponding evidence retrieved are presented in Table 2 & Table 7. After extracting claim decontextualization, we apply ambiguity criteria to identify the main subject and disambiguation cues within each claim, further enhancing their clarity and reducing the potential for ambiguous interpretations. In addition to the decontextualized claims themselves, we use keywords extracted from the corresponding articles to determine the primary subject and the appropriate disambiguation information for each claim. Details of the ambiguity check prompt are provided in the Table 13 itself, and the corresponding two few-shot examples are shown in Figure 12 in the Appendix.

3.3 Human Analysis

Although automated metrics provide valuable information, human evaluation is useful to evaluate the utility of extracted claims. In our evaluation, we focus primarily on precision rather than recall, prioritizing the correctness and grounding of claims in the original abstracts. We made this choice to ensure that the claims are correct and well-supported, rather than attempting to identify all possible claims. To this end, we perform manual verification of the claims by asking three NLP PhD students to exclude those unrelated to specific scientific findings or are ambiguous when interpreted in isolation. Any claim excluded by at least one expert was removed from SciNLP. This initial filtering was performed without reference to supporting evidence or scientific articles. We note that of 7,341 sentences, experts could not distinguish whether 308 sentences could be classified as claims, and these sentences were therefore excluded from further analysis. Following this step, the remaining claims were manually cross-checked against the abstracts, rather than the full articles, due to the substantial time required for full text documents and previous efforts to verify claims using full-document annotation have reported low agreement between annotators (Wadden et al., 2020). We found that for SciNLP, evidence for more than 60% of claims can be found within the abstract.

4 NLP Fact-Checking System Design

Task Formulation Let c denote a scientific claim, t denote the title, $\mathcal{A}$ denote the abstract, $e = \{e_1, e_2, e_3\}$ denote the top three evidence statements, and $\mathcal{K} = \{k_1, k_2, \ldots, k_N\}$ denote the set of keywords extracted for each scientific article. For each claim, we assign a label $y(c, z_i) \in \{\text{SUPPORTS}, \text{REFUTES}\}$ with respect to the claim c. For decontextualization, let dec represent a scientific decontextualized claim, where each dec is further associated with a subject s and disambiguation criteria x, both obtained through the decontextualization process as described in Section 3.2. We concatenate each dec with its corresponding s and x to form the input to the model, which we denote as dec_{s+x}.

Evidence Retrievel (Oracle vs. General) A key step in our fact-checking pipeline is the retrieval of relevant rationale information to assess the veracity and temporal validity of scientific claims. Given the crucial role of rationales in fact verification, previous work has prioritized improving evidence retrieval methods to better inform classification decisions (Wei et al., 2022; Hanselowski et al., 2018; Mishra et al., 2024). In our study, we build on this line of research and use the Facebook Contriever (Izacard et al., 2023), a dense retrieval model to extract the corresponding rationale information from the articles. Given a claim c from the respective article, our evidence retrieval function is defined as (t, $\mathcal{A}$, e, $\mathcal{K}$) = $\texttt{Contriever}(c, article)$, where the title, abstract, and keywords of each article are extracted, while for evidence the three most relevant evidence statements are identified using sentence embeddings that semantically align most with the claim in the SciNLP dataset. We extract the top ten keywords from each article to support the scientific fact-checking task, as they capture domain-specific terminology, including the core entities, methods, and topics of the paper. We refer to this setup as an **oracle setting**, since it assumes ideal knowledge of the source article. In contrast, our second setup, termed the **general setting**, represents a more practical scenario in which the source article is unknown. Here, we apply (t, $\mathcal{A}$, e, $\mathcal{K}$) = `Contriever`(c) setup where the title, abstract, evidence, and keywords for a given claim are not known. We use this setting for rationale search within our corpus of 597 articles. Besides t, $\mathcal{A}$, e, $\mathcal{K}$, SciNLP also includes metadata (e.g., author, year), which are not used in our experiments.

Temporal Prediction We frame temporal prediction of scientific claims as an exploratory analysis with open-ended classification, not as a task with definitive ground-truth annotations. Specifically, we group claims into three heuristic categories for analysis: **Advanced**, for claims describing state-of-the-art approaches that reflect the latest advances in NLP; **Outdated**, for claims referring to methods subsequently surpassed by newer developments; and **Still holds true**, for claims that, although not representing the latest advances, remain valid and accurate despite newer alternatives. These categories are assigned post-hoc based on publication years of articles to examine whether LLM outputs capture temporal shifts in NLP research (see Table 9 for details).

5 Model Inference and Evaluation

We evaluate the verification of the claims in SciNLP across six levels of granularity. For claim-only (c_{only}), we assess the model's ability to support or refute a claim based solely on the claim text. For claim-title ($c + t$), we measure performance using both claim and title, hypothesizing that LLMs may leverage their pretraining to identify relevant papers and verify the validity of the claim. For claim-abstract ($c + \mathcal{A}$), we evaluate performance using the claim and abstract, as 60% of the claims have supporting rationales present within the abstract. In the claim-evidence ($c + e$), we assess LLMs ability to verify claims when explicit evidence is provided, making it a more direct fact-checking task. Furthermore, we examine performance on the concatenated combination of claim, title, abstract, and keywords ($c + t + \mathcal{A} + \mathcal{K}$) to determine whether access to additional rationale context helps LLMs to accurately verify claims. Finally, in the most comprehensive setting, we concatenate the evidence with the previous task as ($c + t + e + \mathcal{A} + \mathcal{K}$), thus providing the model with all available rationale information for a given claim. Each of these experimental variations is performed on both the original dataset and the set of decontextualized claims (dec_{s+x}). Overall performance is measured using accuracy and the macro-F1 score. We focus on macro F1 over micro F1, as micro F1 equals accuracy in binary tasks, while macro F1 better highlights the challenges in correctly predicting "refute" labels. Furthermore, as LLMs behave stochastically, evaluation is often non-trivial. To address this, we repeat each fact-checking task three times and report the majority vote to assess classification performance.

Model	Oracle Setting Acc.	Oracle Setting F1	General Setting Acc.	General Setting F1
Qwen-2.5-7B				
c_{only}	0.6858	0.5802	-	-
$c+t$	0.7901	0.7734	0.7501	0.7381
$c+\mathcal{A}$	0.7850	0.7765	0.6933	0.6914
$c+e$	0.6607	0.6603	0.7371	0.7334
$c+t+\mathcal{A}+\mathcal{K}$	**0.8147**	**0.8014**	0.7266	0.7208
$c+t+e+\mathcal{A}+\mathcal{K}$	0.7993	0.7889	**0.7595**	**0.7530**
Qwen-2.5-72B				
c_{only}	0.6074	0.5563	-	-
$c+t$	0.8344	0.8116	**0.8021**	**0.7824**
$c+\mathcal{A}$	0.8538	0.8412	0.7468	0.7399
$c+e$	0.7006	0.6974	0.7528	0.7481
$c+t+\mathcal{A}+\mathcal{K}$	**0.8726**	**0.8589**	0.7740	0.7636
$c+t+e+\mathcal{A}+\mathcal{K}$	0.8487	0.8384	0.7653	0.7596
Llama-3.3-8B				
c_{only}	0.7025	0.5572	-	-
$c+t$	0.7605	0.6887	0.7585	0.6890
$c+\mathcal{A}$	0.8043	0.7815	**0.7524**	**0.7348**
$c+e$	0.7573	0.7341	0.7971	0.7759
$c+t+\mathcal{A}+\mathcal{K}$	**0.8157**	**0.7899**	0.7749	0.7566
$c+t+e+\mathcal{A}+\mathcal{K}$	0.8075	0.7826	0.7918	0.7733
Llama-3.3-70B				
c_{only}	0.7387	0.6480	-	-
$c+t$	0.8275	0.7961	**0.7931**	**0.7663**
$c+\mathcal{A}$	0.8007	0.7940	0.6995	0.6972
$c+e$	0.7644	0.7387	0.7729	0.7646
$c+t+\mathcal{A}+\mathcal{K}$	0.8309	0.8216	0.7201	0.7154
$c+t+e+\mathcal{A}+\mathcal{K}$	**0.8564**	**0.8452**	0.7562	0.7506

Table 3: Veracity prediction results on SciNLP (Support/Refute) for oracle versus general settings.

Prompt Design Our basic setting adopts a zero-shot approach, utilizing the prompt as detailed in Figure 9 in the Appendix. The prompt is designed according to our problem formulation and the established principles of prompt engineering (Schulhoff et al., 2024). The prompt consists of three

Model	Decontext Acc.	Decontext F1	Oracle Acc.	Oracle F1
Qwen-2.5-7B				
dec_{s+x}	0.6160	0.5416	0.5959	0.5286
$dec_{s+x} + t$	0.7990	0.7910	0.7666	0.7656
$dec_{s+x} + \mathcal{A}$	0.8067	0.8035	0.7936	0.7935
$dec_{s+x} + e$	0.6865	0.6863	0.7086	0.7019
$dec_{s+x} + t + \mathcal{A} + \mathcal{K}$	**0.8338**	**0.8278**	**0.8037**	**0.8031**
$dec_{s+x} + t + e + \mathcal{A} + \mathcal{K}$	<u>0.8136</u>	<u>0.8094</u>	<u>0.7973</u>	<u>0.7972</u>
Qwen-2.5-72B				
dec_{s+x}	0.6168	0.5496	0.5645	0.5482
$dec_{s+x} + t$	0.7988	0.7936	0.7949	0.7908
$dec_{s+x} + \mathcal{A}$	0.8119	0.8053	0.8472	0.8466
$dec_{s+x} + e$	0.7667	0.7641	0.7306	0.7287
$dec_{s+x} + t + \mathcal{A} + \mathcal{K}$	<u>0.8437</u>	<u>0.8471</u>	**0.8562**	**0.8550**
$dec_{s+x} + t + e + \mathcal{A} + \mathcal{K}$	**0.8598**	**0.8557**	<u>0.8511</u>	<u>0.8509</u>
Llama-3.3-8B				
dec_{s+x}	0.5981	0.5121	0.5914	0.5045
$dec_{s+x} + t$	0.7397	0.7234	0.6832	0.6504
$dec_{s+x} + \mathcal{A}$	0.7521	0.7393	0.7770	0.7734
$dec_{s+x} + e$	0.7470	0.7335	0.7297	0.7263
$dec_{s+x} + t + \mathcal{A} + \mathcal{K}$	**0.7671**	**0.7560**	<u>0.7809</u>	<u>0.7760</u>
$dec_{s+x} + t + e + \mathcal{A} + \mathcal{K}$	<u>0.7528</u>	<u>0.7399</u>	**0.7850**	**0.7804**
Llama-3.3-70B				
dec_{s+x}	0.6978	0.6657	0.6455	0.5992
$dec_{s+x} + t$	0.7900	0.7783	0.7784	0.7697
$dec_{s+x} + \mathcal{A}$	0.8338	0.8314	0.8279	0.8278
$dec_{s+x} + e$	0.7892	0.7850	0.7517	0.7512
$dec_{s+x} + t + \mathcal{A} + \mathcal{K}$	<u>0.8428</u>	<u>0.8497</u>	<u>0.8440</u>	<u>0.8439</u>
$dec_{s+x} + t + e + \mathcal{A} + \mathcal{K}$	**0.8681**	**0.8550**	**0.8564**	**0.8563**

Table 4: Comparison of veracity prediction between decontextualized (Decontext) and oracle claims.

components: system, user, and assistant. The system message establishes the context for the model, while the user message specifies the context, guidelines, label definitions, and claim itself, along with relevant rationales when available. The assistant message contains the model's response to the input. To ensure clarity and consistency, we provide detailed definitions of veracity and temporal classification labels. For justification generation in both veracity assessment and temporal classification, we instruct the model to produce justifications that are clear, relevant, consistent, and useful. Additional details can be found in the prompt itself. See Appendix C for details on the models, hardware, and implementation specifics.

Justification Generation We evaluate ROUGE (Grusky, 2023) and BERTScore (Zhang et al., 2019) in our AFC task by comparing various input rationales provided to LLMs with the model-generated justifications across different experimental settings. This is detailed in Appendix D.

6 Results & Discussion

In our initial experiments, we examine models on 7,033 SciNLP claims under a zero-shot baseline across different rationale configurations, as shown in Table 3. Pilot experiments revealed consistent performance improvements with additional rationales as we go down the table. Building on these observations, we evaluated models in both oracle and general settings, testing various rationale distributions. Notably, we observe that including the article title ($c + t$) substantially outperforms the claim-only (c_{only}) configuration in both settings, indicating that the title of the article enhances model reasoning. In the oracle setting, except for Llama-8, we observe a slight difference in performance between the $c + t$ and ($c + \mathcal{A}$ configurations, suggesting that the paper title alone enables the model to leverage its pre-trained knowledge to associate information about well-known studies or topics.

In both oracle and general settings, the inclusion of augmented rationales ($c + t + e + \mathcal{A}$+$\mathcal{K}$) leads to performance gains. The gains are more pronounced in the oracle setting, highlighting the advantage of a richer context. Given that rationale context improves performance in LLMs, our findings align well with prior work (Sahitaj et al., 2025; Samarinas et al., 2021; Atanasova et al., 2022), indicating that models benefit from increased domain knowledge in predicting veracity, which in our case is NLP. In the oracle setting, configurations that incorporate the abstract, ($c + \mathcal{A}$), ($c + t + \mathcal{A}$+$\mathcal{K}$), and ($c + t + e + \mathcal{A}$+$\mathcal{K}$), consistently demonstrate the highest performance gains. This trend aligns with our finding that about 60% of the supporting rationales for the SciNLP claims reside in the abstract, as further confirmed by manual analysis. The smaller models (7B, 8B), as well as the larger Qwen-72B, achieve the best performance under ($c + t + \mathcal{A}$+$\mathcal{K}$), while the Llama-70B performs best with ($c + t + e + \mathcal{A}$+$\mathcal{K}$), suggesting that inclusion of both evidence and contextual cues benefits models more significantly. For example, Llama-70B attains a F1-score of 0.8452 in ($c + t + e + \mathcal{A}$+$\mathcal{K}$), improving from 0.8216 in (c+t+$\mathcal{A}$+$\mathcal{K}$) and 0.7961 in ($c + t$). For Qwen-72B, the highest F1-score of 0.8589 is achieved in ($c + t + \mathcal{A}$+$\mathcal{K}$), followed by 0.8412 in the abstract-only setup ($c + \mathcal{A}$), and 0.8384 in (c+t+e+$\mathcal{A}$+$\mathcal{K}$). It can also be seen that Qwen-72B substantially outperforms all models in ($c + t + \mathcal{A}$+$\mathcal{K}$).

In contrast, in ($c + e$), where models are provided with top-3 evidence statements retrieved by Contriever, consistently shows the lowest performance across all models in oracle settings. We attribute this to retrieval limitations, which, despite being semantically relevant evidence, do not always capture the most contextually precise supporting statements within an article. Consequently, even when direct evidence is retrieved, it can include noisy or partially relevant excerpts, thereby reducing overall performance and alignment in the oracle setting (sample examples in Table 9). Interestingly, unlike in the oracle setting, $(c + e)$ performs better in the general setting, suggesting that multi-document evidence retrieval offers richer and more diverse context, helping models capture a wider range of supporting and opposing arguments. While the oracle setting confines evidence to a single article, limiting factual scope, the general setting benefits from cross-document corroboration, allowing Contriever to aggregate semantically relevant statements from multiple sources (597 articles), enriching contextual grounding. Although this introduces noise, the broader evidence pool appears to enhance cross-perspective reasoning, improve factual alignment, and overall performance.

In the general setting, we further observe that Contriever retrieves about 55% of abstracts and 50% of titles that match the original articles. Although the overall performance of this setting declines compared to the oracle setup, the gap remains modest, suggesting that cross-document retrieval maintains a reasonable degree of reliability even without direct source access. However, both small and large LLMs exhibit comparable performance, showing no substantial gap, suggesting that in multi-document fact-checking, retrieval quality rather than model size primarily limits performance. In other words, small and large models are constrained by the same retrieval noise and evidence incompleteness, leading to a ceiling effect where additional model capacity offers diminishing returns. For example, Llama-8B achieves an F1 of 0.7733 in $(c + t + e + \mathcal{A} + \mathcal{K})$, compared to 0.7506 for Llama-70B. Furthermore, we report the ROUGE and BERTScore for LLM-generated justifications in both settings in Table 8 (Appendix D), complementing the main results of Table 3, with illustrative justification examples in Table 10.

We conduct additional experiments on **decontextualized claims** to assess how a more comprehensive knowledge in claims can influence LLM performance. These experiments follow the same settings used for the oracle and we adopt the same prompt as in our previous evaluations (Figure 9), with the only difference being that decontextualized claims are supplemented with subject and disambiguation criteria (see examples in Table 2). To ensure consistency, we compare decontextualized claims with the same subset of 3,615 oracle claims, with results summarized in Table 4. We observe that decontextualized claims outperform oracle, with all models showing clear performance gains in the first two settings (dec_x & $dec\_x + t$) compared to oracle claims (c_{only} & $c_{only} + t$). For example, Qwen-7B and 72B achieve accuracies of 61.60% and 61.68%, respectively, in the setting dec_{s+x}, compared to 59.59% and 56.45% in the setting c_{only}.

In richer contextual configurations, all models except Llama-8B show consistent performance gains across nearly all settings on decontextualized claims compared to oracle (non-decontextualized) claims. This suggests that rationales retrieved via decontextualized claims are more effective than those derived from oracle claims, particularly in $(dec_{s+x} + e)$ configuration compared to $(c+e)$. We attribute this improvement to the fact that decontextualized claims provide clearer, more precise, and self-contained information, which enhances the retrieval of contextually relevant evidence. In configurations such as $(dec_x + t + \mathcal{A} + \mathcal{K})$, larger models such as Qwen-72B and Llama-70B achieve the highest performance on decontextualized claims, outperforming their regular counterparts. Interestingly, Llama-8B exhibits a different trend, showing limited sensitivity to the added precision of decontextualized inputs — possibly due to its smaller size and reduced ability to leverage fine-grained contextual cues. Overall, our findings highlight the effectiveness of the zero-shot setting and suggest that decontextualization improves veracity prediction for NLP-based claims when evaluated in isolation (dec_{s+x}), i.e., in the absence of detailed rationales or evidence. The small gap between accuracy and macro-F1 indicates balanced performance across the "support" and "refute" classes, indicating no substantial class bias.

6.1 Evaluation with Gemma Models

We further conduct controlled experiments with the Gemma-3 model family (4B and 27B) under General, General-Subset, and Decontext settings to analyze the effect of decontextualization on veracity prediction, as shown in Table 5. **General-Subset** uses the same claims as the Decontext setting but without decontextualization. Because only a sub-

Model	Input	General Acc.	General F1	General-Subset Acc.	General-Subset F1	Input	Decontext Acc.	Decontext F1
Gemma-3-4B	c_{only}	**0.7214**	**0.6730**	0.6731	0.6554	$dec\_s\_+\_x$	0.5335	0.4327
	$c+t+e+\mathcal{A}+\mathcal{K}$	**0.7678**	**0.7201**	0.7174	0.7008	$dec\_s\_+\_x+t+e+\mathcal{A}+\mathcal{K}$	0.5697	0.4542
Gemma-3-27B	c_{only}	0.7632	0.7215	0.7119	0.6995	$dec\_s\_+\_x$	**0.7406**	**0.7263**
	$c+t+e+\mathcal{A}+\mathcal{K}$	0.7216	0.7188	0.7686	0.7667	$dec\_s\_+\_x+t+e+\mathcal{A}+\mathcal{K}$	**0.7898**	**0.7878**

Table 5: Comparison of veracity prediction (Support/Refute) for Gemma models across General, General-Subset, and decontextualized (Decontext) settings.

set of claims could be decontextualized due to resource constraints, this setting enables a controlled comparison on an aligned claim subset. Overall, removing contextual grounding leads to a clear degradation in performance, with the magnitude of the drop depending on the size of the model size and the richness of the input. These experiments are performed using

For Gemma-4B, performance declines substantially under decontextualization. In the General-Subset setting, the (c_{only}) input achieves an accuracy of 0.6731, which drops sharply to 0.5335 in the Decontext (dec_{s+x}) setting. Augmenting the input with title, evidence, abstract, and keywords i.e, ($t+e+\mathcal{A}+\mathcal{K}$) improves performance across all settings; however, even with enriched inputs, decontextualized performance with Gemma-4B (0.5697 accuracy) remains well below the General-Subset (0.7008 accuracy) configuration. This indicates the possibility of reduced model size dependency and increased hallucination to leverage fine-grained contextual cues.

In contrast, Gemma-27B exhibits notably greater robustness, with decontextualized performance remaining consistently high across each input configuration. For example, under claim-only (c_{only}) inputs, accuracy in the decontextualized (dec_{s+x}) setting slightly improves compared to the General-Subset configuration, increasing from 0.7119 to 0.7406 while maintaining a comparable F1 score (0.6995 vs. 0.7263). This trend becomes more pronounced when additional contextual information is provided: the enriched decontextualized input ($dec_{s+x}+t+e+\mathcal{A}+\mathcal{K}$) setting achieves the highest overall performance for Gemma-27B with an accuracy of 0.7898 and F1-score of 0.7878, surpassing both the General and General-Subset settings.

The results with Gemma suggest that, for larger models, decontextualization can act as a regularizing condition rather than a limitation, encouraging the model to rely on semantically salient evidence rather than document-specific structure. Importantly, the setting (dec_{s+x}) reflects a real-world fact-checking scenario, where claims are often evaluated in isolation or with limited contextual grounding. The observed performance gains indicate that Gemma-27B can effectively generalize under such conditions, highlighting its suitability for scientific fact-checking tasks compared to the larger models Qwen-72B and Llama-70B.

Furthermore, the temporal analysis of scientific claims is presented in Appendix E, and the year-wise veracity prediction performance across models is provided in Appendix E.1 (Figure 6).

7 Conclusion

This paper aims to advance scientific AFC in NLP, a task of growing importance for researchers, practitioners, policy makers, and digital media. We introduce SciNLP, the first NLP-based resource for systematically evaluating the veracity of scientific claims. As scientific publications continue to grow exponentially, claim verification remains challenging due to complex scientific language and limited expert annotations. Our study addresses these issues by comprehensively evaluating LLMs to verify foundational and emerging claims of NLP research, an area that has received limited attention. We demonstrate that incrementally adding rationale contexts from articles improves verification of complex scientific claims. Decontextualized claims significantly improve performance for small- and large-scale LLMs in claim-only settings, indicating that incorporating discrete conceptual claim knowledge enhances scientific reasoning. We also analyze the temporal relevance of claims, highlighting LLMs potential to capture paradigm shifts in NLP. Although our findings confirm the applicability of LLMs for automated claim detection and verification, developing more robust evidence retrieval methods remains a challenge for future work.

Limitations

Although this study provides valuable insights on the fact-checking of NLP-focused scientific claims, some limitations should be considered. First, our approach utilizes a uni-hop and multi-document retrieval strategy within the dataset, directly extracting evidence from the source documents. This decision was motivated by our emphasis on scientific claim verification rather than on achieving exhaustive evidence retrieval. Future work could explore more advanced multi-hop or cross-source retrieval techniques to capture supporting rationales from a wider range of documents, thus expanding the contextual grounding of claim verification within SciNLP. Evidence retrieval for NLP-based scientific claims remains particularly challenging, as relevant information is often distributed across multiple sections or papers, and precise matching requires deep domain understanding. For preliminary experiments, we explored the search for evidence from multiple sources using Serper; however, performance gains were limited. We plan to extend this line of investigation in future work. In addition, the field lacks well-annotated fact-checking datasets specifically designed for NLP research. In our study, annotations were generated using GPT-4, making SciNLP the first dataset to include research questions spanning a wide temporal range (2000–2024). Due to budget limitations, decontextualization currently covers roughly half of the dataset, corresponding to large-scale evaluations conducted with proprietary models. Nonetheless, our choice to evaluate open-source models promotes transparency and reproducibility, supporting continued community-driven progress in automated scientific fact-checking.

Moreover, since SciNLP is constructed exclusively from English NLP articles, the findings may not generalize to other languages or scientific domains. Future work could extend the dataset to multilingual and cross-domain settings to assess broader applicability. Despite an extensive review of the literature, some relevant studies may still have been overlooked due to variations in terminology or keywords.

Ethical Considerations

All data in SciNLP are derived from the publicly available ACL Anthology, with strict adherence to the relevant copyright guidelines. Each record explicitly includes the source PDF, title, and authors. The dataset has also been carefully reviewed to ensure ethical compliance, with no privacy risks identified during data collection, processing, or model evaluation. Human evaluations were conducted voluntarily. However, we acknowledge that automated fact-checking models may inherit or amplify existing biases in NLP research, especially when scaled. Therefore, we urge caution in applying this framework in real-world production environments and encourage further investigation into its broader societal and ethical implications.

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.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search. *arXiv preprint arXiv:2407.18940*.

Enes Altuncu, Can Baskent, Sanjay Bhattacherjee, Shujun Li, and Dwaipayan Roy. 2025. Factors: A new dataset for studying the fact-checking ecosystem. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3530–3539.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. 2025. Peerqa: A scientific question answering dataset from peer reviews. *arXiv preprint arXiv:2502.13668*.

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.

Erik Cambria and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, and 1 others. 2024. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121.

Cristina Cornelio, Sanjeeb Dash, Vernon Austel, Tyler R Josephson, Joao Goncalves, Kenneth L Clarkson, Nimrod Megiddo, Bachir El Khadir, and Lior Horesh. 2023. Combining data and theory for derivable scientific discovery with ai-descartes. *Nature Communications*, 14(1):1777.

Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. The next phase of scientific fact-checking: advanced evidence retrieval from complex structured academic papers. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 436–448.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated Justification Production for Claim Veracity in Fact Checking: A Survey on Architectures and Approaches. *Preprint*, arXiv:2407.12853.

Hugging Face. 2021. The ai community building the future. *URL: https://huggingface. co*.

Michael Fire and Carlos Guestrin. 2019. Over-optimization of academic publishing metrics: observing goodhart's law in action. *GigaScience*, 8(6):giz053.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Max Grusky. 2023. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Richard AA Jonker, Tiago Almeida, and Sérgio Matos. 2024. Analyzing a decade of evolution: Trends in natural language processing. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 162–176. Springer.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Hiroaki Kitano. 2021. Nobel turing challenge: creating the engine for scientific discovery. *NPJ systems biology and applications*, 7(1):29.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Steven A Lehr, Aylin Caliskan, Suneragiri Liyanage, and Mahzarin R Banaji. 2024. Chatgpt as research scientist: Probing gpt's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, 121(35):e2404328121.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023a. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023b. Fact-Checking Complex Claims with Program-Guided Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. 2024. Citeme: Can language models accurately cite scientific claims? *Advances in Neural Information Processing Systems*, 37:7847–7877.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in nlp? *arXiv preprint arXiv:2010.03863*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Premtim Sahitaj, Iffat Maab, Junichi Yamagishi, Jawan Kolanowski, Sebastian Möller, and Vera Schmitt. 2025. Towards automated fact-checking of real-world claims: Exploring task formulation and assessment with llms. *arXiv preprint arXiv:2502.08909*.

Chris Samarinas, Wynne Hsu, and Mong-Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 3499–3512.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, and 1 others. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification 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)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Dhara Solanki, Amit Thakkar, Kush Patel, Jigar Sarda, and Akash Kumar Bhoi. 2024. A review on approaches and applications of natural language inference. In *International Conference on Data Analytics & Management*, pages 441–454. Springer.

Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 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*, pages 2652–2664, Dubrovnik, Croatia. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2210.13777*.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Haoran Wang and Kai Shu. 2023. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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 the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, and 1 others. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4).

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A SciNLP Corpus Source

The claims of the corpus are generated using a proprietary GPT-4 model, which is instructed to identify and extract research questions from academic papers. SciNLP is constructed from the following years and resources: 2024[4], 2021[5], 2018[6], 2015[7], 2012[8], 2009[9], 2006[10], 2003[11], and 2000[12].

B SciNLP Topical Coverage

An analysis of the most frequent topics in the SciNLP dataset, as shown in Figure 4, shows clear temporal trends in the focus of scientific claims over the years. Claims from earlier periods (2000, 2003, 2006) predominantly center around topics such as machine translation, argument generation, inverse document frequency, vector-space models, speech recognition, Bayesian frameworks, and WorkNet. These subjects, while foundational, are now considered more traditional or even outdated within the rapidly evolving NLP landscape.

In contrast, claims from the mid-range years (2009, 2012, 2015) increasingly emphasize areas such as discourse knowledge, semantic tagging, statistical machine translation, translation quality, and the early adoption of neural network models. Many of these topics remain relevant and continue to be explored in contemporary NLP research.

Most notably, the most recent years (2018, 2021, 2024) are characterized by claims focusing on state-of-the-art advances, including LSTM networks, word embeddings, transformer architectures, models such as BERT and RoBERTa, multi-genre natural language inference (MNLI), encoder-decoder models, entropy-based approaches, and large language models. These topics reflect the ongoing paradigm shift in NLP, driven by the development and deployment of increasingly sophisticated deep learning models.

C Techinal Specifications

Models Our study evaluates different open-source LLMs with varying parameter sizes and

[4]https://aclanthology.org/volumes/2024.acl-short/
[5]https://aclanthology.org/volumes/2021.acl-short/
[6]https://aclanthology.org/volumes/P18-2/
[7]https://aclanthology.org/volumes/P15-2/
[8]https://aclanthology.org/volumes/P12-2/
[9]https://aclanthology.org/volumes/P09-2/
[10]https://aclanthology.org/volumes/P06-1/
[11]https://aclanthology.org/volumes/P03-2/
[12]https://aclanthology.org/volumes/P00-1/

benchmark capabilities. Specifically, we utilize instruction-tuned models, including Llama-3.3 (8B, 70B), Qwen-2.5 (7B, 72B) and Gemma-3 (4B, 27B).

Hardware We utilize two types of GPUs: the 16-GB NVIDIA V100 and the 40-GB NVIDIA A100. These GPUs are accessed through nodes in a large cluster, and each node is equipped with multiple GPUs. To enhance the speed of inference, some of our experiments leverage data parallelism.

Implementation Details We used PyTorch to implement the models, borrowing from Hugging-Face (Face, 2021) for Llama-3 Meta AI's third-generation LLM family (Grattafiori et al., 2024) instruction tuned variants including Llama-3.3-8B [13], and Llama-3.3-70B [14]. In addition, we use Alibaba's advaced LLM series Qwen-2.5 (Team, 2024) instruction tuned models inlcuding Qwen-2.5-7B [15], Qwen-2.5-72B [16]. Moreover, further experiments are also performed using instruction-tuned Gemma-3-4B- [17], and Gemma-3-27B- [18], spanning a broad spectrum of computational capacities. For GPT-4, we utilized the official OpenAI API.

Year	Original		Decontextualized	
	Support	Refute	Support	Refute
2000	226	130	89	89
2003	212	115	82	81
2006	762	396	290	289
2009	576	282	215	214
2012	745	173	286	173
2015	662	309	243	242
2018	379	433	203	203
2021	729	350	270	269
2024	263	291	139	138
Total	**4,554**	**2,479**	**1,817**	**1,798**

Table 6: Comparison of Support and Refute label counts per year for the original and decontextualized SciNLP.

D LLM-generated Justifications

The ROUGE scores in our evaluation are computed using the official ROUGE implementation from Google Research. Specifically, we use the RougeScorer class, which calculates ROUGE-1, ROUGE-2, and ROUGE-L. In our code, NLTK's Porter stemmer is enabled to normalize morphological variants of words before comparison, ensuring fairer evaluation across different lexical forms. We report all ROUGE and BERTScore from Table 3 by comparing the input rationales provided to the LLMs with the model-generated justifications across oracle and general experimental settings. This evaluation allows us to quantitatively assess how closely the LLM-generated explanations align with the original reference rationales in terms of content overlap and informativeness. In essence, we report ROUGE-1 to capture surface-level lexical overlap (unigrams), ROUGE-2 to measure local phrase similarity (bigrams), and ROUGE-L to assess sentence-level structure through (longest common subsequence) matching. The detailed results are presented in Table 8. For each setting, the model is required to generate a justification that logically connects the provided evidence to the final veracity prediction. To ensure uniformity and interpretability of the final verdict, we define criteria for acceptable justifications, as specified in our AFC prompt (Figure 9) and described in Section 5. This prompt emphasizes four key components of model-generated justifications. Clarity requires that explanations be concise, coherent, and complete. Relevance ensures that justifications remain directly connected to the claim and its supporting evidence, avoiding unnecessary or unrelated details. Consistency emphasizes logical alignment between the justification and the rationale provided, ensuring internal coherence. Utility focuses on the practical value of the explanation to help readers assess the truthfulness or credibility of the claim.

Table 8 presents the ROUGE and BERTScore results for model-generated justifications across all configurations under both oracle and general settings. Overall, all models achieve the BERTScores (0.81–0.86), indicating semantic coherence and alignment with the provided rationale contexts. This may suggest that model scales contribute to contextually grounded explanations. Across input configurations, the $c + e$ consistently shows the highest ROUGE and BERTScore values. This outcome reflects the notion of supplying models with explicit evidence, allowing them to generate justifications that relate to factual phrases from the input. However, although the lexical similarity for the $(c + e)$ configuration is the highest, this setting

[13] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[14] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[15] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
[16] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[17] https://huggingface.co/google/gemma-3-4b-it
[18] https://huggingface.co/google/gemma-3-27b-it

does not necessarily lead to the best performance in veracity prediction. This discrepancy arises because the retrieved evidence, while lexically similar to the claim, often contains noisy or loosely related content that fails to capture the precise reasoning needed for accurate verification. In many cases, the Contriever model retrieves sentences that share overlapping terminology or thematic relevance but lack direct causal or factual alignment with the claim. Consequently, the high ROUGE and BERTScore in $(c + e)$ largely reflect surface-level lexical overlap, while true veracity assessment requires deeper semantic reasoning and contextual grounding.

Configurations involving the abstract $c + \mathcal{A}$ also perform strongly, as abstracts often summarize the key findings of a paper and thus provide concise, relevant information for reasoning. In contrast, setups that incorporate additional metadata such as titles and keywords $c + t + \mathcal{A} + \mathcal{K}$ tend to show slightly lower lexical overlap due to the difference in length of the model input and justification length and broader contextual scope and lexical variability introduced by these elements.

E Temporal Analysis

We advance the AFC task by analyzing scientific claims from NLP research articles to assess their temporal relevance. To address this, we examine whether LLM-based scientific fact-checking can accurately capture paradigm shifts and the evolution of NLP methodologies over time. To this end, we introduce a zero-shot temporal classification experiment, in which models categorize claims into three novel temporal classes without explicitly providing temporal context (without published year information). We used decontextualized claims for this experiment because of their superior veracity performance. SciNLP provides balanced three-year internal sampling from 2000–2024 to fairly capture temporal trends, and we apply the same prompt shown in Figure 9.

Figure 3 shows the temporal distribution of claims as classified by each model with "Advanced" claims concentrated after 2018 and "Still holds true" claims spanning 2009–2018, indicating sustained but not cutting-edge relevance. "Outdated" claims cluster between 2006 and 2018, reflecting major methodological shifts in NLP, with notable differences between models. Qwen-72B shows

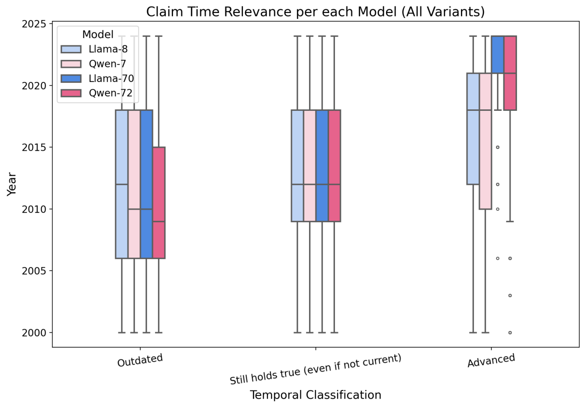

Figure 3: Temporal classifications (Outdated, Still holds true, Advanced) of SciNLP across Llama and Qwen model variants, showing year-wise relevance.

the most accurate temporal classification, especially in "Advanced" and "Outdated" claims, followed by Llama-70B. Smaller models like Qwen-7B and Llama-8B perform similarly on "Still holds true" and "Outdated" claims but struggle with "Advanced" often misclassifying them across 2009–2021. This evaluation highlights that larger LLMs better capture temporal nuances in NLP claims and track methodological shifts over time, while smaller models struggle with fine-grained distinctions.

We further explore the temporal classification dynamics by focusing on a single, high-performing model Qwen-72B to examine how the distribution of temporal claim labels evolves across publication years of NLP articles and data variants. Figure 7 illustrates the year-wise distribution of temporal classifications assigned by this model. Notably, the proportion of claims labeled as "Advanced" (shown by the dark blue bars) increases steadily from 2000 to 2024, mirroring the ongoing innovation and emergence of novel methods in NLP research. In contrast, the incidence of "Outdated" claims declines over the same period, reflecting the field's rapid progression and the obsolescence of earlier approaches. Meanwhile, the "Still holds true" label shows a relatively uniform distribution across years, suggesting that a significant subset of research claims retain their validity and applicability irrespective of the publication year. This analysis underscores the model's sensitivity to evolving research trends and its capacity to capture enduring scientific contributions alongside paradigm shifts. Some sample claims from different years, along with their temporal assessments and temporal justifications produced by Qwen-72B, are presented in Table 11.

E.1 Year-wise Veracity Prediction

In our study, we use the configuration $c + \mathcal{A}$ to present the verification performance of scientific claims in years, as it achieves the optimal results in all models. The oracle setting is chosen to ensure the highest accuracy. Figure 6 illustrates the performance of the F1-macro score intervals over three years of all models under this configuration. It can be seen that, across all models, newer claims, particularly those published after 2015, consistently achieve higher veracity prediction performance. This pattern suggests that recent claims are more linguistically and semantically aligned with the modern pre-trained LLMs representations, reflecting advances in NLP terminology and structure on which these models were trained. In contrast, older claims (for example, 2000–2009) tend to show lower F1 scores, probably due to outdated phrasing, limited contextual detail, or differences in the problem-formulation style used in the early ACL papers, which current LLMs struggle to interpret accurately.

Among the models, Qwen-72B shows the strongest and most stable year-to-year performance, indicating robust generalization across temporal contexts. Llama-70B follows a similar upward trend, while Llama-8B diverges significantly, showing anomalously low performance in 2009, possibly due to overfitting to linguistic patterns not representative of early-era ACL writing. This underperformance suggests that smaller models have limited capacity to generalize between evolving scientific expression and domain-specific shifts. Overall, the improvements observed in recent years indicate that LLMs benefit from exposure to newer linguistic conventions and methodological phrasing in NLP research, allowing them to better capture semantic consistency and factual reasoning in contemporary claims. However, older claims remain challenging due to their lexical and conceptual drift from modern NLP discourse.

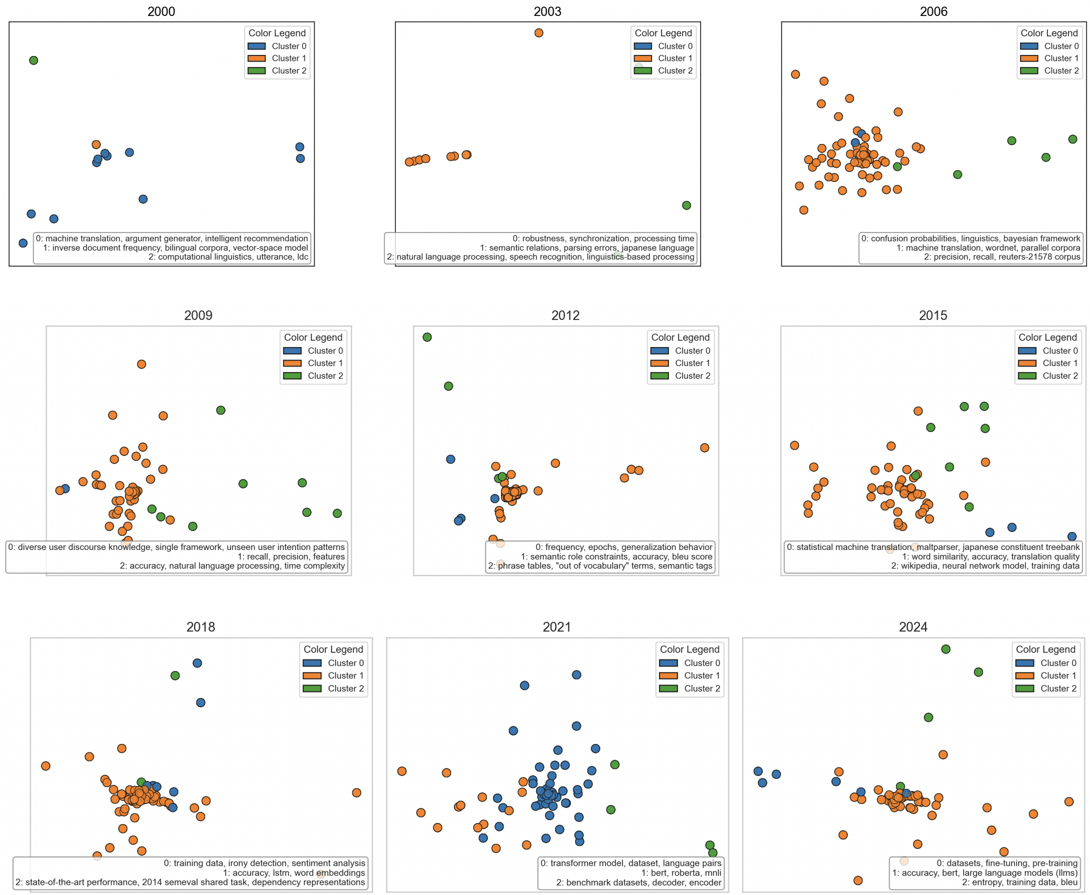

Figure 4: PCA projections of scientific claim keyword clusters of SCINLP for each year. Each subplot displays clustered claims based on their extracted keywords, with the top three keywords per cluster are also shown.

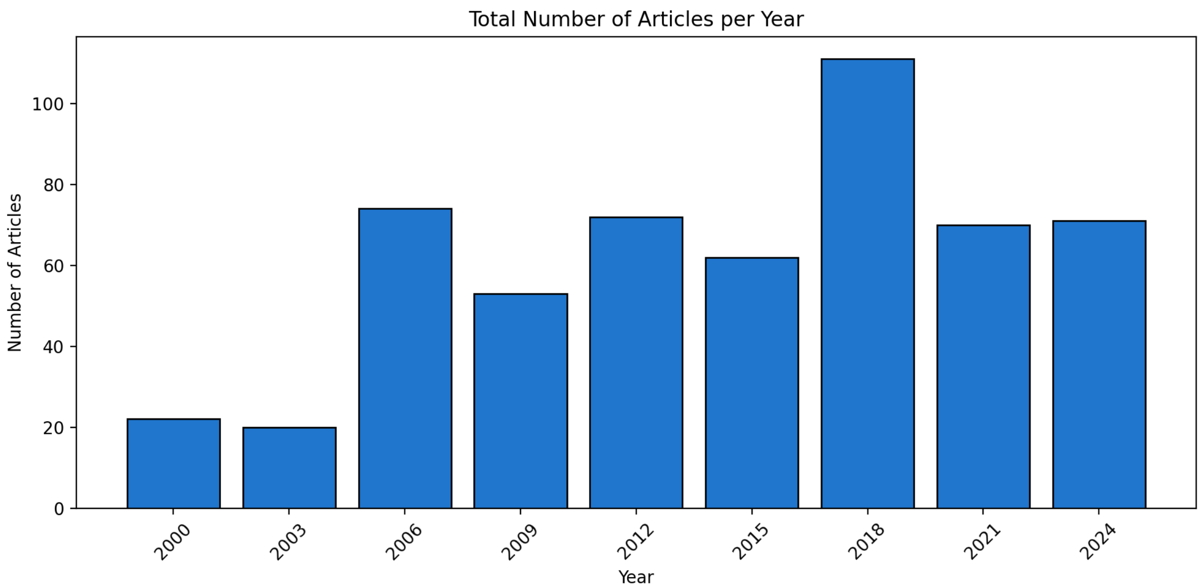

Figure 5: Distribution of number of NLP articles count per year.

Claim	CLAIM DECONTEXTUALIZATION	L	Evidence
Many of the strategies used in the BEETLE II tutorial dialogue system were not successful in improving learning gain.	**Decontexualized Claim:** Many of the error recovery strategies used in the BEETLE II tutorial dialogue system, particularly those dealing with interpretation problems arising from student use of non-standard terminology, were not successful in improving learning gain and user satisfaction. **Subject:** The effectiveness of strategies used in the BEETLE II tutorial dialogue system in improving learning gain. **Disambiguation Criteria:** System Implementation and Learning Outcomes.	S	While typing removes the uncertainty and errors involved in speech recognition, expected student answers are considerably more complex and varied than in a typical spoken dialogue system. We describe an evaluation of an implemented tutorial dialogue system ... which aims to accept unrestricted student input and limit misunderstandings by rejecting low confidence interpretations and employing a range of error recovery strategies depending on the cause of interpretation failure. However, some students in BASE also mentioned that they sometimes were not sure if the system's answer was correcting a problem with their answer, or simply phrasing it in a different way.
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 7: Additional examples from the SciNLP dataset showcasing scientific claims, their decontextualized forms, assigned labels (**L**), and supporting evidence retrieved using FB/Contriever. Labels are denoted as S for *Support* and R for *Refute*.

Model	Oracle Setting				General Setting			
	ROUGE-1	ROUGE-2	ROUGE-L	BERT-Score	ROUGE-1	ROUGE-2	ROUGE-L	BERT-Score
Qwen2.5-7B								
$c+t$	0.3363	0.1558	0.2667	0.8690	0.3335	0.1559	0.2656	0.8683
$c+\mathcal{A}$	0.3354	0.1561	0.2306	0.8374	0.3740	0.1859	0.2570	0.8439
$c+e$	0.3914	0.1745	0.2691	0.8492	0.3971	0.2000	0.2904	0.8551
$c+t+\mathcal{A}+\mathcal{K}$	0.3180	0.1389	0.2145	0.8278	0.3443	0.1579	0.2285	0.8323
$c+t+e+\mathcal{A}+\mathcal{K}$	0.2750	0.1157	0.1818	0.8215	0.2919	0.1321	0.1991	0.8259
Qwen2.5-72B								
$c+t$	0.3007	0.1357	0.2360	0.8669	0.2970	0.1339	0.2326	0.8655
$c+\mathcal{A}$	0.3652	0.1647	0.2368	0.8377	0.3955	0.1875	0.2618	0.8432
$c+e$	0.4245	0.1925	0.2871	0.8510	0.4271	0.2142	0.3068	0.8564
$c+t+\mathcal{A}+\mathcal{K}$	0.3442	0.1482	0.2214	0.8293	0.3639	0.1594	0.2329	0.8324
$c+t+e+\mathcal{A}+\mathcal{K}$	0.2913	0.0980	0.1759	0.8151	0.2964	0.1054	0.1872	0.8182
Llama-3.3-8B								
$c+t$	0.2464	0.1722	0.2202	0.8595	0.2458	0.1712	0.2196	0.8591
$c+\mathcal{A}$	0.4026	0.2007	0.2861	0.8310	0.4206	0.2221	0.3080	0.8367
$c+e$	0.3980	0.2078	0.2972	0.8410	0.4293	0.2356	0.3240	0.8459
$c+t+\mathcal{A}+\mathcal{K}$	0.3346	0.1246	0.2159	0.8127	0.3428	0.1331	0.2254	0.8159
$c+t+e+\mathcal{A}+\mathcal{K}$	0.3130	0.1110	0.1934	0.8073	0.3176	0.1168	0.2034	0.8099
Llama-3.3-70B								
$c+t$	0.2346	0.1277	0.1913	0.8529	0.2327	0.1274	0.1902	0.8521
$c+\mathcal{A}$	0.3706	0.1436	0.2331	0.8228	0.3808	0.1536	0.2428	0.8269
$c+e$	0.3832	0.1526	0.2532	0.8338	0.4070	0.1726	0.2726	0.8380
$c+t+\mathcal{A}+\mathcal{K}$	0.3541	0.1386	0.2245	0.8156	0.3585	0.1429	0.2271	0.8179
$c+t+e+\mathcal{A}+\mathcal{K}$	0.3381	0.1265	0.2071	0.8112	0.3418	0.1325	0.2155	0.8137

Table 8: ROUGE (F1) and BERTScore (F1) across different input settings for each model under Oracle and General settings.

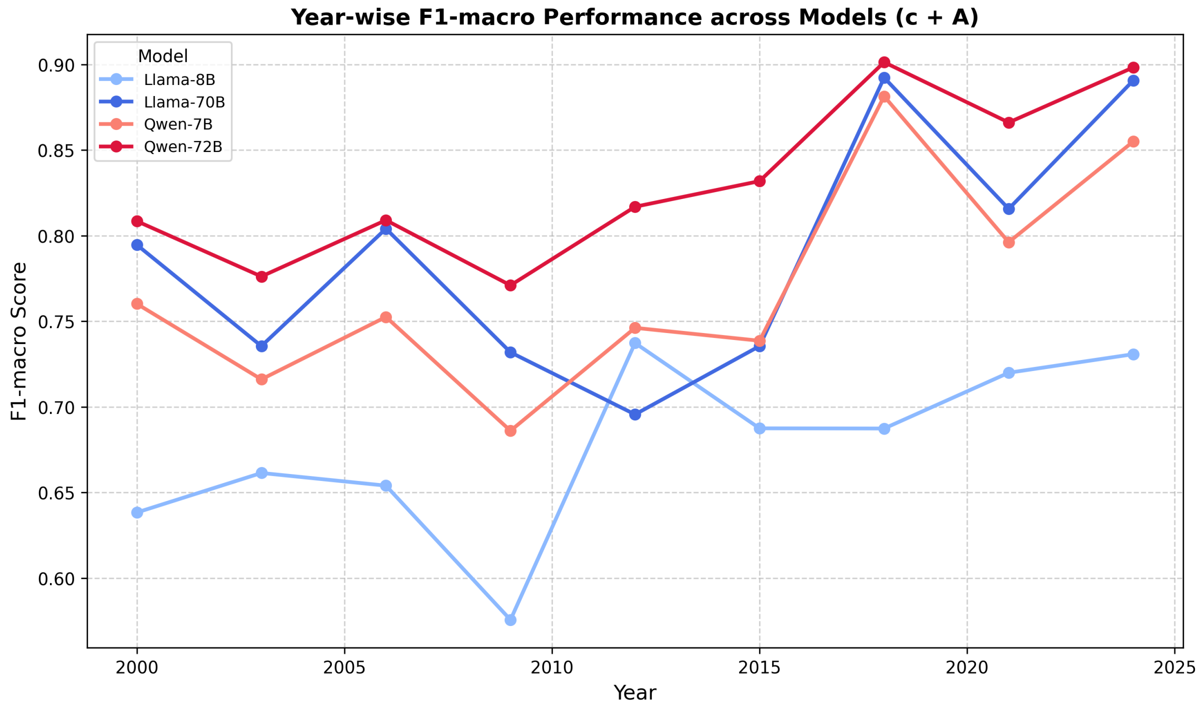

Figure 6: Year-wise performance of each model in oracle setting under the $c + \mathcal{A}$ configuration.

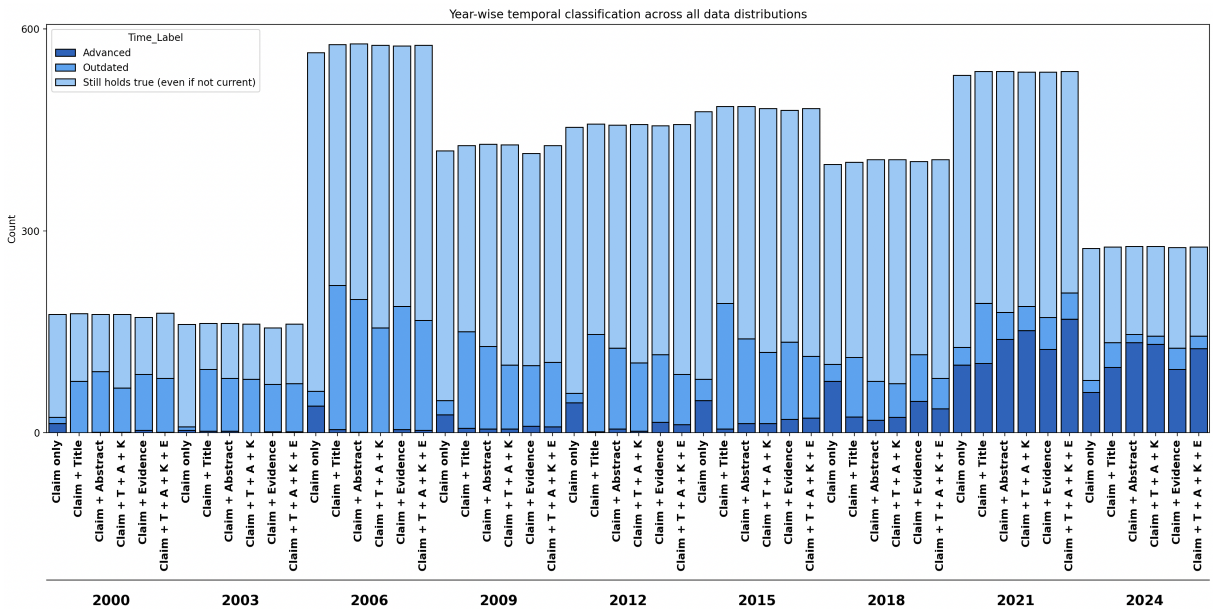

Figure 7: Temporal trend in scientific claim classification by Qwen-2-72B across various data distributions on decontextualized SciNLP.

Claim	Oracle Evidence	L	General Evidence
Despite limitations, phrase-based SMT systems have achieved competitive results in Arabic-to-English benchmark evaluations.	As in related work on syntactic reordering by preprocessing, our method attempts to make Arabic and English word order closer to each other by reordering Arabic VS constructions into SV. (2007) limit reordering to decoding for Chinese-English SMT using a lattice representation. We use a 5-gram language model with modified Kneser-Ney smoothing.	S	With so many classification errors, the effect of this baseline in an SMT system is in doubt, even with a powerful language model. On both the reordering classification and a Chinese-to-English translation task, we show improved performance over a baseline SMT system. Indeed both systems produced similar word accuracy, but our MT system does better in phrase reordering and produces more fluent translations.
The HotFlip method can be used for effective adversarial training.	Furthermore, fast generation of adversarial examples allows feasible adversarial training, which helps the model defend against adversarial examples and improve accuracy on clean examples. While this line of research has recently received a lot of attention in the deep learning community, it has a long history in machine learning, going back to adversarial attacks on linear spam classifiers (Dalvi et al., 2004; Lowd and Meek, 2005). A projected gradient-based approach to create adversarial examples by Madry et al.	S	In particular, adversarial training on real adversarial examples generated by HotFlip, is more effective than training on pseudo-adversarial examples created by adding noise to the embeddings. HotFlip is a method for generating adversarial examples with character substitutions ("flips"). Our work is the first to propose an efficient method to generate real-world adversarial examples which can also be used for effective adversarial training.
There is a class of formal languages that can be recognized by LSTMs but not by GRUs.	The construction of (Siegelmann, 1999) implements pushing 0 into a binary stack by the operation g $\leftarrow$g/4 + 1/4. Another operation of interest is comparing two counters (for example, checking the difference between them). When unbounded computation is allowed, a 2-counter machine has Turing power.	S	This formalism is both elementary and powerful enough to strongly simulate many grammar formal- isms, such as rewriting systems, dependency grammars, TAG, HPSG and LFG. This paper presents a new family of formalisms, Polarized Unification Grammars (PUGs). HPSG (Head-driven Phrase Structure Grammar) There are two ways to translate feature structures (FSs) into PUG.
REPE uses hard contrastive learning objectives to model the alignment between the representations of original sentences and code-switched counterparts.	CS: code-switched sentence. In contrast, our REPE effectively maintains contextual structural information, successfully recognizing fixed expressions like "would like" and "make a stop". In this work, we denote the original and corresponding code-switched sentence as x = w1, w2, .	R	REPE utilizes OT to achieve soft alignment between representations of original and code-switched sentences to preserve structural information within languages. Specifically, REPE introduces optimal transport to facilitate soft alignment between the representations of code-switched and original sentences, thereby preserving structural integrity as much as possible. Besides, REPE introduces multi-view learning to predictions of original and code-switched sentences for further alignment and self-distillation to boost the performance.
SOFT-SC exhibits worse scaling with model size than SC, decreasing performance as model size increases.	While SC does improve the success rate of Mistral-7B on ALFWorld with increasing k, SOFT-SC yields greater performance gains. Soft-SC: Scaling with model size on Bash (test). SOFT-SC exhibits better sample efficiency i.e., produces better performance than SC with fewer samples (cf. This points to additional efficiency gains from SOFT-SC, as it can allow smaller models to replace larger ones."	R	SOFT-SC effectively scales with model size. SOFT- SC improves over SC for all model sizes. SOFT-SC improves black-box models more than SC.
The NCD measure matches the performance of the state-of-the-art MT evaluation measures in English.	However, manual evaluation is important in the comparison of different MT systems and for the validation and development of automatic MT evaluation measures, which try to model human assessments of translations as closely as possible. We compare mNCD against NCD and relate their performance to other MT evaluation measures. (2008), which allows us to measure how well MT evaluation measures correlate with human judgments on the system level.	R	This paper proposes to utilize second-language proiciency tests (SLPTs), such as TOEIC, as the source of the specimens for extrinsic evaluation of MT systems. The main criticisms include: 1) unreliability in evaluating short translations, 2) non-interpretability of the scores beyond numerical comparison, and 3) bias towards statistical MT systems. Task-based mt evaluation: From who/when/where extraction to event understanding.

Table 9: Examples from the SCINLP dataset showcasing scientific claims, their oracle evidence (directly from article), ground labels (**L**), and general evidence (from SCINLP corpus) retrieved using FB/Contriever. Labels are denoted as S for *Support* and R for *Refute*.

Decontextualized Claim	GT	Pred.	Evidence	Justification (Llama-3.3-70B)
The X-FACT dataset, introduced as the largest publicly available multilingual dataset for factual verification of real-world claims, covers 25 typologically diverse languages across 11 language families.	Support	True	First, results from table 3 show that augmenting models with metadata is helpful. This might be due to the evidence-based over-fitting the in-domain data. We left out some fact-checkers because either the claims on their websites were not well specified or the fact-checker did not use any rating scale.	The claim that the X-FACT dataset covers 25 typologically diverse languages across 11 language families is supported by the abstract, which mentions that the dataset contains short statements in 25 languages. This information is clear, relevant, and consistent with the provided evidence, making the claim accurate.
The research demonstrates that low-rank tensors, which require about two orders of magnitude fewer parameters per verb in compositional distributional semantics, cannot match the performance of full, unconstrained-rank tensors in tasks such as sentence similarity and verb disambiguation.	Refute	False	Regardless, we show that the low-rank tensors are able to achieve performance comparable to the full rank for both types of vectors. By fixing a value for R that is sufficiently small compared to S and N (forcing the verb tensor to have rank of at most R), and directly learning the parameters of the low-rank approximation using gradient-based optimization, we learn a low-rank tensor requiring fewer parameters without ever having to store the full tensor. In the process, the model learns vector embeddings for both the SVO triples and for the words in the sentences such that SVO vectors have a high dot product with their contextual word vectors.	The claim that low-rank tensors cannot match the performance of full tensors is contradicted by the evidence provided in the paper. The results show that low-rank tensors can achieve performance comparable to, and occasionally surpassing, the unconstrained-rank tensors on sentence similarity and verb disambiguation tasks. This suggests that the claim is not accurate.
The Simultaneous Machine Translation (SiMT) model does not fit the training data well for hallucination words due to the limited source-side information.	Support	True	While many efforts have been made to enhance performance for SiMT, few of them attempt to understand and analyze hallucination in SiMT. It successfully reduces hallucination words using more target-side context, but also indirectly increases hallucination words using more source-side context, as shown in Figure 5. This means that there maybe an increased likelihood of hallucinations occurring in words that are utilized with limited source-side context Hallucination words use more target context than Non-Hallucination words.	"The claim is supported by the evidence provided, which suggests that the SiMT model's performance on hallucination words is affected by the limited source context. The abstract and the extracted text both mention the issue of hallucination in SiMT due to the absence of source-side information, and the experiments demonstrate that decreasing the over-usage of target-side information can alleviate hallucination.
Soft Self-Consistency (SOFT-SC), a method that replaces self-consistency's discontinuous scoring with a continuous score computed from model likelihoods, exhibits worse scaling with model size than self-consistency (SC), leading to a decrease in performance as the size of large language models increases.	Refute	False	While SC does improve the success rate of Mistral-7B on ALFWorld with increasing k, SOFT-SC yields greater performance gains in 7B, 13B 34B 70B Model Size Success Rate SC Soft-SC Figure 2: Scaling with model size on Bash (test). SOFT-SC exhibits better sample efficiency, i.e., produces better performance than SC with fewer samples. This points to additional efficiency gains from SOFT-SC, as it can allow smaller models to replace larger ones.	The claim 'SOFT-SC exhibits worse scaling with model size than SC, decreasing performance as model size increases' is not supported by the provided evidence. In fact, the evidence suggests that SOFT-SC yields greater performance gains than SC, especially with fewer samples, and can allow smaller models to replace larger ones. This indicates that SOFT-SC may exhibit better scaling with model size than SC, contrary to the claim.

Table 10: Sample decontexualized claims with corresponding ground truth (GT) labels, veracity predictions (Pred.), evidence, and veractiy justification generation produced by Llama-3.3-70B model.

Year	Claim	Veracity	Time Label	Temporal Justification (Qwen-2.5-72B)
2024	*The fine-tuned Flan-T5 model generates biographies based on the given personal attributes.*	Support	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.
2024	*The UltraSparseBERT model uses all of its 4095 neurons for each layer inference.*	Refute	Advanced	UltraSparseBERT represents a cutting-edge approach in NLP, specifically in the domain of sparse modeling and conditional computation for transformer-based models. Its conditional sparsity and the reported speedups over conventional dense models on both CPUs and GPUs highlight its innovation and efficiency gains. As of the time of writing, there are few models that achieve such high levels of sparsity without compromising performance.
2021	*MOLEMAN performs nearest neighbor inference on an index of 700 million mentions.*	Support	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. However, MOLEMAN's methodology remains valid and useful for understanding instance-based approaches. Relevant newer works include those utilizing BERT and its variants for entity linking, which have shown higher accuracy due to their ability to better capture context.
2018	*Generalized chart constraints slow down chart parsers for PCFGs and TAG parsing.*	Refute	Still holds true	While the specific methods mentioned in the claim might not be the latest advancements in NLP, the concept of using constraints to optimize parsing efficiency remains relevant and valid. More recent research has continued to explore and improve upon parsing optimization techniques, indicating that the core idea of using constraints for efficiency is still sound.
2010	*The KA* algorithm makes use of outside edge items which represent the many possible derivations of a sentence.*	Support	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. These newer methods often offer better performance and efficiency compared to traditional algorithmic approaches like KA*. Notable advancements include the use of transformer models for syntactic parsing, which have shown significant improvements in accuracy and speed.
2015	*The paper presents a method to improve the hierarchical phrase-based statistical machine translation system by learning word reorderings for word pairs with different distances.*	Support	Outdated	While the claim was innovative at the time of publication, subsequent advancements in neural machine translation (NMT) models have significantly outpaced traditional statistical machine translation techniques. NMT models, particularly those based on Transformer architectures, can implicitly handle word reorderings without the need for explicit sub-models, offering superior translation quality and efficiency.

Table 11: Sample claims across different years with corresponding veracity label, temporal assessment, and temporal justifications produced by Qwen-2.5-72B model.

Scientific RQ Extraction Prompt

```
SYSTEM:
You are an advanced scientific research question extractor specializing in automated fact-checking.
Your task is to analyze manuscripts and extract scientific research claims or research questions.

Input:
Given the following article [ARTICLE]
Title: [TITLE]


Instructions:
Please perform two tasks:

# 1. Identify **explicitly stated or implied research claims or research questions**, ensuring they are
**detailed, comprehensive, and complex**. Make atleast 12 claims. Do not refer to "the authors" or "the
method" or "the research" in generic terms instead, name the specific technique, dataset, framework, or
concept used in paper—**.".
# - Format each claim as follows:
# **Claim:** [Lengthy, informative statement about the research finding, technique or approach,
hypothesis, empirical evaluation, or contribution.
Avoid vague phrasing. For each claim, create a research question that is fully understandable on its
own, without requiring additional context from the article. Replace any abbreviations, acronyms, or
pronouns with their complete forms to eliminate ambiguity.]
# **Support or Refute Label:** [Support/Refute]
# Generate **at least six reversed claims** with a "Refute" label, ensuring they logically contradict an
existing claim and refrain from the use of explicit negation keywords such as "not". For each refuted
claim, create a research question that is fully understandable on its
own, without requiring additional context from the article.
# **Reason:** [A detailed 4 to 5 sentences justification that directly refers to specific datasets,
architectures, evaluation results, comparisons with baselines, or other methodological insights; with
explicit reference to methodology, data, results, or conclusions that justify your labeling. The reason
should provide strong, context-aware support for the claim.]

# 2. Extract approximately 10-20 of the most relevant and common keywords/frequent words present in the
text (e.g., models used, datasets mentioned, evaluation techniques, frameworks).

# Clearly list these keywords separately under:
# **Keywords Extracted:** keyword1, keyword2, keyword3, ...
```

Figure 8: Prompt used with GPT-4 to extract scientific research questions from ACL Anthology short articles.

DECONTEXTUALIZATION
{system_prompt}

```
You are an expert in scientific claim decontextualization and ambiguity analysis.
```

{user_prompt}

```
TASK 1: DECONTEXTUALIZATION CRITERIA:
Decontextualization adds the right type of information to a CLAIM to make it
standalone.
This process can modify the original CLAIM in the following manners:
- Substituting pronouns or incomplete names with the specific subject being
referred to.
- Including contextual information to provide more context about the subject.
Instructions:
- Identify the "subject" of the claim and locate the claim within the context.
- Use the CONTEXT to substitute any incomplete technique, the research paper,
the proposed method, dataset, approach in the CLAIM.
- If there is no decontextualization necessary, return the original claim and
evidence as is.
- The decontextualization should minimally modify the claim and evidence by only
adding necessary contextual information.
- Refer to the following examples to understand the task and output formats.
- If the label of the CLAIM is Refute, write in the same sense rather than
correcting it but complete the contextual information.
Similarly, generate a decontextualized claim for the following pair of CLAIM and
CONTEXT making minimal alterations to the original structure  of the CLAIM while
ensuring clarity and coherence.
```

INPUT

```
{Claim}:{{claim}}
{Title}:{{title}}
{Abstract}:{{abstract}}
```

Table 12: The task prompt used for decontextualization of the claims.

AMBIGUITY CHECK

{system_prompt}

```
You are an expert in scientific claim decontextualization and ambiguity analysis.
```

{user_prompt}

```
TASK 2: AMBIGUITY CRITERIA:
Ambiguity manifests in diverse forms, including:
- Similar names denoting distinct entities.
- Varied interpretations stemming from insufficient information.
- Multiple understandings arising from vague or unclear information.
Instructions:
- Identify the main SUBJECT within the claim based on the CONTEXT and KEYWORDS provided.
- Determine if the SUBJECT is ambiguous according to the provided AMBIGUITY CRITERIA.
- Utilize your world knowledge and keywords provided to enumerate potential DISAMBIGUATIONS for the identified SUBJECT.
- Specify the TYPE of information employed for disambiguation based on the list of DISAMBIGUATIONS.
- If the SUBJECT does not have ambiguous interpretations, return None
 Similarly generate the subject and disambiguation critera for the following CLAIM, CONTEXT, KEYWORDS, and DECONTEXTUALIZED CLAIM provided in the exact same format as examples
```

INPUT

```
{Claim}:{{claim}}
{CONTEXT} {Title}:{{title}}
{CONTEXT} {Abstract}:{{abstract}}
{Keywords}:{{keywords_extracted}}
{Decontextualized_Claim}:{{decontextualized_claim}}
```

Table 13: The task prompt used for ambiguity checking of the claims.

Fact-Checking Prompt

```
SYSTEM:
You are a helpful assistant for automated fact-checking and temporal analysis of scientific claims.

USER:
You are an intelligent decision support system designed for automated fact-checking and temporal
analysis (“time-shift”) of scientific claims in the field of Natural Language Processing (NLP). For each
claim, assess whether it represents an advanced approach, is outdated, or still holds true even if it is
not the most current method. Respond strictly in the following structured JSON format:

{
  "Veracity": "True or False”,
  "Justification": "Detailed reasoning addressing clarity, relevance, consistency, utility”,
  "Time_Label": "Advanced, Outdated, Still holds true (even if not current)”,
  "Justification_Time_Label": "Detailed rationale explicitly addressing why this specific temporal label
  applies. Focus on clarity, relevance, consistency, and utility.”
}

Definitions for Veracity labels:
- True: "The statement is accurate and there’s nothing significant missing.”
- False: "The statement is not accurate or makes a ridiculous claim.”

Definitions for Time_Label:
- Advanced: The claim describes a cutting-edge or state-of-the-art approach that reflects the latest
advances in NLP.
- Outdated: The claim refers to a method or model that has been surpassed by newer developments and is
no longer commonly used in NLP.
- Still holds true (even if not current): The claim is not the latest, but remains accurate and valid,
even though newer alternatives exist in NLP.

Guidelines for Justification:
- Clarity: Concise, coherent, and complete.
- Relevance: Directly relates to claim and context.
- Consistency: Aligns with evidence provided.
- Utility: Useful for evaluating claim accuracy.

Guidelines for Justification_Time_Label:
- A scientific claim related to natural language processing (NLP) is under analysis.
- Clarity: Is this claim now considered old or outdated based on recent advances in NLP?
- Relevance: Are there newer methods, models, or technologies developed that address the same problem or
task in a different or improved way? If yes, please briefly describe them.
- Consistency: Matches historical/current evidence. Provide references (preferably from recent years) to
papers, tools, or resources that demonstrate the new approaches.
- Utility: If possible, briefly compare the original claim's approach with the newer developments.

Ensure that all the information is correctly placed in a structured JSON format.


INPUT:
Claim: {claim}
Title: {title}
Evidence: {evidence}
Abstract: {abstract}
Keywords: {keywords_extracted}
```

Figure 9: Prompt for the veracity assessment, veracity justification, temporal assessment, and temporal justification of the scientific research claims in SciNLP.

K-shot Examples for Claim Decontextualization.

```
Example 1:
CONTEXT: Title: Improving Arabic-to-English Statistical Machine Translation by Reordering Post-Verbal
Subjects for Alignment.
Abstract: “We study the challenges raised by Arabic verb and subject detection and reordering in
Statistical Machine Translation (SMT). We show that post-verbal subject (VS) constructions are hard to
translate because they have highly ambiguous reordering patterns when translated to English. In addition
implementing reordering is difficult because the boundaries of VS constructions are hard to detect
accurately, even with a state-of-the-art Arabic dependency parser. We therefore propose to reorder VS
constructions into SV order for SMT word alignment only. This strategy significantly improves BLEU and
TER scores, even on a strong large-scale baseline and despite noisy parses.”
CLAIM: Standard phrase-based SMT systems do not capture any generalizations between occurrences in VS
and SV orders. Label: Support
DECONTEXTUALIZED_CLAIM: Standard phrase-based Statistical Machine Translation (SMT) systems do not
capture any generalizations between occurrences of Arabic post-verbal subject (VS) constructions and
subject-verb (SV) orders.

Example 2:
CONTEXT: Title: Sense-Aware Neural Models for Pun Location in Texts.
Abstract: “A homographic pun is a form of wordplay in which one signifier (usually a word) suggests two
or more meanings by exploiting polysemy for an intended humorous or rhetorical effect. In this paper,
we focus on the task of pun location, which aims to identify the pun word in a given short text. We
propose a sense-aware neural model to address this challenging task. Our model first obtains several WSD
results for the text, and then leverages a bidirectional LSTM network to model each sequence of word
senses. The outputs at each time step for different LSTM networks are then concatenated for prediction.
Evaluation results on the benchmark SemEval 2017 dataset demonstrate the efficacy of proposed model.”
CLAIM: The task of identifying the pun word is known as pun location, which is an easy task. Label:
Refute
DECONTEXTUALIZED_CLAIM: The task of identifying the pun word in short texts, known as pun location, is
an easy task, even when using sense-aware neural models such as a bidirectional LSTM network.

Example 3:
CONTEXT: Title: Automatic Extraction of Commonsense LOCATEDNEAR Knowledge
Abstract: “LOCATEDNEAR relation is a kind of commonsense knowledge describing two physical objects that
are typically found near each other in real life. In this paper, we study how to automatically extract
such relationship through a sentence-level relation classifier and aggregating the scores of entity
pairs from a large corpus. Also, we release two benchmark datasets for evaluation and future research.”
CLAIM: The paper proposes a method to automatically extract the commonsense LOCATEDNEAR relation between
physical objects from textual corpora. Label: Support
DECONTEXTUALIZED_CLAIM: The paper proposes a sentence-level relation classification method to
automatically extract the commonsense LOCATEDNEAR relation between physical objects from large textual
corpora.

Example 4:
CONTEXT: Title: PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields
Abstract: “Recent emerged phrase-level topic models are able to provide topics of phrases, which are
easy to read for humans. But these models are lack of the ability to capture the correlation structure
among the discovered numerous topics. We propose a novel topic model PhraseCTM and a two-stage method to
find out the correlated topics at phrase level. In the first stage, we train PhraseCTM, which models the
generation of words and phrases simultaneously by linking the phrases and component words within Markov
Random Fields when they are semantically coherent. In the second stage, we generate the correlation of
topics from PhraseCTM. We evaluate our method by a quantitative experiment and a human study, showing
the correlated topic modeling on phrases is a good and practical way to interpret the underlying themes
of a corpus.”
CLAIM: The correlated topic modeling on phrases is not a practical way to interpret the underlying
themes of a corpus. Label: Refute
DECONTEXTUALIZED_CLAIM: Correlated topic modeling on phrases using the PhraseCTM model is not a
practical way to interpret the underlying themes of a text corpus.

Example 5:
CONTEXT: Title: Semantic Frame Induction using Masked Word Embeddings and Two-Step Clustering
Abstract: Recent studies on semantic frame induction show that relatively high performance has been
achieved by using clustering-based methods with contextualized word embeddings. However, there are two
potential drawbacks to these methods: one is that they focus too much on the superficial information of
the frame-evoking verb and the other is that they tend to divide the instances of the same verb into too
many different frame clusters. To overcome these drawbacks, we propose a semantic frame induction method
using masked word embeddings and two-step clustering. Through experiments on the English FrameNet data,
we demonstrate that using the masked word embeddings is effective for avoiding too much reliance on the
surface information of frame-evoking verbs and that two-step clustering can improve the number of
resulting frame clusters for the instances of the same verb.”
CLAIM: The proposed method uses masked word embeddings of frame-evoking verbs in addition to standard
contextualized word embeddings of frame-evoking verbs. Label: Support
DECONTEXTUALIZED_CLAIM: The proposed semantic frame induction method using masked word embeddings and
two-step clustering employs masked word embeddings of frame-evoking verbs in addition to standard
contextualized word embeddings of frame-evoking verbs.

Example 6:
CONTEXT: Title: UltraSparseBERT: 99% Conditionally Sparse Language Modelling.
Abstract: “Language models only really need to use a tiny fraction of their neurons for individual
```

Figure 10: Few-shot examples supplied to GPT-4 for claim decontextualization.

K-shot Examples for Claim Decontextualization (Continued).

```
inferences. We present UltraSparseBERT, a BERT variant that uses 0.3% of its neurons during inference
while performing on par with similar BERT models. UltraSparseBERT selectively engages just 12 out of
4095 neurons for each layer inference. This is achieved by reorganizing feedforward networks into fast
feedforward networks (FFFs). To showcase but one benefit of high sparsity, we provide an Intel MKL
implementation achieving 78x speedup over the optimized feedforward baseline on CPUs, and an OpenAI
Triton implementation performing forward passes 4.1x faster than the corresponding native GPU
implementation. The training and benchmarking code is enclosed.”
CLAIM: UltraSparseBERT uses the same number of neurons as other BERT models during inference. Label:
Refute
DECONTEXTUALIZED_CLAIM: UltraSparseBERT, a variant of BERT that employs fast feedforward networks for
conditional sparsity, uses the same number of neurons as standard BERT models during inference.
```

Figure 11: Few-shot examples supplied to GPT-4 for claim contextualization (continued).

K-shot Examples for Ambiguity Check

```
Example 1:
##CLAIM##: The researchers propose a new paradigm of grounding comparative adjectives describing colors
as directions in RGB space such that the colors along the vector, rooted at the reference color, satisfy
the comparison. Label: Support
CONTEXT: Title: Lighter Can Still Be Dark: Modeling Comparative Color Descriptions
Abstract: We propose a novel paradigm of grounding comparative adjectives within the realm of color
descriptions.
Given a reference RGB color and a comparative term (e.g., ‘lighter’, ‘darker’), our model learns to
ground the comparative as a direction in the RGB space such that the colors along the vector, rooted at
the reference color, satisfy the comparison. Our model generates grounded representations of comparative
adjectives with an average accuracy of 0.65 cosine similarity to the desired direction of change.
These vectors approach colors with Delta-E scores of under 7 compared to the target colors, indicating
the differences are very small with respect to human perception. Our approach makes use of a newly
created dataset for this task derived from existing labeled color data.
Keywords: {“comparative adjectives”, “color descriptions”, “RGB color space”, “model”, “grounding”,
“vector”, “cosine similarity”, “Delta-E scores”, “dataset”, “labeled color data”, “reference color”,
“target color”, “direction of change”, “deep learning model”, “network architecture”, “training”,
“testing”, “accuracy”, “performance”}
DECONTEXTUALIZED_CLAIM: The paper “Lighter Can Still Be Dark: Modeling Comparative Color Descriptions”
proposes a novel paradigm for grounding comparative adjectives describing colors as directions in RGB
space, such that the colors along the vector, rooted at the reference color, satisfy the comparison.
SUBJECT: grounding comparative adjectives within the realm of color descriptions.
DISAMBIGUATION_CRITERIA: Technique or Approach

Example 2:
##CLAIM##: The paper claims that the use of either the connective or syntactic features alone results in
better disambiguation performance than using both together. Label: Refute
CONTEXT: Title: Using Syntax to Disambiguate Explicit Discourse Connectives in Text
Abstract: Discourse connectives are words or phrases such as once, since, and on the contrary that
explicitly signal the presence of a discourse relation. There are two types of ambiguity that need to be
resolved during discourse processing. First, a word can be ambiguous between discourse or non-discourse
usage. For example, once can be either a temporal discourse connective or a simply a word meaning
“formerly”. Secondly, some connectives are ambiguous in terms of the relation they mark. For example,
since can serve as either a temporal or causal connective. We demonstrate that syntactic features
improve performance in both disambiguation tasks. We report state-of-the-art results for identifying
discourse vs. non-discourse usage and human-level performance on sense disambiguation.
Keywords: discourse connectives, syntactic features, discourse vs. non-discourse, ambiguity, relation
sense, Penn Discourse Treebank (PDTB), Expansion, Comparison, Contingency, Temporal, maximum entropy
classifier, disambiguation, implicit relations, explicit relations, NSF grants.
DECONTEXTUALIZED_CLAIM: The paper “Using Syntax to Disambiguate Explicit Discourse Connectives in Text”
claims that using either connective features or syntactic features alone leads to better disambiguation
performance than using both feature types together.
SUBJECT: Impact of connective and syntactic features, used alone or in combination, on disambiguation of
explicit discourse connectives.
DISAMBIGUATION_CRITERIA: Feature Integration and Experimental Context.
```

Figure 12: Few-shot examples supplied to GPT-4 for ambiguity checking.