WIKISTAR: A System for Shedding Light on the Hidden History of Scientific Wikipedia Articles

Omer Ehrlich[1*] **Nitzan Barzilay**[1*] **Rona Aviram**[3] **Tom Hope**[1,2]
[1] The Hebrew University of Jerusalem [2] Allen Institute for AI (Ai2)
[3] Department of Communication Studies, Ben-Gurion University of the Negev
omer.ehrlich@mail.huji.ac.il, nitzan.barzilay@mail.huji.ac.il

Abstract

Wikipedia plays a key role in shaping public understanding of science, and its openly accessible revision history is a unique record of how scientific knowledge evolves over time. Yet scientifically meaningful revisions are obscured by the sheer volume of routine edits, leaving each article's scientific history hidden. We present WIKISTAR (**S**cientific **T**racking of **A**rticle **R**evisions), an interactive system for exploring scientifically meaningful changes across an article's revision history. Using an LLM classifier with an expert-designed multi-label taxonomy, WIKISTAR first tags edit types such as the addition of technical terms, new research findings, and changes in scientific narrative. Then, through interactive views, an article's full revision history can be traced at any granularity—from aggregate trends that reveal when and in which sections scientific content was added or refined, down to individual edits—showing how scientific knowledge develops at a scale previously impossible. In a user study, experts from three domains found that WIKISTAR surfaced new patterns and research questions and enabled previously impractical analyses. We release our system, code and a human-annotated benchmark.[1]

Figure 1: The WIKISTAR pipeline for tracking the history of scientific edits in a Wikipedia article. (1) Splitting each revision into sections and matching them across revisions to recover the previous version of each edited section. (2) Multi-label edit-type classification of each edit pair. (3) Exploring through interactive views, from aggregate patterns down to individual edits.

1 Introduction

Wikipedia, the world's largest online encyclopedia, is written and continuously rewritten by volunteers around the world. As a widely consulted source, it shapes how the public understands scientific and scholarly topics. Its openly accessible revision history lets scholars trace the development of ideas within and between academic domains, and examine how public knowledge evolves (Benjakob et al., 2023, 2021). Scientific articles on Wikipedia are especially well suited for studying knowledge evolution. Prior work has shown that they closely track developments in the scientific literature through extensive use of reliable academic sources, making them more than public-facing summaries (Simons et al., 2024; Benjakob et al., 2023). Their revision histories therefore potentially constitute a rich longitudinal record of how scientific concepts are introduced, debated, refined, and stabilized.

However, a single article can accumulate thousands of edits over its lifetime—the "AI" article alone has over 18,000 section edits—and scientifically significant edits make up only a small fraction,

*Equal contribution.
[1] Code 🤗 WIKISTAR-BENCH dataset 🤗 WIKISTAR demo

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 414–423
October 24-29, 2026 ©2026 Association for Computational Linguistics

buried among routine changes such as rephrasing and formatting. Isolating meaningful edits manually is highly labor-intensive, so scientifically significant edits remain hidden and inaccessible.

Prior work classifies Wikipedia edits (Yang et al., 2017; Rajagopal et al., 2022) and builds tools for exploring authorship and editorial conflicts (Viégas et al., 2004; Flöck and Acosta, 2014; Guo et al., 2023), but both lines target general-purpose editing at the whole-page or token level. We instead work at the section level and focus on scientific significance, which localizes each edit, reveals how a section's scientific content develops over time, and drives the interactive views of our system.

We present **WIKISTAR (Scientific Tracking of Article Revisions)**, a system that detects, quantifies, and contextualizes scientific textual changes in Wikipedia articles over time (Figure 1). We introduce the task of *scientific edit classification*: labeling section-level edits with the scientifically meaningful changes they make, using a taxonomy of ten expert-designed labels. WIKISTAR extracts section-level edits, classifies each edit against the taxonomy using an LLM, and presents the results through interactive views that trace how an article's scientific content evolves across time and sections (Figure 2). This makes the scientific edit history of Wikipedia accessible to Wikipedia researchers and more broadly to scientists, journalists, and others interested in how science is represented online.

In a user study, experts in Wikipedia editing, science journalism, and the history and philosophy of science found that WIKISTAR surfaced patterns they could not easily discover manually, prompted new research questions, and enabled analyses that were previously impractical.

Finally, we release WIKISTAR-BENCH, a human-annotated dataset of section-level Wikipedia edits labeled for scientific significance.

In summary, our main contributions:

(1) We create WIKISTAR, a system for exploring how a Wikipedia article's scientific content evolves over time. We validate WIKISTAR in a user study with domain experts from three domains, who found it highly useful.

(2) We introduce the task of scientific edit classification of Wikipedia article sections, grounded in our taxonomy of ten labels for scientifically salient changes, and release WIKISTAR-BENCH, a human-annotated dataset of 1,387 section edits spanning three domains.

2 Scientific Edit Classification

In this section, we first introduce the task of *scientific edit classification* of Wikipedia edits. To support this task, we release WIKISTAR-BENCH, a human-annotated benchmark, and use it to compare seven LLMs as classifiers.

As discussed in the Introduction, since scientifically significant edits are only a small fraction of an article's edits, we define *scientific edit classification* as labeling each Wikipedia edit with the types of scientific change it makes. We cast this as a multi-label classification task over section-level edits. Given a section and its preceding version, which we call a *section edit pair*, the task is to assign every applicable label from our taxonomy of ten scientific edit types (§2.2), or the *Non-Scientific Edit* label when no scientific change occurs.

We choose the section as our unit of granularity for two reasons. First, it carries the right amount of context for classification: whole revisions inflate prompt length and reduce output consistency (Shivashankar and Steinmetz (2024); Fieblinger et al. (2024)), while individual sentences discard the context needed to judge whether a change is scientifically meaningful. Second, sections localize each edit to a specific aspect of the article, letting WIKISTAR track how scientific content is distributed across sections and develops over time.

2.1 Section Identification and Linking

Turning an article's revision log into section-level edits requires two operations: parsing each revision into its constituent sections, and pairing each edited section with its previous version—or, for a newly created section, recognizing that it has none.

Acquiring section text: Using the MediaWiki API (Wikimedia Foundation, 2025), we retrieve every revision of the article as raw wikitext with its metadata (revision ID, timestamp, editor). Each revision is split into sections by a regex-based parser that follows Wikipedia's markup conventions. As each revision contains the article's full section list, most of it unchanged from the previous one, we retain only the sections that changed, so each entry corresponds to a single edited section.

Matching section edit pairs: The *scientific edit classification* task requires the version of each section before and after an edit, which we call a *section edit pair*. Matching versions is challenging because sections are not stable: editors rename, split and merge them, yet Wikipedia stores each revision

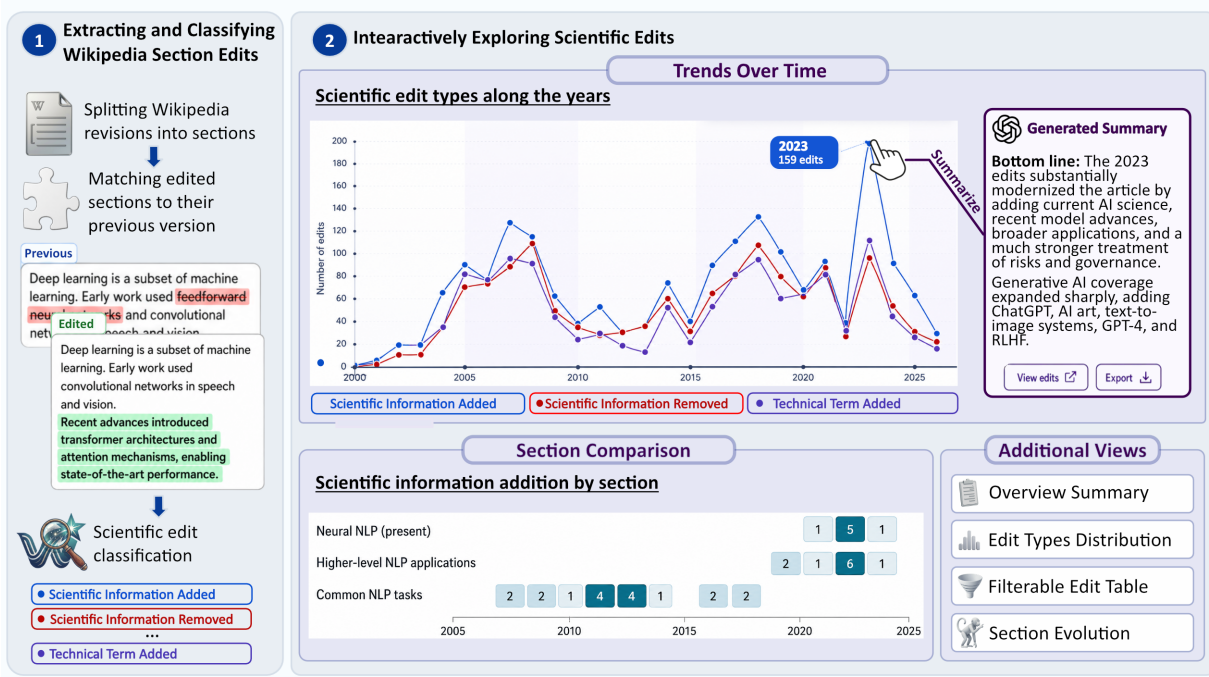

Figure 2: Overview of the WIKISTAR system. **(1) Extraction and classification:** The system splits each Wikipedia revision into sections, matches each edited section to its previous version, and classifies each edit pair into scientific edit types. **(2) Interactive exploration:** Interactive views let users explore the classified history: The *Over Time* view shows temporal trends in edit types, and the *Section Comparison* view compares them across selected sections. Selecting a point or cell within those views generates a summary of its edits, with access to the underlying edits.

as flat full-article text with no identifier linking section versions. To locate a section's earlier version, we search backward through revisions and compute two similarity measures against candidate sections: **title similarity** (normalized Levenshtein distance) and **content similarity** (cosine similarity over embeddings of the section content). We match the section to the most recent candidate for which either measure exceeds a threshold. This simple approach worked well in practice; more advanced matching is left to future work.

2.2 Scientific Edits Taxonomy

A closed set of labels is what lets WIKISTAR aggregate, filter, and visualize edits by type. We designed a taxonomy of ten scientifically meaningful edit types—such as *New Scientific Information* and *Change in Scientific Narrative*—plus a *Non-Scientific Edit* label, built by a scientific-Wikipedia expert and refined iteratively against real edits. Full definitions appear in Appendix A, with annotated examples for select types in Appendix B.

Some labels in the taxonomy may appear replaceable by rule-based detectors, but each requires judgment that rules do not capture. *Wikilink Added* could be approximated by detecting new square-bracketed text, but such a rule would also count existing links that were only fixed or reformatted. *Academic References Added* is similar: references share a common structure, but the citation format does not indicate whether the source is scientific.

2.3 WIKISTAR-BENCH

We curate WIKISTAR-BENCH, the first human-annotated dataset of scientifically labeled section-level Wikipedia edits, released as an evaluation resource for scientific edit classification. It contains 1,387 examples across three scientific domains—Biology (821), Computer Science (288), and Neuroscience (278)—each spanning at least 9 popular Wikipedia pages. Each example is a single section edit pair with its metadata and human-annotated gold labels from our taxonomy.

2.4 Benchmark construction

We first applied LLaMA 3.3 70B (Dubey et al., 2024) to all section-edit pairs extracted from the complete revision histories of 30 Wikipedia articles. From these model-labeled edits, we sampled examples for human annotation to obtain broad coverage across source articles and predicted edit types. The final gold-label supports are not uni-

form because model predictions can differ from human annotations and examples may receive multiple co-occurring labels. This sampling procedure intentionally enriches scientifically meaningful edits rather than preserving their natural prevalence. Non-Scientific Edit accounts for about 31% of WIKISTAR-BENCH, whereas most edits in complete article histories are non-scientific.

Annotation process: The WikiStar-Bench dataset was labeled by four annotators: three STEM graduate students and an assistant professor who holds a PhD in biology and researches scientific Wikipedia. Annotators labeled each section edit pair independently, assigning every applicable label from our 10-label taxonomy, or the *Non-Scientific Edit* label if none applied.

Annotator agreement: To ensure reliability of human annotation, all annotators tagged a shared subset of 104 examples. Annotator agreement using mean pairwise Cohen's κ (Cohen, 1960) is high ($\kappa = 0.89$), indicating strong overall agreement. Agreement is naturally lower on more subjective and interpretive labels (such as *Change in Scientific Narrative* with $\kappa = 0.7$ and *Scientific Clarification Added* with $\kappa = 0.8$).

2.5 Results

Experimental settings: Against WIKISTAR-BENCH's gold labels, we report per-label precision and recall, and macro- and micro-F1 (our primary metric). We evaluate seven LLMs as classifiers: GPT-5.4, GPT-5-mini, GPT-5.1 (OpenAI, 2025a), GPT-4o (Hurst et al., 2024), o3-mini (OpenAI, 2025b), LLaMA 3.3 70B (Dubey et al., 2024), and Qwen3-Next-80B (Yang et al., 2025). Scientific edits are rare in real article histories, so WIKISTAR-BENCH was built to over-sample them. Because they are also the harder labels to classify, the reported scores reflect a harder setting than a full article history.

Prompt refinements: Our final prompt opens with a general task explanation, then gives each label a definition and a single minimal example. The most effective refinements over a naive one-sentence-definition prompt were asking the model to first decide whether the edit is scientific at all, and flagging common mistakes in the definitions; these raise GPT-5.4's macro-F1 from 0.69 to 0.82. Few-shot demonstrations did not help (macro-F1 drops to 0.78). Appendix D reports per-model results across prompt versions.[2]

	GPT-5.4			Human
Label	P	R	F1	F1
New Sci. Information	.88	.74	.80	.93
Sci. Information Removed	.84	.92	.87	.94
Sci. Clarification Added	.73	.72	.72	.86
Technical Terms Added	.80	.93	.86	.93
Researcher Names Added	.93	.82	.87	.92
Change in Sci. Narrative	.54	.70	.61	.72
Academic References Added	.90	.86	.88	.97
Academic References Removed	.84	.86	.85	.95
Wikilink Added	.95	.83	.89	.97
Quantitative Information	.75	.86	.80	.95
Non-Scientific Edit	.83	.87	.85	.91
Macro avg	.82	.83	.82	.91
Micro avg	.84	.83	.83	.92

Table 1: Per-label precision, recall, and F1 for GPT-5.4 on WIKISTAR-BENCH (n=1,387), and the human agreement ceiling—mean pairwise F1 across annotators on overlapping examples (n=104).

Aggregate results: Of the seven models, GPT-5-mini and GPT-5.4 are the two best-performing, with comparable scores (macro-F1 0.83 vs. 0.82; micro-F1 0.86 vs. 0.83). The deployed demo uses GPT-5-mini due to its overall comparable results and lower operational cost. Results for all models are in Appendix C. Across labels, GPT-5.4 trails the human agreement ceiling (macro-F1 0.82 vs. 0.91), showing that scientific edit classification is not trivial and leaves room for improvement.

Per-Label results: Table 1 reports per-label results for GPT-5.4. F1 varies substantially by label: strongest on objective labels with explicit textual signals (such as *Wikilink Added* with 0.89, and *Academic References Added* with 0.88), where remaining errors mostly concern the scientific qualifier, e.g., counting a book as an academic reference. It is weakest on the interpretive labels (e.g., *Change in Scientific Narrative* at 0.61), where the model tends to over-extend the label beyond its intended scope, often misclassifying *Scientific Clarifications* as narrative changes. This tracks lower human agreement on those labels (Section 2.4).

Domain-Specific Results: We report GPT-5.4's per-label F1 for each domain in WIKISTAR-BENCH: Biology, Computer Science, and Neuroscience in Table 2. Macro-F1 is stable across all domains. As on the full benchmark, objective

[2]See prompt at `create_prompt.py` in our repository.

Edit Type	Biology 820 examples				Computer Science 288 examples				Neuroscience 278 examples			
	P	R	F1	#	P	R	F1	#	P	R	F1	#
New Scientific Information	.88	.77	.82	307	.93	.73	.81	135	.86	.69	.77	145
Scientific Information Removed	.89	.88	.88	108	.80	.98	.88	50	.79	.92	.85	66
Scientific Clarification Added	.72	.76	.74	149	.77	.71	.74	69	.70	.63	.66	51
Scientific Technical Terms Added	.77	.91	.83	214	.86	.95	.90	109	.80	.96	.87	101
Researcher Names Added	.92	.85	.89	81	1.00	.64	.78	28	.90	.86	.88	42
Change in Scientific Narrative	.52	.73	.60	22	.62	.50	.55	16	.54	.87	.67	15
Academic References Added	.94	.86	.90	139	.80	.86	.83	14	.82	.89	.85	45
Academic References Removed	.87	.85	.86	47	.83	.83	.83	6	.79	.89	.84	26
Wikilink Added	.95	.84	.89	234	.96	.86	.91	109	.95	.78	.85	117
Add./Mod. of Quantitative Information	.77	.87	.81	121	.63	.71	.67	42	.81	.96	.88	45
Non Scientific Edit	.90	.89	.89	295	.70	.88	.78	67	.73	.79	.76	72
Macro Avg	.83	.84	.83	1717	.81	.79	.79	645	.79	.84	.81	725
Micro Avg	.85	.84	.85	1717	.83	.81	.82	645	.81	.82	.82	725

Table 2: **Performance Across Domains of GPT-5.4 on WIKISTAR-BENCH Per-Label.** Precision (P), Recall (R), F1, and support (#; number of annotated instances) for the ten scientific edit types and the *Non-Scientific Edit* label.

labels retain high performance throughout, while interpretive labels are weakest in every domain. This consistency indicates that the taxonomy and prompts transfer across scientific domains rather than overfitting to any single one.

3 WIKISTAR: System Overview

WIKISTAR extracts section edit pairs from an article's revision history, classifies each with our multi-label taxonomy (see §2.2), and turns the labeled edits into interactive views that let users explore the article's scientific content across time and sections (see Figure 2). Each component of WIKISTAR addresses a concrete question researchers pose about an article's scientific history: what changed, when, where, and what type of change. We introduce these components through the following scenario:

Consider a historian of science who wants to know how Wikipedia's "Artificial Intelligence" article has changed over time: which ideas were treated as established, when new topics entered, where editors' attention concentrated. The answer lives in the revision history, but reading it by hand is impractical, as it accumulated over 18,000 section edits across two decades.

3.1 History Overview

The *History Overview* view is an LLM-generated summary of how the article's scientific content has evolved. Classification produces a short explanation of what changed in each edit; the *Overview* gathers these explanations for every edit labeled *New Scientific Information* or *Change in Scientific Narrative*—the two labels capturing new or revised scientific claims—and composes them chronologically into a single structured account. A brief summary highlights the most significant changes; clicking `show more` expands it into a period-by-period narrative of the article's lifetime.

The historian begins with the History Overview: the AI article opens with broad definitions of the field, grows more technical as deep learning enters, and turns recently to societal impact—two decades of edits, visible at a glance.

3.2 Trends Over Time

The *Trends Over Time* view provides an interactive line plot comparing the frequency of selected edit types across time (see Figure 2, *Trends Over Time*). By default, the view aggregates yearly across all sections, but supports filtering along several variables to suit each user's focus. Users can select the time frame, the temporal resolution, and the article's sections to include. Clicking a point on the graph generates a summary of the edits it represents (those of a given edit type within a specific period), and the *summarize all time periods* button extends this feature to an entire edit-type series at once; both ground the aggregated trends in the edits' textual content. Below the summary, additional options let users inspect the underlying edit text or export several types of information.

Narrowing the Trends Over Time view to the Applications section, the historian notices a sharp rise around 2023 in new scientific information and technical terms, with few academic references. Click-

ing the peak, they find the summary ties the surge partly to the public release of LLMs, but also, unexpectedly, to editors choosing to add decades-old milestones like Deep Blue and Watson.

3.3 Per-Section Edit Heatmap

The *Section Comparison* view complements the aggregated *Trends Over Time* view with fine-grained comparison across sections. For any selected edit type, it presents a heatmap where each row is a section and color intensity shows how that edit type trends within the section over time (Figure 2, *Section Comparison*). Like the Over Time view, it supports filtering and generates an exportable summary when a cell is clicked. Together, these enable both cross-section comparison and targeted drill-down: users can single out a section, edit type, and period, then surface the actual edits.

Switching to the Section Comparison view, the historian sees where *each change happened, not just when:* Change in Scientific Narrative *concentrates in the early 2000s in* Philosophy *and* History*, while* New Scientific Information *surges after 2020 in* Applications *and* Misinformation*—a shift from debating what AI is to tracking what it does.*

3.4 Section Evolution

The *Section Evolution* view shows how the article's sections evolved. It lays out the sections as rows and time periods as columns, coloring each cell to show what happened to that section in that period: whether it first appeared, was edited in a scientifically meaningful way, or was renamed. This allows users to inspect the development of the structure of the article as it changed throughout the years.

3.5 Edit Table

This view presents a filterable table of all edits across the article's history, together with their metadata. Users can locate specific edits by filtering on metadata fields (e.g. section names and time range) or by free-text search. Then, users see a table of section edits. Each row shows the section text before and after the edit, the edit's metadata, the scientific edit labels the model assigned along with its explanation for each assignment.To support any type of further analysis, the table can be exported.

4 User Study

We conduct a user study with three participants, chosen to span complementary perspectives on how scientific knowledge is produced and evolves on Wikipedia. **P1** is deeply involved in the Wikipedia community as a writer and editor, and is a PhD candidate in Physics. **P2** is a journalist covering cybersecurity and disinformation, with an MA in the history of science. In their work, they frequently treat Wikipedia's edit record as a lens for examining how knowledge is constructed, contested, and manipulated. **P3** is a researcher in the philosophy and history of science, focused on how scientific ideas and models move between fields.

Each participant received an introduction to WIKISTAR, then had 15 minutes to freely explore a scientific Wikipedia article of their choice. They were asked to think aloud. Afterwards, they rated nine statements about the system's utility on a 5-point Likert scale (Table 3). The full script and participant rankings are available in this repository.

Statement Examples	Avg.
I find the types of edits WIKISTAR tracks useful and insightful.	5.00
WIKISTAR would let me do research or analysis that wasn't practical for me before.	5.00
Using WIKISTAR, I noticed potentially interesting things that I could not have easily discovered myself without it.	4.67
WIKISTAR prompted new questions about the evolution of the Wikipedia article I explored that I would not have thought to ask.	5.00
WIKISTAR could allow me to confirm or quantify things I had previously only suspected.	4.33
WIKISTAR could surface interesting evidence about how science is presented to the public.	4.67

Table 3: Participant agreement ratings on system-utility statements, collected via a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). **Avg.** denotes the mean rating across participants. A subset of the 9 rated statements is shown; see full set in this repository.

Tracing a subject across fields. **P3** examined the "Ising model" article, a physics formalism that later spread to other fields. Puzzled that the *Section Comparison* view framed its neuroscience content as "social sciences," they found the answer in *Section Evolution*—the section was renamed "Neuroscience" only in 2022:

> *"It's very interesting to see how this comes together and explains the structure of the article."*

Surfacing invisible structural history. **P1** used WIKISTAR to examine the "Chaos Theory" article, whose foundational research largely predates Wikipedia. They wanted to trace how its structure changed across twenty-five years, and while

inspecting the *Section Evolution* view, they noted:

> *"It's kind of a dream for anyone who discusses and thinks of Wikipedia as a source of statistical data. This is much richer and more detailed than any other tool we currently have in our toolbox as Wikipedia authors and editors. This is really helpful and useful."*

From trends to edits. **P2** used WIKISTAR to examine the "Vaccine" article, tracing scientific debate and controversy across its history. While looking at the *Section Comparison* view, one cell sparked their interest, showing markedly more added scientific content than its neighbors. On discovering that they could generate a summary of the edits composing that cell, they enthused:

> *"That's seriously gorgeous! It adds tons and tons of value. This is a very good analysis that I would actually want and need for my research. It helps me actually do my work."*

They described the experience as "almost like guided reading." Turning to the *Trends Over Time* view, **P2** experimented with the summarization feature across several points on the plot, then chose *View source edits* to inspect the data behind the summaries and reflected:

> *"Now I can connect my pipeline of thinking and research! This option to look at the metadata and the text of the edits is important, because I need to have the data itself and validate it."*

5 Related Work

A growing body of work studies how scientific knowledge is represented and diffuses on Wikipedia (Teplitskiy et al., 2015). Most relevant to us, Benjakob et al. (2023) treat the revision history as a historiographical record of science, reading editorial revisions as intentional acts that mark how knowledge is introduced, refined, and restructured over time. Studies of this kind remain rare because they are manual and labor-intensive, requiring close reading of long revision histories. A single article accumulates thousands of revisions, most of them unrelated to scientific content, so the effort grows with the length of the history and confines such studies to a few articles. In parallel, prior work classifies Wikipedia edits using general edit-type and intention taxonomies (Liu and Ram, 2011; Daxenberger and Gurevych, 2012; Yang et al., 2017), capturing generic operations rather than scientifically meaningful change, while another line visualizes edit provenance and activity, from token-level authorship (Flöck and Acosta, 2014) to edit wars and vandalism (Guo et al., 2023). The first group labels the form of an edit, so two edits that both add a citation receive the same label whether one records a replicated result or repairs a broken link. Neither classifies the *scientific* content of edits, and both operate at the page or token level rather than the section. The page level is too coarse to indicate which part of an article changed, and the token level splits a change into insertions and deletions that must be reassembled before it can be interpreted. WIKISTAR is the first to do both at the section level: it matches each edited section to its previous version and classifies the change into ten scientific types. A separate line defines edit-intention taxonomies for revisions to scientific papers themselves (Jiang et al., 2022; Ruan et al., 2024), tracking how the science evolves rather than how it is represented. There, revisions come from the authors and arrive in discrete versions, while Wikipedia articles are revised by many editors with different aims and no version boundary between rounds of changes. In Wikipedia, scientific edits are buried among many unrelated ones and must first be isolated, which is what WIKISTAR supports. Working at the section level lets users move between one section's history and whole-article trends, making scientific edit analysis accessible, at scale, to anyone interested in how science evolves and is represented online.

6 Conclusion

We presented WIKISTAR, a system for tracking how the scientific content of Wikipedia articles develops over time. WIKISTAR extracts section-level edits, classifies each into an expert-designed taxonomy of ten scientifically meaningful edit types, and presents the results through interactive views that let users move from high-level trends down to the individual edits behind them. We released WIKISTAR-BENCH, the first human-annotated benchmark for scientific edit classification, and used it to compare seven LLMs.

A user study with participants from Wikipedia editing, science journalism, and the history and philosophy of science suggests WIKISTAR is useful across these varied perspectives, enabling new kinds of research into how the representation of science on Wikipedia changes over time.

References

Omer Benjakob, Rona Aviram, and Jonathan Sobel. 2021. Citation needed? wikipedia and the covid-19 pandemic. *BioRxiv*.

Omer Benjakob, Olha Guley, Jean-Marc Sevin, Leo Blondel, Ariane Augustoni, Matthieu Collet, Louise Jouveshomme, Roy Amit, Ariel Linder, and Rona Aviram. 2023. Wikipedia as a tool for contemporary history of science: A case study on crispr. *PLoS One*, 18(9):e0290827.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured wikipedia articles. In *International Conference on Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models.

Romy Fieblinger, Md Tanvirul Alam, and Nidhi Rastogi. 2024. Actionable cyber threat intelligence using knowledge graphs and large language models. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 100–111. IEEE.

Fabian Flöck and Maribel Acosta. 2014. WikiWho: Precise and efficient attribution of authorship of revisioned content. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pages 843–854. ACM.

Yuhan Guo, Qin Han, Yuke Lou, Yiming Wang, Can Liu, and Xiaoru Yuan. 2023. Edit-history vis: An interactive visual exploration and analysis on wikipedia edit history. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*, pages 157–166. IEEE.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arxivedits: Understanding the human revision process in scientific writing. *ArXiv*, abs/2210.15067.

Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manag. Inf. Syst.*, 2:11:1–11:23.

OpenAI. 2025a. Introducing gpt-5. Accessed: 2026-06-02.

OpenAI. 2025b. Openai o3-mini system card. Accessed: 2026-02-18.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: a dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Re3: A holistic framework and dataset for modeling collaborative document revision. *ArXiv*, abs/2406.00197.

Kanchan Shivashankar and Nadine Steinmetz. 2024. Contri (e) ve: Context+ retrieve for scholarly question answering. *arXiv preprint arXiv:2409.09010*.

Arno Simons, Wolfgang Kircheis, Marion Schmidt, Martin Potthast, and Benno Stein. 2024. Who are the "heroes of crispr"? public science communication on wikipedia and the challenge of micro-notability. *Public Understanding of Science (Bristol, England)*, 33:918 – 934.

Misha Teplitskiy, Grace Lu, and Eamon Duede. 2015. Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, 68.

Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 575–582, Vienna, Austria. ACM.

Wikimedia Foundation. 2025. MediaWiki Action API. `https://www.mediawiki.org/wiki/API:Main_page`. Accessed: 2026-07-08.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010.

A Appendix - Edit-Type Taxonomy

Label	Definition
New Scientific Information	Addition of information in the current revision absent from the previous one.
Scientific Information Removed	Removal of at least one complete sentence carrying scientifically meaningful content (factual claims, findings, or definitions) present before but not now; excludes vandalism reverts and non-substantive wording.
Scientific Clarification Added	Modification or expansion of scientifically relevant information already present in the previous revision.
Scientific Technical Terms Added	Addition or modification of domain-specific jargon (field-specific terms, concepts, or processes) not present before and unfamiliar to general audiences.
Researcher Names Added	Inclusion of researcher names not present in the previous revision.
Change in Scientific Narrative	Change in scientific narrative or perspective—a verb tense, framing, or viewpoint shift—relative to the previous revision.
Academic References Added	Addition of references to published academic papers (identifiers such as DOI, PMID, PMC, ISBN, arXiv, or ISSN) not present before.
Academic References Removed	Removal of such academic-paper references present in the previous revision but no longer present.
Wikilink Added	Addition of a link to another Wikipedia page not present in the previous revision.
Add./Mod. of Quantitative Information	Addition or modification of statistics or numeric results (metrics, percentages, counts, ratios, confidence intervals, p-values, or mathematical expressions) not present before.
Non Scientific Edit	No scientifically meaningful modification between revisions; assigned alone.

Table 4: The WIKISTAR taxonomy: ten labels capturing scientifically meaningful edit types and a *Non-Scientific Edit* label. Definitions provided in our prompt released with our our code repository.

B Appendix - Edit Examples

Edit type	Example (*Previous Section Text* → *Edited Section Text*)
New Sci. Information	[. . .] modulated by external cues, primarily daylight. → [. . .] modulated by external cues, primarily daylight. **They allow organisms to anticipate and prepare for environmental changes.** (*Circadian rhythm*, lead section; revision of 2008-05-06.)
Technical Terms Added	[. . .] a protruding tongue and an enlarged tongue near the tonsils, a short neck → [. . .] an enlarged tongue near the tonsils**, or macroglossia**, a short neck [. . .] (*Down syndrome*, "Characteristics" section; revision of 2009-07-22.)
Quantitative Information	[. . .] of the thousands of enumerated ports, *about 250* well-known ports are reserved by convention → [. . .] **1024** well-known ports are reserved by convention [. . .] (*Port (computer networking)*, lead section; revision of 2014-11-30.)
Change in Sci. Narrative	*Prev.*: It is **easier** to obtain high-quality sequence data when the desired DNA is purified [. . .] from any contaminants. *Curr.*: It is **only possible** to obtain high-quality sequence data when the desired DNA is relatively pure, [. . .] free from other contaminants. (*DNA sequencing*, "Large-scale sequencing strategies" section; revision of 2007-02-07.)

Table 5: Before/after excerpts for selected WIKISTAR edit types; changed span in **bold**, [. . .] marks elided text.

C Appendix - Per-Model Results on WIKISTAR-BENCH

Table 6 reports per-label precision, recall, and F1 for every model we evaluated on the full benchmark, extending the single-model view of Table 1. Metrics are computed against the gold labels.

	Precision							Recall							F1						
Label	GPT-5 mini	GPT-5.4	GPT-5.1	LLaMA3.3	o3-mini	GPT-4o	Qwen3	GPT-5 mini	GPT-5.4	GPT-5.1	LLaMA3.3	o3-mini	GPT-4o	Qwen3	GPT-5 mini	GPT-5.4	GPT-5.1	LLaMA3.3	o3-mini	GPT-4o	Qwen3
New Sci. Information	**.93**	.88	.86	.82	.91	.92	.87	**.93**	.74	.92	.65	.69	.48	.83	**.93**	.80	.89	.72	.78	.63	.85
Sci. Information Removed	.81	**.84**	.76	.59	.81	.81	.46	.83	**.92**	.68	.33	.68	.29	.09	.82	**.87**	.72	.42	.74	.43	.15
Sci. Clarification Added	.69	.73	.54	.68	.75	**.80**	.53	**.81**	.72	.75	.59	.39	.27	.66	**.75**	.72	.63	.63	.51	.41	.58
Technical Terms Added	.88	.80	.80	.87	**.92**	**.92**	.66	.83	**.93**	.83	.62	.41	.36	.84	**.86**	**.86**	.81	.72	.57	.52	.74
Researcher Names Added	.77	**.93**	.79	.83	.90	.87	.68	**.95**	.82	.91	.79	.78	.61	.94	.85	**.87**	.84	.81	.84	.72	.79
Change in Sci. Narrative	.41	**.54**	.22	.21	.34	.34	.25	.74	.70	**.87**	.55	.57	.57	.63	.52	**.61**	.35	.30	.43	.42	.36
Academic Refs Added	**.93**	.90	.69	.59	.88	.79	.48	.93	.86	.93	.94	.88	.60	**.98**	**.93**	.88	.79	.73	.88	.68	.64
Academic Refs Removed	**.89**	.84	.75	.53	.85	.65	.38	.82	**.86**	.59	.37	.46	.31	.45	**.85**	**.85**	.66	.44	.60	.42	.41
Wikilink Added	.94	.95	.93	.91	.96	**.97**	.82	**.88**	.83	.71	.60	.52	.24	.41	**.91**	.89	.81	.72	.67	.39	.54
Quantitative Information	.80	.75	.69	.77	**.91**	.80	.59	**.89**	.86	.86	.62	.62	.44	.78	**.84**	.80	.76	.69	.74	.56	.67
Non-Scientific Edit	**.94**	.83	.86	.81	.86	.63	.77	.78	**.87**	.67	.59	.78	.85	.43	**.85**	**.85**	.75	.68	.82	.72	.56
Macro avg	.82	.82	.72	.69	**.83**	.77	.59	**.85**	.83	.79	.60	.62	.46	.64	**.83**	.82	.73	.62	.69	.54	.57
Micro avg	.86	.84	.74	.75	**.87**	.77	.65	**.86**	.83	.80	.61	.62	.46	.64	**.86**	.83	.77	.67	.72	.57	.64

Table 6: **Per-label performance on WIKISTAR-BENCH.** Precision (P), Recall (R), and F1 for each model; best value per row-block in **bold**. *LLaMA3.3* and *Qwen3* denote LLaMA-3.3-70B and Qwen3-80B. The LLaMA, GPT-4o, and o3-mini results are evaluated on N=1,364, N=1,358, and N=1,357 examples respectively; small per-model differences reflect responses dropped after parsing.

D Appendix - Result Comparison Between Prompt Versions

	GPT-5-mini		GPT-5.4		LLaMA	
Label	Naive	Ref.	Naive	Ref.	Naive	Ref.
New Sci. Information	.72	**.93**	.77	**.81**	.73	**.84**
Sci. Information Removed	.69	**.82**	.79	**.87**	.46	**.49**
Sci. Clarification Added	.58	**.75**	.62	**.73**	**.63**	.62
Technical Terms Added	.75	**.86**	.78	**.86**	**.71**	**.71**
Researcher Names Added	.58	**.85**	.71	**.87**	.73	**.81**
Change in Sci. Narrative	.40	**.52**	.32	**.61**	**.29**	.28
Academic References Added	.63	**.93**	.64	**.88**	.63	**.72**
Academic References Removed	.70	**.85**	.72	**.85**	.46	**.50**
Wikilink Added	.86	**.91**	.91	.89	**.81**	.74
Quantitative Information	.78	**.84**	.69	**.80**	**.76**	.73
Non Scientific Edit	.56	**.85**	.63	**.85**	**.63**	**.63**
Macro F1	.66	**.83**	.69	**.82**	.62	**.64**
Micro F1	.68	**.86**	.72	**.83**	.67	**.70**

Table 7: Per-label F1 for the *naive* prompt (a single-sentence definition per label) versus our *refined* prompt (explicit inclusion/exclusion rules, per-label examples, and an up-front scientific/non-scientific decision), across three backbone models. **Bold** marks each refined score that improves on its naive counterpart. *LLaMA* denotes LLaMA-3.3-70B.