HindiSciVerify: A New Dataset for Evidence-based Scientific Claim Verification in Hindi

Anonymous ACL submission

Abstract

A s scientific communication becomes more global, there's a growing need for automated tools to verify scientific claims across different languages and complex evidence types. Most existing research has focused on English datasets accompanying true or false labels, but real-world use cases are never straightforward. The evidence for these scientific claims is often too technical, involving highly scientific jargon and a deep understanding of domain-specific concepts, which requires domain-grounded specific models that, in turn, are time-consuming. Therefore, to bridge the existing gaps, we propose HindiSciVerify - a new dataset designed to verify scientific claims in Hindi. The paper further describes how we extracted and annotated claims and the corresponding evidence related to these claims and built baseline models to tackle the claim verification task. Each claim is labeled as supported, refuted, or not enough information(neutral), based on evidence retrieved from the corresponding text. The findings reveal the difficulties of scientific fact-checking and show how HindiSciVerify can serve as a highly sound resource for training models ,benchmarking other models and further research in scientific claim verification.

1 Introduction

The proliferation of digital information has made the ability to distinguish credible scientific information from misinformation a paramount societal challenge, particularly in contexts like public health where misinformation can have dire consequences (Islam et al., 2020). To combat this, automated fact-checking has emerged as a vital research area, with scientific claim verification standing out as a particularly demanding task. At its core lies Evidence-Claim Matching (Wadden et al., 2020), a process that requires nuanced reasoning to determine whether scientific evidence truly supports or refutes a given statement. This is inherently non-trivial, as it often demands multi-hop reasoning across disparate text snippets and a deep understanding of complex terminology where a single word choice can entirely invert a claim's meaning.

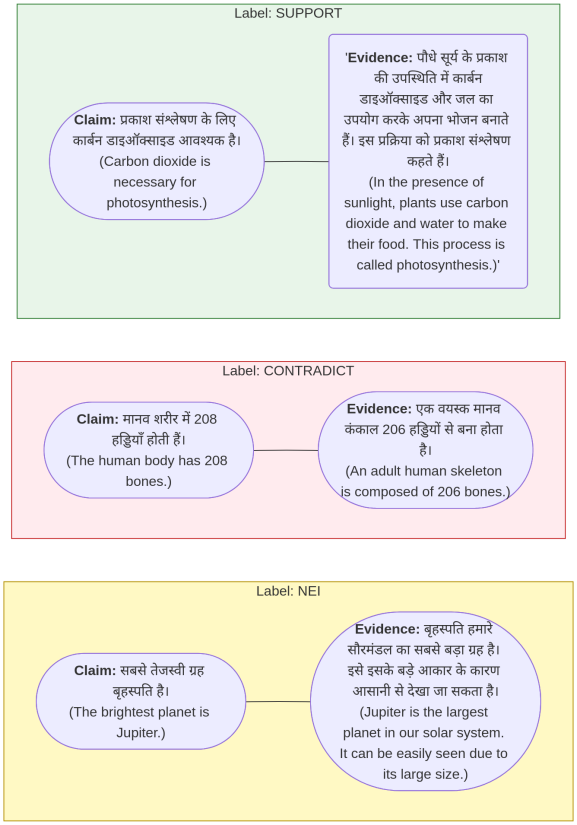

Figure 1: Examples from the HindiSciVerify dataset showcasing the three target labels.

Significant progress has been made in high-resource languages like English (Li et al., 2024), supported by foundational datasets such as FEVER (Thorne et al., 2018) and domain-specific benchmarks like SciFact (Wadden et al., 2020). However, these advancements have not been mirrored in the Indic landscape, where a lack of specialized corpora creates a profound technological and ac-

cessibility gap for hundreds of millions of Hindi speakers, who face unique challenges due to linguistic and morphological complexities (Naik and Chakravarthy, 2016). The lack of dedicated resources means that automated systems cannot effectively serve these populations, leaving them more vulnerable to the spread of unverified information.

This paper aims to bridge that gap by introducing **HindiSciVerify**, the first large-scale, three-way annotated dataset for scientific claim verification in Hindi. We develop a novel, semi-automatic data curation pipeline that leverages the capabilities of modern Large Language Models (LLMs) for claim generation, evidence retrieval, and counterfactual data creation. Our methodology employs advanced techniques such as Parameter-Efficient Fine-Tuning (PEFT) and the modular composition of LoRA adapters to create a high-quality, balanced dataset of 24,000 instances derived from Indian science textbooks.

We conduct extensive experiments to establish strong performance baselines on HindiSciVerify using state-of-the-art models, including Mixture-of-Experts (MoE) architectures. Furthermore, we provide a deep analysis of the challenges inherent in this task, exploring cross-lingual generalization, the efficacy of our data curation choices through ablation studies, and a detailed analysis of model failure modes, particularly in the context of Evidence-Claim Matching.

Our primary contributions are threefold:

1. We release **HindiSciVerify**, a new, large-scale benchmark for scientific claim verification in Hindi.
2. We present a scalable and robust methodology for creating complex NLP datasets in low-resource languages using a cascade of LLMs and PEFT techniques.
3. We provide a comprehensive analysis of model performance, cross-lingual transfer capabilities, and task-specific challenges, offering valuable insights for future research.

2 Related Work

Automated fact verification has matured with foundational datasets like FEVER (Thorne et al., 2018) for general claims and SciFact (Wadden et al., 2020) for scientific claims. This line of work has highlighted persistent challenges in the field, such as the need for multi-hop reasoning across evidence (Lu et al., 2023), robust numerical reasoning (Cobbe et al., 2021), and effective evidence retrieval from large corpora (Karpukhin et al., 2020). Our work builds on these established paradigms but addresses a critical and undeserved area.

While global research has flourished, the Indic landscape faces a unique ”digital divide” where general-purpose models often fail to grasp the technical nuances of Hindi scientific prose. Recent efforts like IndicClaimBuster (Pal et al., 2025) and FactDRIL (Singhal et al., 2022) have begun to address this by curating multi-domain datasets specifically for Indian languages. These benchmarks, alongside reasoning frameworks like IndicXNLI (Aggarwal et al., 2022), highlight the difficulty of transferring logical verification across language barriers. Ultimately, the shift toward specialized corpora like HealthVer (Sarrouti et al., 2021) underscores that verifying a scientific claim requires more than just translation; it requires a model that understands the high-stakes vocabulary of the domain itself.

This gap is most pronounced in morphologically rich languages like Hindi; while not low-resource in a general sense, it remains significantly under-resourced in scientific datasets, fact-verification frameworks, and technical research literature (Karthika et al., 2025). Consequently, while foundational multilingual models exist, large-scale datasets for complex reasoning tasks in Indic languages remain scarce. To bridge this gap efficiently, our methodology leverages Parameter-Efficient Fine-Tuning (PEFT) techniques, particularly LoRA (Hu et al., 2022) and adapter composition (Pfeiffer et al., 2021), as a computationally tractable approach to specialize large models for this new domain and language.

3 Methodology

Our methodology advances the framework of automated scientific claim generation, bridging the gap between established English-centric benchmarks and the unique linguistic demands of the Hindi pedagogical domain. Building upon the foundational logic of SciClaimHunt (Kumar et al., 2025), we implement a multi-stage pipeline designed to capture the semantic precision required for scientific discourse while navigating Hindi's morphological complexity. By integrating a semi-automated curation workflow with Parameter-Efficient Fine-Tuning (PEFT) (Lialin et al., 2023), we develop

a computationally sustainable approach to domain specialization that preserves model performance even in the face of data scarcity.

3.1 Corpus Curation and Preprocessing

Data Source: To establish a robust factual baseline, we utilized the science curricula from the National Council of Educational Research and Training (NCERT) for grades 9–12, which serves as the authoritative "gold standard" for scientific communication in Hindi at the school level. These materials are meticulously curated through a multi-tiered review process involving academic experts and pedagogical researchers, ensuring that the technical terminology remains linguistically precise and conceptually accurate. By spanning approximately 5,000 pages of physics, chemistry, and biology, this corpus provides the dense, high-stakes vocabulary necessary to evaluate how models handle nuanced scientific reasoning in a primary Indic language.

Text Extraction & Tokenization: To ensure our dataset maintained the highest linguistic integrity, we employed a high-fidelity OCR pipeline specifically optimized for the structural nuances of the Devanagari script, such as its complex ligatures and vowel signs (matras). This was followed by a sophisticated tokenization process using the IndicNLP library (Kunchukuttan, 2020), which allowed us to move beyond simple whitespace splitting and respect the rich morphological boundaries of the Hindi language. By treating tokenization not just as a technical step but as a foundational linguistic task, we ensured that the resulting scientific corpora were primed for the deep semantic reasoning required in automated fact-verification.

3.2 Semi-Automatic Dataset Curation Pipeline

Our pipeline generates a three-label dataset (SUPPORT, CONTRADICT, NEI), following the standard set by FEVER (Thorne et al., 2018).

3.2.1 Stage 1: Generation of Supported Claims and Evidence

Claim Identification Model. We developed a specialized binary classifier from a **Llama 3.1 8B** model (Grattafiori et al., 2024) fine-tuned using LoRA (Hu et al., 2022). We merged two LoRA adapters: a domain adapter (ΔW_D) trained on SciARK (Fergadis et al., 2021) and a language adapter (Mirashi et al., 2024) (ΔW_L) trained on a Hindi corpus. The final weight update is a weighted composition:

$$W_{final} = W_0 + \gamma \Delta W_D + (1 - \gamma) \Delta W_L \quad (1)$$

where the merging coefficient $\alpha = 0.8$.

Evidence Retrieval and Re-ranking. For each claim, we implemented a dual-stage retrieval pipeline that balances computational efficiency with deep semantic understanding by first mapping assertions and candidate sentences into a high-dimensional vector space. To prune the vast NCERT search space, we utilized cosine similarity to identify a manageable pool of the 120 most relevant matches based on their vector embeddings ($\mathbf{e}_c$) and each sentence ($\mathbf{e}_s$):

$$\text{sim}(c, s) = \frac{\mathbf{e}_c \cdot \mathbf{e}_s}{\|\mathbf{e}_c\| \|\mathbf{e}_s\|} \quad (2)$$

This initial selection was then subjected to a rigorous re-ranking phase using a **Llama 3.3 70B bnb4bit** model, which functioned as a high-precision semantic judge to evaluate the nuanced logical relationship between the claim and the retrieved evidence.

3.2.2 Stage 2 & 3: Counterfactual Claim Generation

To generate challenging instances (Kaushik et al., 2020) we use the in-context learning (ICL) ability of the **Llama 3.3 70B bnb4bit** model. We employ a three-stage pipeline to ensure that generated claims require deep semantic reasoning rather than simple pattern matching.

Generation of Refuted Claims. To create **CONTRADICT** instances, we utilize 4-shot prompting to apply controlled semantic perturbations (Wright et al., 2022). This includes *Qualitative Inference Swaps* (reversing trends), *Relationship Inversions* (swapping subject/object roles), and *Quantitative Modifications*. Crucially, we enforce a "balanced language" constraint (Jansen et al., 2025) that forbids explicit negation markers like "not" or "never," ensuring the claims remain indistinguishable from valid assertions without evidentiary context.

Generation of Neutral (NEI) Claims. For **NEI** instances, we implement a novel **Evidence Mismatching** technique. We first use a BERT-based dense retriever to fetch topically related but non-verifying evidence passages. These are paired with

Attribute	FEVER	SciFact	SciFact-OPEN	SciClaimHunt	HindiSciVerify (Ours)
Language	English	English	English	English	**Hindi**
Domain	General	Scientific	Scientific	Scientific	**Scientific (Pedagogical)**
Task	3-way Class.	3-way Class.	3-way Class.	2-way Class.	**3-way Class.**
Total Claims	185k	1.1k	234	108k	**24k**
Evidence Source	Sentences	Abstracts	Full Papers	Full Papers	**Full Chapters**

Table 1: A quantitative comparison of HindiSciVerify with existing large-scale fact verification datasets.The proposed work introduces the first large-scale benchmark for this task in Hindi, uniquely sourced from pedagogical textbook chapters.

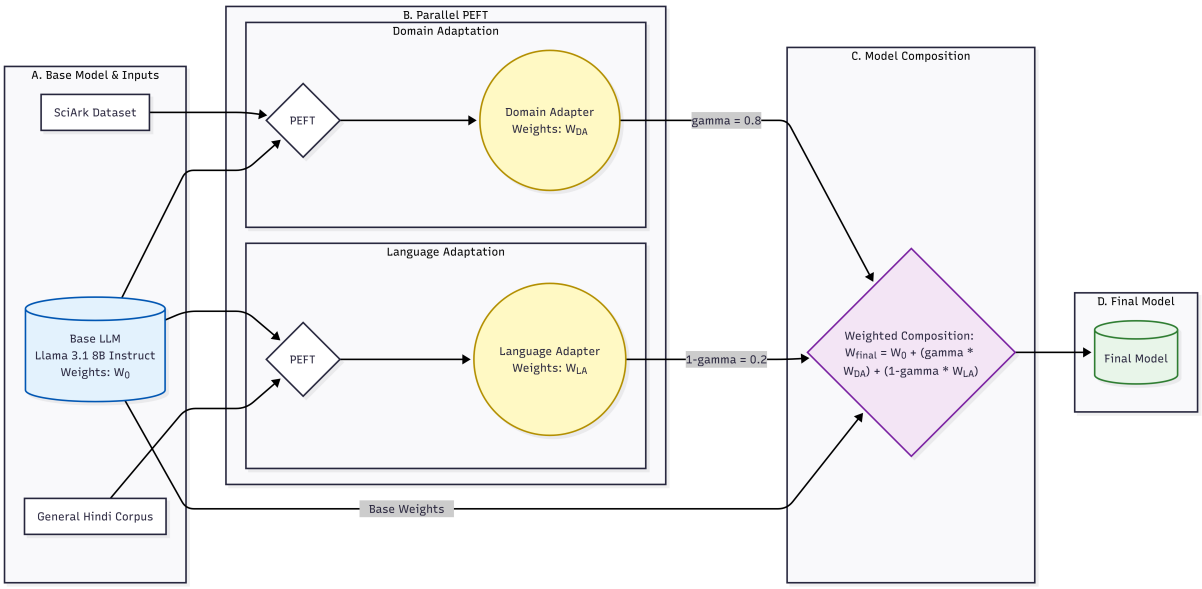

Figure 2: The LoRA adapter merging process.

the claim and refined using a 4-shot ICL prompt to ensure the evidence is contextually relevant yet provides "Not Enough Information" for verification. This rigorous filtering step ensures the "objectivity" of the dataset remains intact preventing label noise that often arises from automated retrieval methods

Consistency Verification. To minimize label noise, all generated pairs undergo a final verification loop. A high-parameter LLM **Llama 3.3 70B bnb4bit** acts as a "judge" (Ortega and Gomez-Perez, 2025), retaining only those instances where the predicted label aligns with the intended category (*Refute* or *NEI*) with high confidence.

3.3 Final Dataset Composition

Our final dataset, **HindiSciVerify**, consists of 24,000 instances, balanced across the three labels.

4 Experiments

To rigorously evaluate our dataset, we established in-domain baselines using **mBERT** (Devlin et al.,

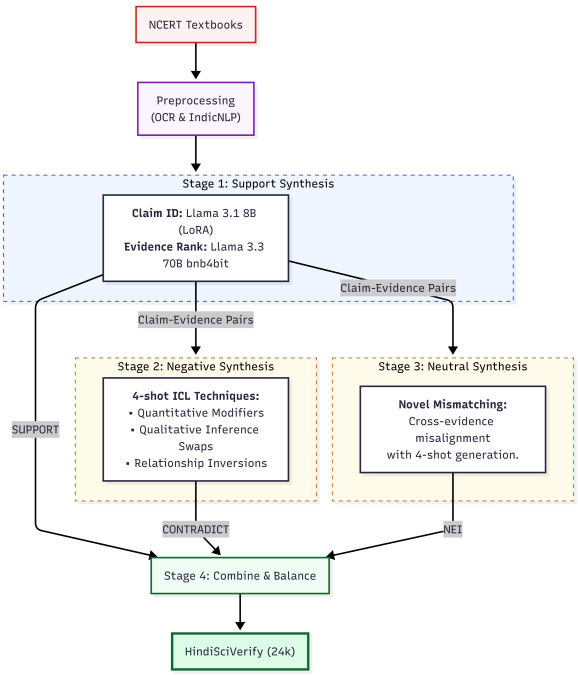

Figure 3: The semi-automatic data curation pipeline for HindiSciVerify.

2019), **XLM-R-Large** (Conneau et al., 2020), and **OLMo-7B-MoE** (Muennighoff et al., 2025). We also analyzed model generalization. Full implementation details and our generalization analysis can be found in Appendix A.2 and A.4, respectively.

5 Results and Discussion

5.1 Main Results

Table 2 summarizes in-domain performance. As hypothesized, the OLMo-7B-MoE model achieves the highest F1-score, suggesting the MoE architecture is well-suited for the task.

Model	Acc	Prec	Rec	F1
mBERT-base	73.4	74.8	72.6	73.7
XLM-RoBERTa-Large	80.6	80.9	80.2	80.5
OLMo-7B-MoE	**88.6**	**88.7**	**88.2**	**88.4**

Table 2: In-domain performance on the HindiSciVerify test set.

5.2 Ablation Study Results

Table 3 validates our methodological choices. The merged adapter model for claim identification significantly outperforms single-adapter versions. Removing the evidence re-ranking step causes a noticeable performance drop, confirming its importance.

Configuration	F1-score
Adapter Merging Study	
Domain Adapter Only	77.2
Language Adapter Only	81.2
Merged Adapters (Our Method)	**88.1**
Evidence Re-ranking Study	
No Re-ranking (Top-k Similarity)	74.6
With Re-ranking (Our Method)	**88.4**

Table 3: Results of the ablation studies on final task performance.

5.3 Generalization and Error Analysis

Our cross-lingual experiments (Appendix A.4) confirm the difficulty of zero-shot transfer. A quantitative error analysis (Table 4) reveals that Numerical and Multi-Sentence Reasoning are the most frequent failure modes. A qualitative example is in Appendix A.5.

5.4 Cross-Lingual Generalization

To test, whether our model learned the abstract task of scientific verification, we evaluated our best Hindi-tuned model on several English datasets. As shown in Table 5, our model experiences a sharp drop in performance when tested on external benchmarks like *SciFact* and *SciFact-Open*. This decline is likely due to the significant domain shift between our NCERT-based training data, which prioritizes pedagogical clarity and established consensus, and the dense, specialized nature of research abstracts. While the former focuses on clear, structured knowledge, datasets derived from sources like *S2ORC* involve a level of lexical density and niche terminology that falls outside standard educational curricula.

Error Category (%)	Description
Numerical (28%)	Model struggles with claims involving quantities or percentages.
Multi-Sentence (35%)	Correct label requires synthesizing info across multiple evidence sentences.
Other (37%)	Negation, semantic ambiguity, etc.

Table 4: A breakdown of primary error types.

Our findings align with Wadden et al. (2022a), who observed F1 scores dropping by 15 points when moving models from controlled environments to open-domain research literature. Professional academic text is often characterized by "hedged" language and granular specificity mismatches—where evidence is far more detailed than the claim itself. Consequently, a model trained on the structural regularity of textbooks finds it difficult to navigate the more complex and often noisy discourse patterns found in raw scientific publications.

Evaluation Dataset	Accuracy	F1-Score
SciFact (English)	65.7	64.1
SciFact-OPEN (English)	62.1	60.5
SciClaimHunt (English)†	68.2	67.5

Table 5: Zero-shot performance of our Hindi-tuned OLMo-MoE model on English datasets. †For 2-label evaluation, any 'NEI' prediction was treated as incorrect.

6 Conclusion

Through this work, we present **HindiSciVerify**, the first large-scale dataset for scientific claim ver-

ification in Hindi. We present a novel data curation pipeline that leverages modern LLMs and demonstrate through extensive experiments the specific difficulties that arise when verifying technical claims in a morphologically rich language like Hindi. We hope this dataset serves as a practical benchmark and can be extended to a scalable methodology for creating complex NLP datasets in other under-resourced indic languages.

Ethics Statement

The corpus for this work was derived from publicly available science textbooks published by the National Council of Educational Research and Training (NCERT).Since the data is drawn entirely from official curricular material it does not involve any personal information or privacy sensitive content. The resulting **HindiSciVerify** dataset will be released under the Creative Commons(CC) license.

Limitations

We acknowledge that our dataset is sourced exclusively from the pedagogical domain. Although it provides a structured and reliable starting point, it does not yet cover more specialized domains, such as medical literature or peer-reviewed scientific publications, which often involve greater technical depth, nuanced keywords, and a higher form of abstraction. Future work could expand to more complex sources, such as research articles and incorporate multi-modal evidence like diagrams and charts. Furthermore, our pipeline's application to other low-resource languages remains an area for future exploration.

References

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu and 1 others. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Md Saiful Islam and 1 others. 2020. COVID-19 related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621.

Peter Jansen, Samiah Hassan, and Ruoyao Wang. 2025. Matter-of-fact: A benchmark for verifying the feasibility of literature-supported claims in materials science. *Preprint*, arXiv:2506.04410.

Vladimir Karpukhin and 1 others. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

N J Karthika, Maharaj Brahma, Rohit Saluja, Ganesh Ramakrishnan, and Maunendra Sankar Desarkar. 2025. Multilingual tokenization through the lens of indian languages: Challenges and insights. *Preprint*, arXiv:2506.17789.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. *Preprint*, arXiv:1909.12434.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. Sciclaimhunt: A large dataset for evidence-based scientific claim verification. *Preprint*, arXiv:2502.10003.

Anoop Kunchukuttan. 2020. The IndicNLP library. Available at `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Preprint*, arXiv:2404.11553.

Vladislav Lialin, Omkar Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2023. Multi-hop evidence retrieval for cross-document relation extraction. *Preprint*, arXiv:2212.10786.

Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. 2024. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. *Preprint*, arXiv:2401.02254.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, and 5 others. 2025. OLMoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*.

Manali Naik and V. Srinivasa Chakravarthy. 2016. A comparative study of complexity of handwritten bharati characters with that of major indian scripts. *Preprint*, arXiv:1609.09227.

Raúl Ortega and Jose Manuel Gomez-Perez. 2025. SciClaims: An end-to-end generative system for biomedical claim analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 141–154, Suzhou, China. Association for Computational Linguistics.

Pritam Pal, Shyamal Krishna Jana, and Dipankar Das. 2025. IndicClaimBuster: A multilingual claim verification dataset. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2478–2489, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Jonas Pfeiffer and 1 others. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the EACL*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1322–1331.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

David Wadden, Shufan Lin, Kyle Lo, Iz Beltagy, 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)*.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022a. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

David Wadden and 1 others. 2022b. SciFact-Open: An open-domain dataset for scientific claim verification. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

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.

A Appendix

A.1 Dataset Statistics

Table 6 provides a detailed breakdown of the HindiSciVerify dataset composition.

Split	Support	Contradict	NEI	Total
Train	6,400	6,400	6,400	19,200
Dev	800	800	800	2,400
Test	800	800	800	2,400
Total	**8,000**	**8,000**	**8,000**	**24,000**

Table 6: Distribution of labels across the Train, Dev, and Test splits.

A.2 Implementation Details

Experiments were conducted using Hugging Face Transformers. Models were fine-tuned on a single NVIDIA RTX 6000 Ada Generation GPU with

48GB of VRAM. We used the AdamW optimizer with a linear learning rate schedule. Hyperparameters are in Table 7.

Hyperparameter	mBERT	XLM-R	OLMo-7B-MoE
Learning Rate	2e-5	2e-5	1e-5
Batch Size	16	16	4
Epochs	20	10	3
Optimizer	AdamW	AdamW	AdamW

Table 7: Hyperparameters used for fine-tuning the baseline models.

A.3 Details on Text Extraction and Tokenization

OCR Methods: Converting the scanned NCERT textbook pages into machine-readable(digital) text was a highly important first step in the whole pipeline. We considered using Tesseract OCR, due to its maturity and broad language support. However, our experiments revealed that its performance on the specific Devanagari fonts and page layouts used in the NCERT books was highly inconsistent. We therefore went ahead for a modern open-source OCR method (Paddle OCR) that provided greater accuracy and robustness in recognizing and correctly segmenting the Hindi text from our specific corpus, ensuring higher quality input for the subsequent downstream processing stages.

Hindi Tokenization: Tokenizing Hindi text is non-trivial due to its rich morphology and complex word formations (e.g., sandhi). Standard whitespace-based tokenizers are often insufficient. To address this, we utilized the **IndicNLP** library (Kunchukuttan, 2020), a specialized toolkit for Indian languages. This library provided us with a robust and linguistically-aware tokenizer that correctly handled the nuances of Devanagari(hindi) script, which was essential for generating clean sentence and word boundaries for our claim and evidence extraction tasks.

A.4 Generalization Analysis

Objective: Analyze model generalization across linguistic boundaries.

A.4.1 Zero-Shot Transfer: English-to-Hindi

We evaluate pre-trained mBERT ,XLM-R and OLMOE-7B directly on the **HindiSciVerify** test set to test their inherent cross-lingual transfer capabilities.

A.4.2 Exploratory Study: Zero-Shot Transfer from Hindi-to-English

We take our best model fine-tuned only on **HindiSciVerify** and evaluate it on the English SciFact test set to probe for any abstract, language-agnostic task learning.

A.5 Qualitative Error Analysis

Qualitative analysis reveals that model failures stem primarily from a lack of semantic grounding in some examples, where the system relies on word matching rather than logical reasoning. As shown in Table 8, the model struggles with cross-unit numerical mappings, fails to synthesize facts split across sentence boundaries, and frequently overlooks critical negation markers like nahi(negation).

Category	Content
Numerical	**Claim:** जल का क्वथनांक 100°C होता है। (Water boils at 100°C.) **Evidence:** समुद्र तल पर, पानी 212°F पर उबलता है। (At sea level, water boils at 212°F.) **Model/Gold:** CONTRADICT / SUPPORT
Multi-Sentence	**Claim:** क्रिया और प्रतिक्रिया बल एक ही वस्तु पर कार्य करते हैं। (Action and reaction forces act on the same object.) **Evidence:** [1] प्रत्येक क्रिया की समान और विपरीत प्रतिक्रिया होती है। [2] ये बल हमेशा दो अलग-अलग वस्तुओं पर कार्य करते हैं। ([1] Every action has an equal/opposite reaction. [2] These forces act on different objects.) **Model/Gold:** SUPPORT / CONTRADICT
Negation	**Claim:** लोहा का उपयोग अर्धचालक बनाने में होता है। (Iron is used to make semiconductors.) **Evidence:** सिलिकॉन एक प्रमुख अर्धचालक है; हालाँकि गैलियम एक विकल्प है, लेकिन लोहा नहीं। (Silicon is a major semiconductor; while Gallium is an option, Iron is not.) **Model/Gold:** NEUTRAL / CONTRADICT

Table 8: Representative failure cases in the HindiSciVerify dataset, highlighting specific challenges in technical Hindi reasoning.

A.6 Performance on Counterfactual Subsets

To further analyze model performance, we evaluated our best model (OLMo-7B-MoE) on subsets of the test set, grouped by the generation method. As shown in Table 9, the model performs best on claims created via Entity Swapping (Relationship inversions) but struggles with Causal Reversal (Qualitative swapping), indicating that our techniques successfully create distinct levels of reasoning difficulty.

Generation Method	Acc.	F1
Numerical Perturbation *(Quantitative swapping)*	86.1	86.7
Entity Swapping *(Relationship inversions)*	90.3	90.0
Causal Reversal *(Qualitative swapping)*	83.2	83.1

Table 9: Model performance on subsets of contradictory claims.

B Rationale for Native Data Sourcing

To ensure the ecological validity and rigorousness of our Hindi scientific claim verification benchmark, we elected to construct a native dataset sourced from NCERT textbooks rather than relying on machine-translated versions of existing English benchmarks (e.g., SciFact). This decision is grounded in three primary methodological concerns regarding cross-lingual transfer in specialized domains:

1. Mitigation of "Translationese" Artifacts Neural models often exploit "translationese"—unnatural phrasing and syntactic interference from source languages—as spurious cues (Wadden et al., 2022a). By curating native Hindi text, we ensure our model learns from authentic scientific discourse rather than overfitting to the predictable idiosyncrasies of a translation engine.

2. Preventing Semantic Drift Automated translation frequently suffers from "semantic drift," mapping precise technical terms to generic equivalents and introducing label noise. Using native NCERT data prevents this "specificity mismatch" (Wadden et al., 2022b), ensuring claims remain semantically aligned with their evidence through standardized terminology like *urja grah*.

3. Expert-Curated Gold Standards NCERT textbooks provide a pre-validated corpus that avoids the "domain shift" seen in synthetic datasets (Jansen et al., 2025). This grounding ensures that terminology and factual density reflect the standardized scientific knowledge expected of our target demographic.

Experimental Setup for Cross-Lingual Baselines To measure the actual impact of native training data, we designed a controlled experiment that isolates the source of our data (Native Hindi vs. Machine-Translated English) to measure the "domain gap." This setup allows us to empirically test how translation artifacts and specificity mismatches affect verification performance.

Model Architecture Selection We chose two distinct architectures to ensure our findings hold across different modeling paradigms:

- **Llama-3-8B (Generative):** We used **Llama-3-8B Instruct** as our primary verifier. As noted by Ortega and Gomez-Perez (2025), it achieves high performance on specialized claim tasks while remaining efficient enough for single-GPU deployment.
- **MultiVerS (Discriminative):** We included **MultiVerS** (Wadden et al., 2022a) to align our work with established benchmarks. It is specifically designed to handle long-range dependencies in scientific abstracts (Wadden et al., 2022b).

Translation and Semantic Drift Analysis To see if English resources can stand in for native data, we used a "Translate-Train" approach. We translated the **SciFact** (Wadden et al., 2020) and **SciClaimHunt** (Kumar et al., 2025) datasets into Hindi using a state-of-the-art NMT system (**Llama 3.3 70B bnb4bit**). This helps us quantify *specificity mismatch*—where automated translation might turn precise scientific terms from research papers into generic equivalents, failing to capture the pedagogical accuracy of NCERT textbooks.

B.1 Human Annotation and Quality Assessment

To assess the quality of our dataset, we performed a human annotation study. Four Hindi-proficient annotators with computer science backgrounds evaluated a balanced, random sample of 500 claim-evidence pairs on four criteria (Wright et al., 2022): **Fluency**, **Atomicity**, **De-Contextualization** and

Case Study: Semantic Drift in Machine Translation

Source (English):
Claim: The external force on a body always arises due to some other body.
Evidence: The earth pulls a stone downwards due to gravity. The answer according to Newton is: Yes, the stone does exert an **equal and opposite force** on the earth. We do not notice it since the earth is very massive.

Target (Native Hindi / NCERT - Gold Standard):
Claim: किसी **पिण्ड (Object/Body)** पर लगने वाला बाह्य बल सदैव ही किसी अन्य **पिण्ड** के कारण होता है।
Evidence: पृथ्वी गुरुत्वीय बल के कारण किसी पत्थर को अधोमुखी दिशा में खींचती है | परंतु न्यूटन के अनुसार: हाँ, पत्थर भी पृथ्वी पर परिमाण में समान **(Equal in Magnitude)** तथा दिशा में विपरीत **(Opposite in Direction)** बल **(Force)** लगाता है|

Target (Machine Translated - Baseline):
Claim: किसी **शरीर (Biological Body)** पर बाहरी **ताकत (Strength)** हमेशा किसी दूसरे **शरीर** के कारण उत्पन्न होती है।
Evidence: पृथ्वी गुरुत्वाकर्षण के कारण पत्थर को नीचे की ओर खींचती है |न्यूटन के अनुसार उत्तर है | हाँ, पत्थर पृथ्वी पर एक समान **(Same)** और उल्टा **(Reverse)** जोर **(Effort/Push)** डालता है

Figure 4: Example of **Specificity Mismatch** introduced by machine translation. The Native text uses the precise physics term *Pind* (Object) and *Bal* (Force). The translated text drifts into biological/colloquial domains, using *Shareer* (Human Body) and *Taakat* (Physical Strength), rendering the claim unverifiable against scientific evidence.

Model & Data Source	P	R	F1	Δ F1
Llama 3-8B (Generative)				
Native Hindi (NCERT)	**85.8**	**85.6**	**85.7**	–
Translated SciFact†	68.3	69.4	68.8	-19.7%
Translated SciClaimHunt‡	72.1	73.2	72.7	-15.2%
MultiVerS (Discriminative)				
Native Hindi (NCERT)	78.4	76.2	77.3	–
Translated SciFact†	62.5	60.1	61.3	-20.7%
Translated SciClaimHunt‡	65.8	63.4	64.6	-16.4%

Table 10: Cross-architecture comparison on the Hindi Test Set. †Wadden et al. (2020); ‡Kumar et al. (2025). Across both paradigms, native training significantly outperforms translated baselines.

Scientific Concept	Native Hindi (Gold)	Translated (Error)	Failure Mode
Body (Physics)	**पिण्ड** (*Pind*) *(Inanimate Object)*	शरीर (*Shareer*) *(Biological Body)*	**Domain Shift** (Physics → Biology)
Force ($F = ma$)	**बल** (*Bal*) *(Vector Quantity)*	ताकत (*Taakat*) *(Physical Strength)*	**Specificity Mismatch** (Scientific → Colloquial)

Table 11: Qualitative analysis of translation artifacts. The translated model fails to retrieve evidence because it searches for colloquial terms (e.g., *Strength*) rather than the specific scientific nomenclature (e.g., *Force*) present in the corpus.

Faithfulness. As shown in Table 12, the inter-annotator agreement scores fall within the "substantial agreement" strong outcome for this subjective task. The high agreement on Atomicity (Fleiss' $\kappa = 0.82$) and De-Contextualization (Krippendorff's $\alpha = 0.78$) confirms our pipeline generates clear, self-contained claims, while the substantial agreement on Faithfulness validates the dataset's reliability for complex reasoning.

Metric	Fluency	Atomicity	De-Context.	Faithfulness
Krippendorff's α	0.70	0.70	0.78	0.69
Fleiss' Kappa	0.70	0.82	0.76	0.69

Table 12: Inter-annotator agreement scores.

B.2 LoRA Adapter Merging Methodology

To create a highly specialized claim identification model, we employed a LoRA adapter merging technique. We began with the base weights of the Llama 3.1 8B model (W_0) and independently fine-tuned two LoRA adapters: a domain adapter (ΔW_D) trained on the SciARK dataset (Fergadis et al., 2021) to learn the structure of scientific claims, and a language adapter (ΔW_L) trained on a general Hindi corpus (Mirashi et al., 2024) to improve linguistic fluency. We used standard LoRA configurations for fine-tuning, with a rank (`r`) of 128 for the domain adapter and 64 for the language adapter. The final model weights were then composed by taking a weighted sum of these adapters and adding them to the base model, as shown in Equation 1.The rationale behind PEFT merging is to combine complementary capabilities—domain-specific understanding and linguistic proficiency—while retaining the efficiency of parameter-efficient fine-tuning, reducing training cost and preserving reusable modularity, so different language or domain adapters can be flexibly swapped or recombined when needed.

B.3 Attention Visualization for Model Interpretability

The following figures demonstrate our model's attention mechanism across different scenarios, providing insight into its reasoning process for each provided claim and evidence pair.

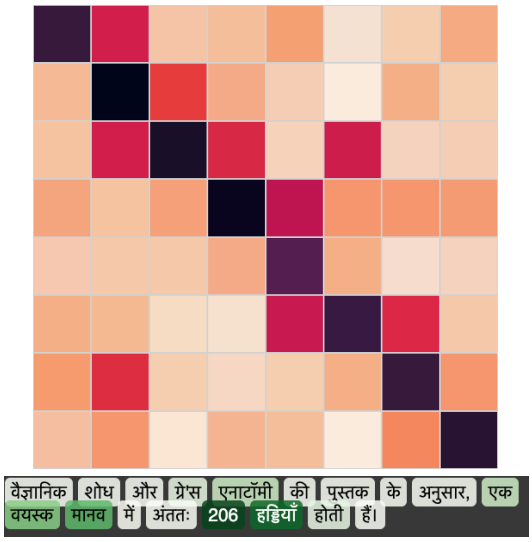

Figure 5: **SUPPORT Scenario:** For the claim "मानव शरीर में 206 हड्डियाँ होती हैं।" (The human body has 206 bones.), the model correctly places high attention on the phrase "206 हड्डियाँ" (206 bones) in the evidence, demonstrating successful grounding on the key numerical fact for verification.

Figure 6: **CONTRADICT Scenario:** When presented with the claim "मैरी क्यूरी ने गणित में नोबेल पुरस्कार जीता।" (Marie Curie won the Nobel Prize in mathematics.), the model focuses on the semantic conflict between "गणित" (mathematics) in the claim and the subjects "भौतिकी" (physics) and "रसायन विज्ञान" (chemistry) in the evidence, which is the correct basis for a refutation.

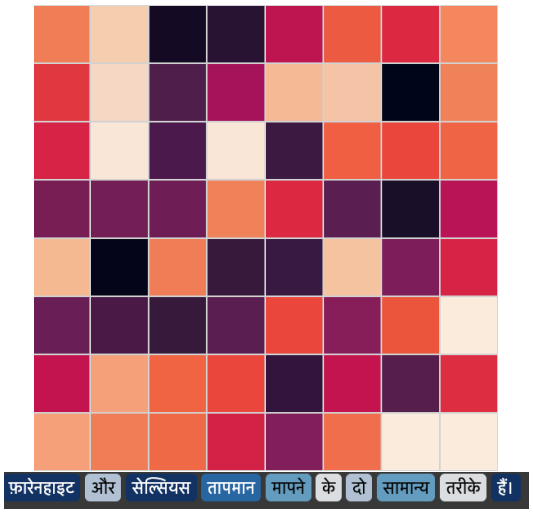

Figure 7: **NEI Scenario:** For the claim that Celsius and Fahrenheit are equal at -40, the evidence is only topically related. The heatmap shows diffuse attention, correctly indicating that the evidence lacks any specific, verifiable information (like the number -40) to either support or refute the claim.

C Prompting Methodology

This section details the structure of the 4-shot prompts used for generating counterfactual claims.

C.1 Prompt for Contradictory and Neutral Claims

To ensure the dataset challenges the boundaries of automated reasoning, we utilized the Llama 3.3 70B model to synthesize sophisticated `CONTRADICT` and `NEUTRAL` (NEI) instances that move beyond simple lexical overlap. While contradictions were crafted to present scientifically plausible falsehoods that require deep domain knowledge to debunk, the NEI (Not Enough Information) instances were engineered to be topically relevant yet factually "unsupported" by the specific local evidence. This dual-pronged strategy ensures that the model is evaluated not just on its ability to spot errors, but on its capacity to recognize the inherent limits of the provided evidentiary scope without falling into the trap of hallucination.

Prompt Template: Generation of Refuted (Negative) Claims

System Instruction: You are an expert scientific annotator specializing in generating high-quality "Refuted" (False) claims in Hindi. Your task is to generate a scientifically plausible but factually incorrect claim based on a given valid sentence.

You must apply one of the following ”Perturbation Strategies” from the MuSciClaims benchmark to create the falsehood:

1. **Quantitative Modification:** Alter specific numerical values, dates, or statistical figures (e.g., change ”206 bones” to ”306 bones”).

2. **Qualitative Inference Swap:** Reverse the direction of a trend or effect (e.g., change ”increases” to ”decreases” / ”badhta hai” to ”ghatta hai”).

3. **Relationship Inversion:** Swap the relationship between two entities (e.g., change ”A is stronger than B” to ”B is stronger than A”).

CONSTRAINT (Balanced Language): Do NOT simply add negation words like ”nahi” (not), ”never”, or ”false” to the sentence. The negative claim must sound like a natural, positive assertion that happens to be factually wrong.

Few-Shot Examples:
Example 1 (Quantitative Modification):
Sentence: मानव शरीर में 206 हड्डियाँ होती हैं।
Evidence: वयस्क मानव कंकाल प्रणाली 206 हड्डियों से बनी होती है। हालाँकि जन्म के समय शिशुओं में लगभग 300 हड्डियाँ होती हैं, लेकिन विकास के दौरान कई हड्डियाँ आपस में जुड़ जाती हैं
NegativeSentence: मानव शरीर में 306 हड्डियाँ होती हैं।

Example 2 (Entity/Relationship Swap):
Sentence: प्रकाश संश्लेषण के लिए कार्बन डाइऑक्साइड, पानी और सूर्य के प्रकाश की आवश्यकता होती है।
Evidence: प्रकाश संश्लेषण वह प्रक्रिया है जिसमें हरे पौधे सूर्य के प्रकाश की उपस्थिति में कार्बन डाइऑक्साइड और पानी का उपयोग करके अपना भोजन (ग्लूकोज) बनाते हैं और ऑक्सीजन छोड़ते हैं।
NegativeSentence: प्रकाश संश्लेषण के लिए ऑक्सीजन, पानी और सूर्य के प्रकाश की आवश्यकता होती है।

Example 3 (Relationship Inversion):
Sentence: पृथ्वी सूर्य के चारों ओर घूमती है।
Evidence: सौर मंडल में सूर्य केंद्र में स्थित है और पृथ्वी सहित सभी ग्रह अण्डाकार कक्षाओं में सूर्य की परिक्रमा करते हैं। पृथ्वी को एक चक्कर पूरा करने में लगभग 365 दिन लगते हैं।
NegativeSentence: सूर्य पृथ्वी के चारों ओर घूमता है।

Example 4 (Qualitative Inference Swap):
Sentence: व्यायाम करने से हृदय गति बढ़ती है।
Evidence: शारीरिक गतिविधि के दौरान, मांसपेशियों को अधिक ऑक्सीजन की आवश्यकता होती है। इस मांग को पूरा करने के लिए, हृदय तेजी से पंप करना शुरू कर देता है, जिससे हृदय गति बढ़ जाती है।
NegativeSentence: व्यायाम करने से हृदय गति कम हो जाती है।

Example 5 (Entity Mutation):
Sentence: पानी दो हाइड्रोजन परमाणुओं और एक ऑक्सीजन परमाणु (H_2O) से बना है।
Evidence: पानी का एक अणु दो हाइड्रोजन परमाणुओं और एक ऑक्सीजन परमाणु के सहसंयोजक बंधन से बनता है। इसका रासायनिक सूत्र (□₂□
NegativeSentence: पानी दो हाइड्रोजन परमाणुओं और दो ऑक्सीजन परमाणुओं (H_2O_2) से बना है।

Final Task:
Sentence: {datapoint.claim}
Evidence: {datapoint.evidence}
NegativeSentence:

Prompt Template: Generation of Neutral (Hard Negative) Claims

System Instruction: You are an expert in scientific claim verification. Your task is to generate ”Neutral” (Not Enough Information) claims in Hindi.
A Neutral claim must satisfy the ”Hard Negative” criteria:

1. **Topical Similarity:** It must share the same key entities and scientific vocabulary as the Positive Claim (it should look related).

2. **No Logical Link:** It must be impossible to prove or disprove the claim using *only* the provided Evidence.

3. **Factuality:** The statement itself should be scientifically correct in the real world, but *irrelevant* to the specific evidence provided.

Few-Shot Examples:
Example 1 (History vs Theory):
Positive Claim: अल्बर्ट आइंस्टाइन ने सापेक्षता का सिद्धांत विकसित किया था।
Evidence: अल्बर्ट आइंस्टाइन ने 1905 में विशेष सापेक्षता का सिद्धांत प्रकाशित किया जो समय और स्थान के संबंध को बताता है।
Neutral Claim: अल्बर्ट आइंस्टाइन को 1921 में भौतिकी के नोबेल पुरस्कार से सम्मानित किया गया था।
(Reasoning: Same entity [Einstein], true fact, but Evidence discusses ”Relativity,” not ”Nobel Prize.”)

Example 2 (Discovery vs. Usage):
Positive Claim: मैरी क्यूरी ने रेडियम और पोलोनियम की खोज की।
Evidence: मैरी क्यूरी और उनके पति पियरे क्यूरी ने 1898 में यूरेनियम अयस्क से रेडियम और पोलोनियम को अलग किया।
Neutral Claim: रेडियम का उपयोग घड़ी की सुइयों को चमकदार बनाने के लिए किया जाता था।
(Reasoning: Same entity [Radium], but Evidence discusses ”Discovery,” not ”Usage.”)

Example 3 (Structure vs. Composition):
Positive Claim: मानव शरीर में 206 हड्डियाँ होती हैं जो शरीर को संरचनात्मक समर्थन प्रदान करती हैं।
Evidence: वयस्क मानव शरीर में आमतौर पर 206 हड्डियाँ होती हैं, जबकि नवजात शिशुओं में लगभग 300 हड्डियाँ होती हैं जो बाद में विकास के दौरान आपस में जुड़ जाती हैं।
Neutral Claim: मानव कंकाल प्रणाली में हड्डियों के अलावा कार्टिलेज, लिगामेंट्स और टेंडन्स जैसे ऊतक भी शामिल हैं।
(Reasoning: Same topic [Skeleton], true fact, but the

Evidence discusses ”Bone Count,” not ”Other Tissues” like cartilage.)

Example 4 (State Change vs. General Effect):
Positive Claim: पानी तीन अवस्थाओं में मौजूद होता है: ठोस, द्रव और गैस।
Evidence: वैज्ञानिक प्रमाण यह स्थापित करते हैं कि पानी तापमान के आधार पर अपनी अवस्था बदलता है। 0°C से नीचे यह ठोस (बर्फ) होता है और 100°C से ऊपर यह गैस (भाप) बन जाता है।
Neutral Claim: तापमान में परिवर्तन पानी के स्वरूप को प्रभावित कर सकता है।
(Reasoning: Closely related topic, but the Evidence lists specific ”State Thresholds” (0°C/100°C), while the Neutral claim makes a vague general statement not explicitly proven by the specific figures in the evidence.)

Final Task:
Positive_Claim: {datapoint.claim}
Evidence: {datapoint.evidence}
Neutral_Claim: