SCIENCEMETER: Tracking Scientific Knowledge Updates in Language Models

Yike Wang **Shangbin Feng** **Yulia Tsvetkov** **Hannaneh Hajishirzi**
University of Washington
yikewang@cs.washington.edu

Abstract

Large Language Models (LLMs) are increasingly used to support scientific research, but their knowledge of scientific advancements can quickly become outdated. We introduce SCIENCEMETER, a new framework for evaluating scientific knowledge update methods over scientific knowledge spanning the past, present, and future. SCIENCEMETER defines three metrics: *knowledge preservation*, the extent to which models' understanding of previously learned papers are preserved; *knowledge acquisition*, how well scientific claims from newly introduced papers are acquired; and *knowledge projection*, the ability of the updated model to anticipate or generalize to related scientific claims that may emerge in the future. Using SCIENCEMETER, we examine the scientific knowledge of LLMs on claim judgment and generation tasks across a curated dataset across ten domains. We evaluate five representative knowledge update approaches and find that the best-performing knowledge update methods can preserve only 85.9% of existing knowledge, acquire 71.7% of new knowledge, and project 37.7% of future knowledge, underscoring that developing robust scientific knowledge update mechanisms is both crucial and challenging. [1]

1 Introduction

LLMs are being widely used to aid scientific research (Liang et al., 2024; Luo et al., 2025; Hsu et al., 2024; Qi et al., 2023; Jansen et al., 2025), with the potential to enable even greater future discoveries (Ahn, 2024; Si et al., 2024). However, due to the rapid pace of scientific advancements (Kuhn, 1962) and the static nature of pretrained LLMs (Bommasani et al., 2021), their scientific knowledge quickly becomes stale. We posit that effective scientific knowledge updates in LLMs must do more than simply adding new information, but preserve existing knowledge, incorporate new findings, and enable generalization to reason about future or yet-undiscovered knowledge. Although generic update strategies have been explored, e.g., continual pre-training (Gururangan et al., 2020), instruction-tuning (Wei et al., 2021), or RAG (Shi et al., 2023), it is unclear whether these methods sufficiently support these goals.

To fill this gap, we propose SCIENCEMETER—a new framework for evaluating how LLMs update and reason over scientific knowledge. As shown in Figure 1, our approach centers on tracking scientific knowledge updates as trajectories along three axes: *preservation* of prior knowledge (the parametric knowledge already encoded in the LLM), *acquisition* of new knowledge introduced through knowledge update methods, and *projection* of future knowledge not yet available to the model but can be inferred. Past discoveries serve as the foundation for future advancements and remain valuable for researchers seeking historical context, validation, or reinterpretation of previous findings, while the latter evaluates the utility of knowledge updates in enabling models to internalize new knowledge, moving beyond mere factual memorization to understand the underlying principles and patterns that govern such knowledge. This capability can facilitate hypothesis generation (Qi et al., 2023) and the formulation of novel ideas (Si et al., 2024)—key future usages of AI for science.

Inspired by SciFact (Wadden et al., 2020), SCIENCEMETER operationalizes scientific knowledge as *atomic scientific claims*, i.e., atomic verifiable statements expressing a finding about one aspect of a scientific entity or process, which can be verified against a single source. While prior work in general domains often represents knowledge as factoid information or structured entity tuples (Tang et al., 2024; Yin et al., 2024), we argue that scientific claims are more appropriate and meaningful units of knowledge in scientific contexts, as they

[1] The code and dataset are available at http://github.com/yikee/ScienceMeter.

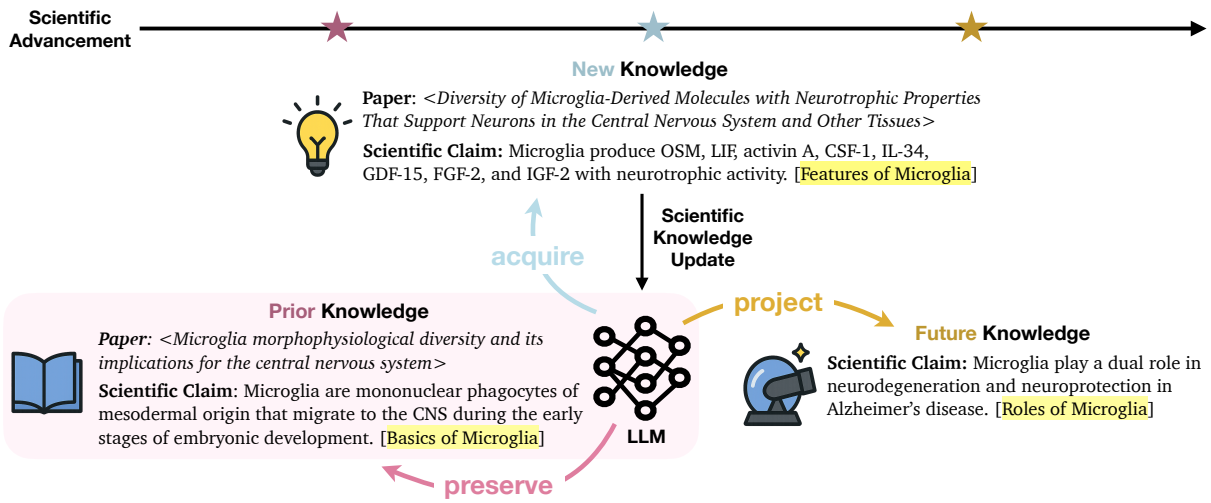

Figure 1: We propose an evaluation framework, SCIENCEMETER, along with novel metrics to quantify the reliability and usefulness of scientific knowledge updates in LLMs: preservation of existing scientific claims and their linkage to existing literature, acquisition of new claims from emerging research, and projection of future claims. For example, when we update an LLM with a new paper introducing the features of Microglia, our framework evaluates the acquisition of this new knowledge, as well as the preservation of existing knowledge about the fundamentals of Microglia learned from previous literature, and the ability of LLMs to effectively use its parametric knowledge to extrapolate future knowledge on Microglia, such as its potential roles in Alzheimer's disease.

better capture the core insights and implications of research beyond isolated numerical values.

In SCIENCEMETER, we curate a dataset spanning ten rapidly evolving scientific fields. As LLMs become increasingly integrated into these domains, it is essential to evaluate whether existing knowledge update methods can support their progress. Related scientific literature is grouped chronologically to represent prior, new, and future knowledge. To evaluate scientific knowledge, we focus on two tasks, *claim judgment* and *claim generation*, and emphasize both *factual accuracy* and the model's *confidence* to better reflect the rigor of the scientific domain. Specifically, we categorize the model's knowledge of a claim as *correct* (factually accurate and confident), *incorrect* (factually inaccurate and confident), or *unknown* (not confident) and quantify the percentage of two types of errors in preservation, acquisition, and projection, respectively.

We evaluate LLMs' scientific knowledge updates using five methods spanning training, inference, or both. Experimental results across standard and frontier models highlight that the best-performing knowledge update method achieve on average only 85.9% on knowledge preservation, 71.7% on knowledge acquisition, and 37.7% on knowledge projection in the domain of *Computer Science*. While inference-time update methods tend to be effective for large models, smaller models require training-based approaches to achieve comparable performance. Cross-domain analysis reveals that performance on these objectives is correlated, with knowledge preservation and projection heavily influenced by domain volatility, while the availability of domain knowledge during pretraining has limited impact. Moreover, even applying on domain-adapted scientific LLMs struggle to balance these objectives, underscoring persistent challenges in updating scientific knowledge in LLMs.

2 SCIENCEMETER

To systematically evaluate scientific knowledge updates in LLMs, SCIENCEMETER integrates three core components: (1) a carefully curated multi-domain dataset of scientific papers and scientific claims (§2.1); (2) evaluation of a model's scientific knowledge through claim judgment and generation tasks, assessed by both factual accuracy and model confidence (§2.2); and (3) novel metrics for evaluating knowledge update methods that aggregate claims from past/present/future data sets (§2.3). An overview of SCIENCEMETER is in Figure 2.

2.1 Dataset Construction

Paper Collection We identify 10 rapidly evolving scientific domains, including *Computer Science*, *Medicine*, *Biology*, *Materials Science*, *Psychology*, *Business*, *Political Science*, *Environmental Science*, *Agricultural and Food Sciences*, and *Education*. For each domain, we randomly retrieve 1,000 journal or conference papers (excluding review or survey papers) published at least three

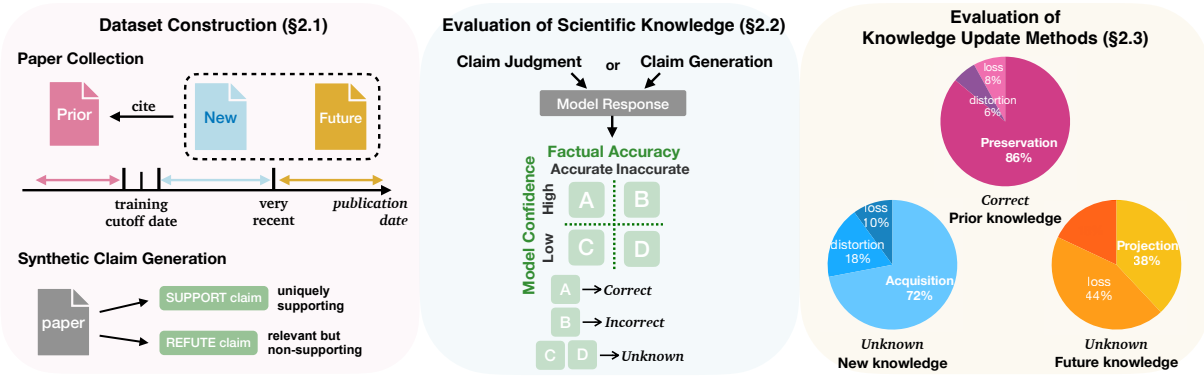

Figure 2: An overview of SCIENCEMETER: (1) We curate chronologically organized datasets of scientific papers and claims; (2) define claim judgment and generation tasks to evaluate scientific knowledge, incorporating both factual accuracy and model confidence; and (3) introduce metrics for evaluating scientific knowledge updates.

months before the knowledge cutoff date of the given model through Semantic Scholar (Ammar et al., 2018), an openly accessible academic search platform that indexes publisher-authorized metadata and papers. To obtain more recent knowledge on the subjects relevant to each paper, we additionally retrieve papers that cite the original paper and were published at least three months after the knowledge cutoff date. We also set a recent cutoff date, beyond which papers serve as a proxy for future knowledge. In total, we constructed 5,148 triplets of $(p_{\text{prior}}, p_{\text{new}}, p_{\text{future}})$, each representing a prior, new, and future version of scientific knowledge on the same topic (Appendix B). An additional filtering will be applied to these papers based on the actual model knowledge (Section 2.3).

Synthetic Claim Generation We invite GPT-4o to synthetically generate one SUPPORT (uniquely supporting) scientific claim and one REFUTE (relevant but non-supporting) scientific claim for each paper, resulting in a total of 15,444 (p, c_{SUPPORT}) and 15,444 (p, c_{REFUTE}) tuples. Expert evaluation by PhD students in their respective fields shows that, on average, 90.2% of the generated claims strictly comply with the specified SUPPORT or REFUTE rules, while 98.7% broadly align with these rules. Further details are provided in Appendix B.

2.2 Evaluation of Scientific Knowledge

Now we want to evaluate a model's scientific knowledge on the collected papers, specifically the scientific claims made in each paper. We propose two tasks, judgment and generation, and categorize the model's knowledge of a claim as *correct*, *incorrect*, or *unknown*, based on both the response's *factual accuracy* and the model's *confidence*.

Task Formulation

- **Claim Judgment** To evaluate a model's knowledge of p_{prior} or p_{new}, we frame the task as a claim verification problem: given the title t of a prior or new scientific paper and its associated claim c, the ground-truth label is $y(c, t) \in \{\text{SUPPORT}, \text{REFUTE}\}$. The model is instructed to predict a label $\hat{y}(c, t)$ for each (c, t) pair. To evaluate a model's knowledge of p_{future}, we adapt the task into a classification setting: given a claim c associated with a "future" paper, the ground truth label $y(c) \in \{\text{SUPPORT}, \text{REFUTE}\}$ indicates whether its associated "future" paper supports or refutes the claim c. The model is asked to predict a label $\hat{y}(c)$ based solely on its internalized knowledge or extrapolative reasoning, without access to the actual paper.
- **Claim Generation** The generation task poses a greater challenge. To evaluate a model's knowledge of p_{prior} or p_{new}, the model is given the title t of a prior or new scientific paper p and instructed to generate a supporting claim $\hat{c}(t)$ such that $y(\hat{c}, t) = \text{SUPPORT}$. To evaluate a model's knowledge of p_{future}, the model is given the subject s of a "future" scientific paper p (with title t) and tasked with generating a supporting claim $\hat{c}(s)$ such that $y(\hat{c}, t) = \text{SUPPORT}$.

We primarily evaluate a model's scientific knowledge based on its ability to correctly recall the claim from a minimal cue (the paper's title), as it is the premise and a crucial step for downstream tasks such as question-answering and RAG. Prior work (Petroni et al., 2019; Zhao et al., 2024b; Wang et al., 2024d) shows that strong recall correlates with higher QA and reasoning performance, making it a reliable proxy for real-world utility.

Task Evaluation Our evaluation methodology emphasizes both *factual accuracy* and the model's *confidence*. We present our choices of measurement methods in Section 3.2. By combining factual accuracy with model confidence, we categorize the model's knowledge of a claim as *correct* (factually accurate and confident), *incorrect* (factually inaccurate and confident), or *unknown* (not confident). We argue that when confidence is low, even a factually accurate answer is not reliable as it may result from hallucination or random chance. This categorization enables a detailed analysis of the impact on scientific knowledge following knowledge updates, as discussed in the next section.

2.3 Evaluation of Knowledge Update Methods

Given the set of papers and associated claims, along with the model's knowledge about each claim (*correct*, *incorrect*, or *unknown*), we can systematically evaluate how a given knowledge update method impact the model's prior, new, and future scientific knowledge. Specifically, we define three core metrics, Knowledge Preservation, Knowledge Acquisition, and Knowledge Projection, as well as two associated error categories, `distortion` and `loss`, as detailed below.

Let $\mathcal{P}_{\text{prior}}$, $\mathcal{P}_{\text{new}}$, and $\mathcal{P}_{\text{future}}$ be sets of prior, new, and future scientific documents in a particular scientific domain. $\mathcal{P}_{\text{new}}$ is introduced using the given scientific knowledge update method f. While documents are collected in such a way that $\mathcal{P}_{\text{prior}}$ is highly likely to appear in the training data of the model, whereas $\mathcal{P}_{\text{new}}$ and $\mathcal{P}_{\text{future}}$ are not (as described in Section 2.1), for each task we only take into account scientific claims in $\mathcal{P}_{\text{prior}}$ that are not *unknown* and claims in $\mathcal{P}_{\text{new}}$ and $\mathcal{P}_{\text{future}}$ that are *unknown* to the model before knowledge updates. Specifically, let g represent either the claim judgment or generation task presented in Section 2.2. Then, given a pre-trained language model LM and a knowledge update method f, we define:

- **Knowledge Preservation** as the percentage of claims in $\mathcal{P}_{\text{prior}}$ that remain *correct*. The proportion of previously *correct* claims that become *incorrect* is referred to as a `distortion` in preservation, while the proportion that becomes *unknown* is considered as `loss`.
- **Knowledge Acquisition** as the proportion of claims in $\mathcal{P}_{\text{new}}$ that the model LM correctly acquires through f, i.e., changing from *unknown* to *correct*. Similarly, the proportion of *unknown* claims that become *incorrect* is referred to as a `distortion` in acquisition, while the proportion that stays as *unknown* is referred to as `loss`.
- **Knowledge Projection** as the percentage of claims in $\mathcal{P}_{\text{future}}$ that the model LM successfully projects after update f, i.e., claims changing from *unknown* to *correct*. Because some incorrect projections may become correct over time, the true magnitude of Knowledge Projection is likely higher than our current estimate. We define the proportion of *unknown* claims that remain *unknown* as `loss`.

We provide the detailed formulas for each metric in Appendix D. The sum of Knowledge Preservation, `distortion`, and `loss` equals one, and the same holds for Acquisition. Among the error types, `distortion` is considered more severe than `loss` in both Preservation and Acquisition scenarios. This is because generating a factually inaccurate response with high confidence (e.g., stating "this claim is SUPPORT for sure" to a REFUTE claim) is more problematic than producing a low-confidence response, regardless of its factual accuracy (e.g., "maybe this claim is SUPPORT/REFUTE"). An optimal scientific knowledge update method should aim to collectively maximize Knowledge Preservation, Acquisition, and Projection.

3 Experiment Settings and Results

In this section, we evaluate five knowledge update methods, covering training, inference, or a combination of both. Due to cost constraints, we conduct experiments in the *Computer Science* domain in this section, and present cross-domain analysis in Section 4. Experiments on both standard and frontier models, along with three confidence measurement approaches, demonstrate the challenge of developing a reliable scientific knowledge update method capable of meeting all three objectives.

3.1 Models

We aim to evaluate the performance of various knowledge update methods on a widely used, reasonably sized model and a frontier large model. As a representative of commonly used mid-sized models, we select LLaMA3.1-8B-Instruct (Grattafiori et al., 2024), and OLMo2-32B-Instruct (OLMo et al., 2024) serves a representative of frontier models, given computational constraints and the limited openness of commercial models. Notably, OLMo2-32B-Instruct has demonstrated frontier

performance while requiring only one-third of the compute of other open-weight models and outperforming GPT-4o mini (OpenAI, 2024). See Appendix G for results on scientific LLMs.

3.2 Factual Accuracy and Model Confidence

Factual Accuracy For the Claim Judgment task, we map the model's predictions $\hat{y}(c, t)$ and $\hat{y}(c)$ to the set {SUPPORT, REFUTE} using manually identified answer patterns, and compare them against the ground-truth labels $y(c, t)$ and $y(c)$, respectively. For the Claim Generation task, we assess the factual accuracy of the generated claim $\hat{c}$ by inviting GPT-4o to determine whether $y(\hat{c}, t) = \text{SUPPORT}$, based on the abstract of the corresponding paper p.

Model Confidence Given the absence of a validation set, we estimate confidence levels using three rule-based measurement methods and finalize the decision through majority voting.

- **More Information** Following existing prompt-based solutions (Feng et al., 2023, 2024b), we append a prompt asking whether more information is needed to answer a given question: *"Do you need more information to answer this question? (Yes or No)"*. Indicating the need for more information suggests a lack of confidence.
- **Consistency** We paraphrase the question three times, sample one response per version, and classify the model as confident only if all three responses give the same binary answer.
- **Linguistic Confidence** Given only the model's response, we prompt GPT-4o (OpenAI, 2024) with the following question: *"Do you think the model is confident about its answer? (Yes or No)"*, aiming to capture implicit linguistic markers of confidence, such as assertive phrasing, authoritative tone, and decisive language in the response. We also conduct a human evaluation of linguistic confidence. The confidence classification consistency between three human evaluators and GPT-4o is 75.9%, thereby validating the effectiveness of GPT-4o as a judge in this task.

All three methods are used as confidence measurement for the judgment responses, while **More Information** is used for generation responses.

3.3 Knowledge Update Methods

We experiment with five knowledge update methods that update new scientific knowledge (i.e., $\mathcal{P}_{\text{new}}$) at either the training stage, the inference stage, or both. $\mathcal{P}_{\text{new}}$ is split into training and test sets, and only the test set will be evaluated. Following previous work on scientific domains (Wadden et al., 2020; Newman et al., 2024), we use abstracts of papers in $\mathcal{P}_{\text{new}}$ instead of full papers, as they typically contain sufficient information and are easier to fit within the model's context window.

Training. Through training, we update the model parameters by minimizing loss defined by different training objectives. Only LoRA adapters (Hu et al., 2021) are trained for all training baselines, with the training duration set to 1 epoch for autoregressive training and 4 epochs for SFT.

Continual Pre-training (CNT PRETRAIN). $\mathcal{P}_{\text{new}}^{\text{test}}$ is introduced through autoregressive training (Gururangan et al., 2020), minimizing the standard next-token prediction loss: $-\frac{1}{|d|}\sum_t \log p_\theta(d_t|d_{<t})$.

Standard Instruction-tuning (INST TUNE). The model is first trained autoregressively on both $\mathcal{P}_{\text{new}}^{\text{train}}$ and $\mathcal{P}_{\text{new}}^{\text{test}}$, and then fine-tuned (Wei et al., 2021) on training QA by minimizing the answer prediction loss given the question: $-\frac{1}{|a|}\sum_t \log p_\theta(a_t|q, a_{<t})$.

Pre-instruction-tuning (PRE INST TUNE). Jiang et al. (2024b) exposes LLMs to QA pairs before continued pre-training on documents. Specifically, the model is instruction-tuned on training QA along with $\mathcal{P}_{\text{new}}^{\text{train}}$ prior to autoregressively trained on $\mathcal{P}_{\text{new}}^{\text{test}}$.

Inference (INFER). The success of in-context learning (Brown et al., 2020) highlights the potential for introducing new knowledge at inference time, offering a more cost-efficient approach. Many existing knowledge augmentation methods, including RAG (Shi et al., 2023), search engines (Press et al., 2022), and multi-LLM collaborations (Feng et al., 2023, 2024a), leverage this strategy to provide additional information. In our setting, we add corresponding p_{new} in $\mathcal{P}_{\text{new}}^{\text{test}}$ to the prompt text.

Training + Inference (INST TUNE + INFER). Following Tang et al. (2024), we also explore whether combining training and inference-time methods can yield improved performance. Specifically, we integrate standard instruction-tuning with the inference-time approach.

Method	Claim Judgment Task								Claim Generation Task							
	Pres	Dist	Loss	**Acqu**	Dist	Loss	**Proj**	Loss	**Pres**	Dist	Loss	**Acqu**	Dist	Loss	**Proj**	Loss
LLAMA3.1-8B-INSTRUCT																
CNT PRETRAIN	85.0	5.5	9.5	37.3	29.9	32.8	34.5	48.3	53.3	30.0	16.7	53.1	42.0	**5.0**	11.8	70.6
INST TUNE	**86.3**	**4.1**	9.6	38.9	28.3	32.8	24.1	41.3	**72.2**	**17.8**	**10.0**	56.1	38.2	5.7	**29.4**	64.7
PRE INST TUNE	59.0	38.3	**2.7**	**64.2**	**26.8**	9.0	44.9	48.2	63.3	23.3	13.3	56.1	37.4	6.5	11.8	64.7
INFER	68.6	17.8	13.6	43.2	50.8	**6.0**	**48.3**	13.7	14.4	62.2	23.3	**84.4**	**8.4**	7.3	23.5	**5.9**
INST TUNE + INFER	69.9	19.2	10.9	41.8	43.3	15.0	44.9	**6.8**	12.2	58.9	28.9	76.0	11.5	12.6	17.6	17.6
OLMO2-32B-INSTRUCT																
CNT PRETRAIN	89.4	**0.0**	10.6	18.7	40.7	40.7	16.6	63.8	82.5	17.5	**0.0**	68.3	31.7	**0.0**	13.1	71.5
INST TUNE	89.5	0.9	9.6	20.3	35.6	44.2	13.8	68.5	**85.8**	**14.2**	**0.0**	67.7	32.3	**0.0**	18.9	71.3
PRE INST TUNE	89.4	0.9	9.7	17.0	39.9	43.2	17.6	65.7	84.2	15.8	**0.0**	68.3	31.7	**0.0**	18.6	63.9
INFER	**99.1**	0.9	**0.0**	**57.7**	**3.3**	**39.0**	**35.3**	**15.6**	42.9	55.8	1.3	79.3	9.9	10.8	**37.6**	**13.7**
INST TUNE + INFER	96.1	0.9	2.9	46.6	6.8	46.7	33.3	18.7	41.7	57.1	1.3	**80.5**	**8.7**	10.8	30.4	26.4

Table 1: Preservation, Acquisition, and Projection performance of knowledge update methods in the domain of *Computer Science*. Best results in **bold** and second best in underline. Since we take model confidence into account (Section 2), the random baseline is below 50. Higher values of preservation, acquisition, and projection are better, while lower values of `distortion` and `loss` are preferred. All methods fail to meet objectives collectively.

3.4 Results

No knowledge update method can simultaneously achieve all three objectives. As shown in Table 1, the best-performing knowledge update methods, averaged across tasks and models, preserve only 85.9% of existing knowledge, acquire 71.7% of new knowledge, and project 37.7% (or more) of future knowledge. However, we fail to find a method that can achieve all three objectives collectively. We additionally provide results on factual accuracy alone, without taking confidence into account, in Appendix F for reference. Overall, standard instruction-tuning and inference methods remain as the strongest methods across all five. Enabling models to project future knowledge presents a new challenge for knowledge update. As LLMs become increasingly integrated into scientific workflows such as hypothesis and idea generation, it becomes critical to develop update methods that not only integrate new claims but also enable models to anticipate and reason about future advancements.

Inference-based methods work for larger models, whereas smaller models require training to achieve comparable performance. For instance, in the claim judgment task, OLMo2-32B achieves inference-time performance that is 10.9% higher than training-based methods in Knowledge Preservation on average, whereas the inference-time method on LLaMA-8B performs 10.7% worse than training-based methods. This discrepancy arises in part from the larger models' capacity to effectively filter out irrelevant or noisy contextual information during inference. With their greater representational power and more robust internal attention mechanisms, larger models are less sensitive to distractions in the prompt or context (Brown et al., 2020; Askell et al., 2021; Min et al., 2022), allowing them to incorporate new knowledge with minimal degradation of prior understanding. From a computational perspective, inference-based update is significantly more cost-effective than training (Houlsby et al., 2019; Rebuffi et al., 2017; Lester et al., 2021), especially when updates are frequent. Notably, combining inference with additional training does not lead to performance improvements over inference alone, suggesting diminishing returns from training once a model has sufficient capacity to leverage inference-based strategies effectively. For smaller models, however, training remains a necessary step to compensate for their limited ability to generalize and disambiguate new knowledge in context. More in Appendix I.3.

In the more challenging task, `distortion` is significantly greater than `loss`. Specifically, in the Generation task, the amount of `distortion` is, on average, three times higher than `loss` in both Preservation and Acquisition. As we discussed in Section 2.3, `distortion` is considered more severe than `loss` in both scenarios. This observation suggests a significant challenge for knowledge update methods, as they may introduce errors even

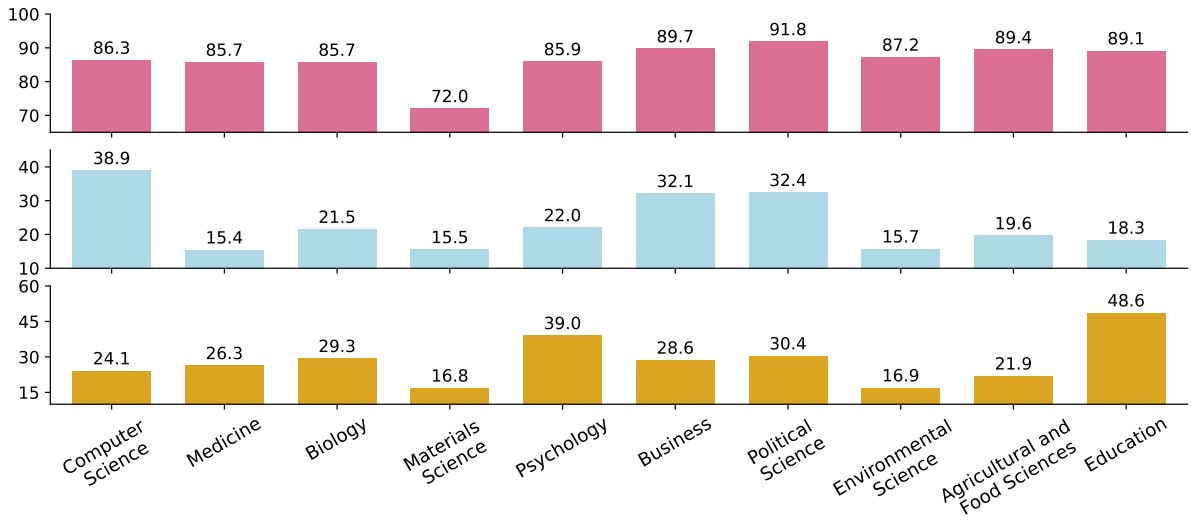

Figure 3: Preservation, Acquisition, and Projection performance of Standard Instruction-tuning on LLAMA3.1-8B in the claim judgment task. The performance in *Materials Science* and *Environmental Science* is poor across all three objectives, whereas *Political Science* and *Education* show relatively strong performance in all three.

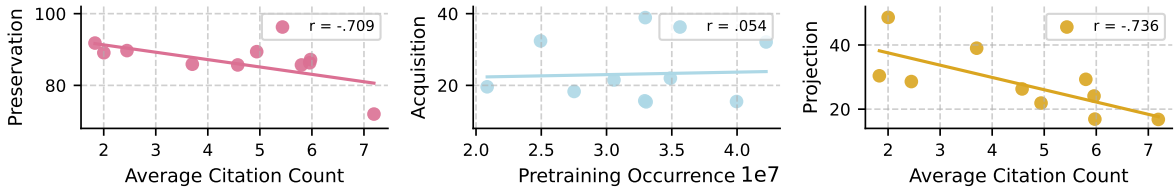

Figure 4: The correlation between Preservation, Acquisition, Projection and average citation count, as well as pretraining occurrence. The closer the points are to the best-fit line, the stronger the correlation.

when they should remain cautious. To address this issue, future developments in knowledge update methods could incorporate an abstention mechanism, avoiding updating models' representations if they lack confidence or certainty about the new content. Such a mechanism would allow models to opt out of updates when faced with ambiguous or uncertain knowledge, helping to preserve accuracy and reduce the propagation of errors.

4 Cross-domain Analysis

In this section, we further break down performance by scientific domain. As shown in Figure 3, performance varies significantly across domains. While over 90% of scientific knowledge in *Political Science* is preserved, only 72% of *Materials Science* knowledge can be retained. Similarly, while 48.6% of scientific knowledge in *Education* can be projected, this drops to just 16.8% in *Materials Science*. We also notice that these three capabilities appear to be correlated. Performance in certain domains tends to be consistently poor across all three tasks, for example, *Materials Science* and *Environmental Science*, whereas domains such as *Political Science* exhibit relatively strong performance across all three objectives.

We hypothesize that the domain performance might be influenced by two key factors:

First, the nature of the domain, specifically the stability or volatility of domain knowledge. Knowledge preservation and projection may be more challenging in domains with higher volatility (such as *Materials Science*, with frequent breakthroughs or shifting paradigms that rapidly change the state of knowledge) compared to those with greater stability (such as *Political Science*, with long-established theories and principles that evolve slowly over time). To test this hypothesis, we randomly retrieve 1,000 papers published during the same time period in each domain, and calculate the average citation count for these papers (Appendix H), under the assumption that higher average citation counts reflect higher knowledge volatility.

Second, the availability of domain knowledge in the pretraining corpus. We posit that if domain knowledge appears frequently or widely in the pretraining corpus, knowledge acquisition might be easier. To assess this possibility, we collect the 100 tokens that appear least frequently in the abstracts of these 1,000 papers in each domain, as they tend to be specialized tokens unique to each domain. We then use Infini-gram (Liu et al., 2024a) to count the occurrence of these tokens in the pretraining

corpus Dolma-v1.7 (Soldaini et al., 2024), and use the average occurrence of domain-specific tokens as a proxy for the availability of domain knowledge in the pretraining data (Appendix H).

As shown in Figure 4, our analysis reveals a strong relationship between the ability to preserve and project scientific knowledge and the dynamics of the domain. Specifically, the Pearson correlation coefficients (Pearson, 1895) between average citation count and Knowledge Preservation and Projection are -0.709 and -0.736 respectively, both indicating a significant relationship. Highly dynamic domains with frequent updates may lead to more knowledge conflicts (Wang et al., 2023), making preservation and projection particularly challenging. In contrast, the correlation between pretraining availability and Knowledge Acquisition is relatively weak, indicating that pretraining alone may have a limited impact on how well models adapt to evolving scientific information.

5 Related Work

LLMs for Scientific Advancements Recent research has demonstrated the significant potential of LLMs in driving scientific advancements across various domains, revolutionizing the way researchers approach complex problems and innovate in their respective fields (Liang et al., 2024; Luo et al., 2025; Hsu et al., 2024; Qi et al., 2023; Jansen et al., 2025; Ahn, 2024; Si et al., 2024; Starace et al., 2025). Researchers utilize off-the-shelf LLMs (AI4Science and Quantum, 2023), domain-specific scientific LLMs (Zhang et al., 2024b), or LLMs augmented with external resources (Asai et al., 2024) to assist in scientific research. Studies show that current LLMs can be useful across various stages of the research cycle (Luo et al., 2025), including literature review (Hsu et al., 2024; Agarwal et al., 2024), hypothesis proposing (Qi et al., 2023), idea generation (Si et al., 2024), and experiment planning and implementation (Jansen et al., 2025; Huang et al., 2023). However, to the best of our knowledge, we are the first to examine whether LLMs can stay up to date and continuously contribute to the advancement of these fields.

Evaluation of Knowledge Updates in LLMs Our evaluation of scientific knowledge updates differs from existing work on evaluation of knowledge updates in LLMs (Li and Goyal, 2025; Sun et al., 2025; Wang et al., 2024a) in three key aspects. First, most prior work primarily regards the effective incorporation of new information as the only objective (Jang et al., 2021; Ovadia et al., 2023; Jiang et al., 2024b; Zhang et al., 2024a; Tang et al., 2024; Yin et al., 2024; Zhao et al., 2024a; Jang et al., 2022), with few also considering the preservation of old knowledge (Jiang et al., 2024b; Zhang et al., 2024a). However, they rely on generic benchmarks such as Natural Questions (Kwiatkowski et al., 2019) and CommonsenseQA (Talmor et al., 2018), which evaluate the retention of general world knowledge rather than the preservation of knowledge related to the newly updated information. And we further introduce a new evaluation dimension, evaluating the utility of knowledge updates for reasoning, hypothesis generation (Qi et al., 2023), and the creation of novel ideas (Si et al., 2024), which are the key future applications of AI in science. Second, existing approaches heavily rely on Wikipedia as the data source and assess knowledge at the factoid level (e.g., names, locations) (Jang et al., 2021; Ovadia et al., 2023; Jiang et al., 2024b; Zhang et al., 2024a; Tang et al., 2024; Yin et al., 2024; Zhao et al., 2024a; Jang et al., 2022), whereas we extend evaluation to natural language representations, which better capture the core insights and implications of research beyond isolated numerical values as well as the complexity of real-world knowledge. Third, prior work on knowledge alignment primarily focuses on temporal alignment (Zhao et al., 2024a; Yin et al., 2024; Jang et al., 2022), aiming to align knowledge to specific timestamps, such as associating a president with a particular year, while our goal is to align scientific claims with scientific literature.

6 Conclusion

In this work, we investigate scientific knowledge updates of LLMs and propose that an effective and reliable update method should be able to preserve existing scientific knowledge, acquire new scientific knowledge, and project future scientific knowledge, which are crucial for the continual use of LLMs in evolving scientific fields. To this end, we introduce an evaluation framework SCIENCEMETER with rich datasets of scientific papers across domains, new tasks and evaluation of scientific knowledge, and new metrics for evaluating knowledge update methods. With comprehensive experiments on frontier general-purpose and science-focused LLMs, we find that achieving these objectives remains an open research challenge, underscoring the need for further exploration in this direction.

Limitations

Real scientific advancement is far more complex In this work, we model scientific advancement as a linear timeline spanning existing, new, and future developments. However, genuine scientific progress is considerably more complex in reality. New advancements often emerge from the convergence of multiple research trajectories across diverse domains. Future work should aim to capture this multidimensional nature of scientific progress.

Beyond scientific claims While this work focuses on scientific claims as the fundamental unit of analysis for evaluating scientific knowledge and scientific knowledge updates, we recognize that scientific knowledge operates at multiple meaningful levels of granularity. Other critical dimensions worthy of investigation include the paper-level and researcher-level, which could be potential directions for future research.

Contradictory claims When evaluating future scientific knowledge using claim classification tasks, we acknowledge that, theoretically, there is a chance that some claims may contradict past findings. However, empirical evidence suggests such occurrences are rare. Moreover, our claims are sufficiently detailed and comprehensive, making it unlikely that identical or highly similar claims have appeared in prior literature.

Disentangling knowledge from instruction-following and reasoning capabilities Separating the "knowledge" of LLMs from their instruction-following and reasoning abilities is challenging, particularly if we define "knowledge" as the content they generate. Prior work has attempted to assess LLMs' knowledge using cloze-style tasks (Jang et al., 2021) at inference time, which rely more on raw knowledge and less on instruction-following ability. However, such formats are limited to evaluating factoid knowledge. In this work, we define scientific knowledge as scientific claims and propose claim judgment and generation tasks to evaluate it. While these tasks are effective for assessment and analysis, we acknowledge that model performance on them still depends, to some extent, on instruction-following and reasoning capabilities.

Ethics Statement

We envision certain potential ethical risks of SCIENCEMETER. For example, when evaluating "future" scientific claims, the framework risks creating ethical dilemmas regarding premature validation of unproven hypotheses. This becomes particularly problematic when assessing claims in sensitive domains (e.g., climate science or medical research) where premature endorsement could influence policy or clinical decisions.

However, SCIENCEMETER also provides significant ethical benefits by introducing systematic transparency to scientific knowledge assessment. SCIENCEMETER can help surface meritorious but underrecognized research directions, and these features may ultimately promote more equitable and evidence-based scientific progress when implemented with appropriate ethical safeguards.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300), and NSF IIS-2044660. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This work was also supported in part by Azure credits from Microsoft Accelerate Foundation Models Research.

References

Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*.

Sangzin Ahn. 2024. The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions. *The Korean journal of physiology & pharmacology: official journal of the Korean Physiological Society and the Korean Society of Pharmacology*.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, and 1 others. 2018. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and

1 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Simone Clemente, Zied Ben Houidi, Alexis Huet, Dario Rossi, Giulio Franzese, and Pietro Michiardi. 2025. In praise of stubbornness: The case for cognitive-dissonance-aware knowledge updates in llms. *arXiv preprint arXiv:2502.04390*.

Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, Tomas Pfister, Yejin Choi, and Yulia Tsvetkov. 2024a. When one llm drools, multi-llm collaboration rules. *arXiv preprint arXiv:2403.17852*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models. *arXiv preprint arXiv:2305.09955*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024b. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

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*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Guoxiu He, Xin Song, and Aixin Sun. 2025. Knowledge updating? no more model editing! just selective contextual reasoning. *arXiv preprint arXiv:2503.05212*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Lu Wang, and Aakanksha Naik. 2024. Chime: Llm-assisted hierarchical organization of scientific studies for literature review support. *arXiv preprint arXiv:2407.16148*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Benchmarking large language models as ai research agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. *arXiv preprint arXiv:2204.14211*.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2021. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. 2025. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems*, 37:10088–10116.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, and 1 others. 2024a. Learning to edit: Aligning llms with knowledge editing. *arXiv preprint arXiv:2402.11905*.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen-tau Yih, and Srinivasan Iyer. 2024b. Instruction-tuned language models are better knowledge learners. *arXiv preprint arXiv:2402.12847*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Aochong Oliver Li and Tanya Goyal. 2025. Memorization vs. reasoning: Updating llms with new knowledge. *arXiv preprint arXiv:2504.12523*.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, and 1 others. 2024. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024a. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Tianci Liu, Zihan Dong, Linjun Zhang, Haoyu Wang, and Jing Gao. 2025. Mitigating heterogeneous token overfitting in llm knowledge editing. *arXiv preprint arXiv:2502.00602*.

Zeyu Leo Liu, Shrey Pandit, Xi Ye, Eunsol Choi, and Greg Durrett. 2024b. Codeupdatearena: Benchmarking knowledge editing on api updates. *arXiv preprint arXiv:2407.06249*.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. Arxivdigestables: Synthesizing scientific literature into tables using language models. *arXiv preprint arXiv:2410.22360*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *ArXiv*, abs/2210.03350.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

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*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. 2023. Honeybee: Progressive instruction finetuning of large language models for materials science. *arXiv preprint arXiv:2310.08511*.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson,

and 1 others. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Chen Sun, Renat Aksitov, Andrey Zhmoginov, Nolan Andrew Miller, Max Vladymyrov, Ulrich Rueckert, Been Kim, and Mark Sandler. 2025. How new data permeates llm knowledge and how to dilute it. *arXiv preprint arXiv:2504.09522*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yugang Jiang, and 1 others. 2024. Evowiki: Evaluating llms on evolving knowledge. *arXiv preprint arXiv:2412.13582*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Changyue Wang, Weihang Su, Hu Yiran, Qingyao Ai, Yueyue Wu, Cheng Luo, Yiqun Liu, Min Zhang, and Shaoping Ma. 2024a. Lekube: A legal knowledge update benchmark. *arXiv preprint arXiv:2407.14192*.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024b. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37:53764–53797.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024c. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024d. Unveiling factual recall behaviors of large language models through knowledge neurons. *arXiv preprint arXiv:2408.03247*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. 2024. History matters: Temporal knowledge editing in large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19413–19421.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Yipeng Zhang, Haitao Mi, and Helen Meng. 2024a. Self-tuning: Instructing llms to effectively acquire new knowledge through self-teaching. *arXiv preprint arXiv:2406.06326*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A Smith. 2024a. Set the clock: Temporal alignment of pretrained language models. *arXiv preprint arXiv:2402.16797*.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024b. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. *arXiv preprint arXiv:2406.12277*.

A Related Work (Cont.)

We distinguish our evaluation of knowledge preservation from Catastrophic Forgetting (CF) in Continual Learning (CL) (Kirkpatrick et al., 2017; Jang et al., 2021; Clemente et al., 2025), as our setting involves multiple training stages and methods. Another relevant line of work is knowledge editing (Meng et al., 2022b; Wang et al., 2024c; Meng et al., 2022a; Zhang et al., 2024b; Liu et al., 2024b; Wang et al., 2024b; Jiang et al., 2024a; Liu et al., 2025; He et al., 2025), which aims to replace incorrect existing knowledge, whereas our goal is to integrate new knowledge without altering the model's understanding of previously learned scientific literature.

B Dataset Details

B.1 Date Cutoffs

Table 2 presents the specific date cutoffs used to construct the dataset for all models in this study. A three-month buffer accounts for potential discrepancies between a paper's online availability and its official publication date, allowing a more accurate representation of respective knowledge.

B.2 Metadata

Table 3 shows the number of papers in each domain in our dataset.

B.3 Synthetic Claims

To generate synthetic claims for each paper, we prompt GPT-4o with the instructions detailed in Table 4. We explored various alternative methods

Model	Cutoff	Prior Knowledge	New Knowledge	Future Knowledge
LLAMA3.1-8B	Dec 2023	2022.10.1 - 2023.9.30	2024.3.1 - 2024.11.30	2024.12.1 - 2025.2.1
OLMO2-7B	Dec 2023	2022.10.1 - 2023.9.30	2024.3.1 - 2024.11.30	2024.12.1 - 2025.2.1
OLMO2-32B	Dec 2023	2022.10.1 - 2023.9.30	2024.3.1 - 2024.11.30	2024.12.1 - 2025.2.1
HONEYBEE	Oct 2023	2022.8.1 - 2023.7.31	2024.1.1 - 2024.11.30	2024.12.1 - 2025.3.1

Table 2: Date cutoffs used to distinguish prior, new, and future knowledge when constructing the dataset for different models.

Domain	Paper Count
Computer Science	835
Medicine	480
Biology	351
Materials Science	559
Psychology	491
Business	503
Political Science	409
Environmental Science	455
Agricultural and Food Sciences	533
Education	532
SUM	5148

Table 3: Paper Count in each domain after filtering out papers without citation information or abstracts. Originally, 1,000 papers were retrieved per domain.

for synthetic claim generation, such as retrieving a relevant paper and using its SUPPORT claim as the REFUTE claim for the given paper. However, the method we ultimately adopted, despite its simplicity, yielded the best results. Additionally, we control the granularity of claims by constraining their length to approximately 15 words, ensuring that they are neither overly simplistic nor excessively verbose (e.g., the entire abstract).

We further conduct an expert evaluation of our synthetic claims. We invite 20 PhD students across ten domains and assign each student 50 papers within their respective area of expertise. For each paper, we provide the title, abstract, and two synthetic claims, and they are instructed to classify each claim into one of the following categories:

- **Uniquely Supported** – The claim can only be verified by the given paper.
- **Broadly Supported** – The claim is supported by the given paper but is likely validated by other sources as well.
- **Not Supported** – The claim is not supported by the given paper.

The results are presented in Table 5, showing that, on average, 90.2% of claims strictly adhere to the rule, while 98.7% broadly meet our expectations. While the results demonstrate the effectiveness of our synthetic claims, there is still room for improvement, so we try to collect author-annotated claims in Section B.4.

B.4 Author-annotated Claims

As the original authors of research papers possess the most relevant expertise for claim annotation, we aim to collect author-annotated claims in the *Computer Science* domain under a limited annotation budget. Specifically, we develop a scientific claim annotation interface (Figure 5 and Figure 6) and email the first authors of 500 papers from our *Computer Science* dataset. Although some emails were undeliverable, we successfully received responses from 43 authors, resulting in 86 annotated scientific claims (43 labeled as SUPPORT and 43 as REFUTE). Out of the 43 responses, 38 authors accepted the proposed SUPPORT claims and 29 authors accepted the proposed REFUTE claims. For the rejected cases, the authors edited and revised the claims.

Our evaluation using LLaMA-8B on this author-annotated dataset (Table 6) reveals no statistically significant performance difference compared to synthetic claims. This finding empirically validates the effectiveness of our synthetic claim generation methodology, suggesting that the synthetic claims maintain comparable quality to human expert annotations while offering scalability advantages.

C Claim Judgment and Generation Tasks

We present the prompts used for the claim judgment and generation tasks in Table 7 and Table 8.

Prompt: SUPPORT Claim Generation

System Prompt
You are an expert scientific research assistant.

User Prompt
Please identify and extract the main scientific claim that is uniquely supported by the given paper. A scientific claim is a atomic verifiable statements expressing a finding about one aspect of a scientific entity or process, which can be verified against a single source.

Prompt: REFUTE Claim Generation

System Prompt
You are an expert scientific research assistant.

User Prompt
Please identify and extract a scientific claim that is relevant but not supported by the given paper. A scientific claim is a atomic verifiable statements expressing a finding about one aspect of a scientific entity or process, which can be verified against a single source.

Table 4: Prompt templates for synthetic claim generation.

Domain	Label	Uniquely Supported	Broadly Supported	Not Supported
Computer Science	SUPPORT	92%	8%	0%
	REFUTE	0%	10%	90%
Medicine	SUPPORT	86%	12%	2%
	REFUTE	2%	8%	90%
Biology	SUPPORT	86%	14%	0%
	REFUTE	0%	12%	88%
Materials Science	SUPPORT	88%	8%	4%
	REFUTE	2%	14%	84%
Psychology	SUPPORT	98%	2%	0%
	REFUTE	4%	10%	86%
Business	SUPPORT	94%	4%	2%
	REFUTE	0%	18%	82%
Political Science	SUPPORT	86%	10%	4%
	REFUTE	2%	10%	88%
Environmental Science	SUPPORT	98%	2%	0%
	REFUTE	0%	2%	98%
Agricultural and Food Sciences	SUPPORT	94%	4%	2%
	REFUTE	0%	6%	94%
Education	SUPPORT	88%	10%	2%
	REFUTE	0%	6%	94%

Table 5: Expert evaluation results on synthetic claims, averaged across two experts per domain.

Overview

Thank you for participating in our study!

We invite you to provide two scientific claims about your paper:

- **Main supported claim**: The main scientific claim that is uniquely supported by your paper's findings.
- **Relevant but unsupported claim**: A scientific claim that is related to the topic of your paper, but which your paper does not support.

Note: A scientific claim is a atomic verifiable statements expressing a finding about one aspect of a scientific entity or process, which can be verified against a single source. Aim for approximately 20 words per claim.

We draft both claims for you. Please feel free to edit, refine, or rewrite them as needed, or accept the draft claims if they fully satisfy the requirements.

Paper Information

Title	**Unveiling Performance Challenges of Large Language Models in Low-Resource Healthcare: A Demographic Fairness Perspective**
Publication Date	2024-11-30
Link	Semantic Scholar
Abstract	This paper studies the performance of large language models (LLMs), particularly regarding demographic fairness, in solving real-world healthcare tasks. We evaluate state-of-the-art LLMs with three prevalent learning frameworks across six diverse healthcare tasks and find significant challenges in applying LLMs to real-world healthcare tasks and persistent fairness issues across demographic groups. We also find that explicitly providing demographic information yields mixed results, while LLM's ability to infer such details raises concerns about biased health predictions. Utilizing LLMs as autonomous agents with access to up-to-date guidelines does not guarantee performance improvement. We believe these findings reveal the critical limitations of LLMs in healthcare fairness and the urgent need for specialized research in this area.

Figure 5: An example of the scientific claims author annotation interface.

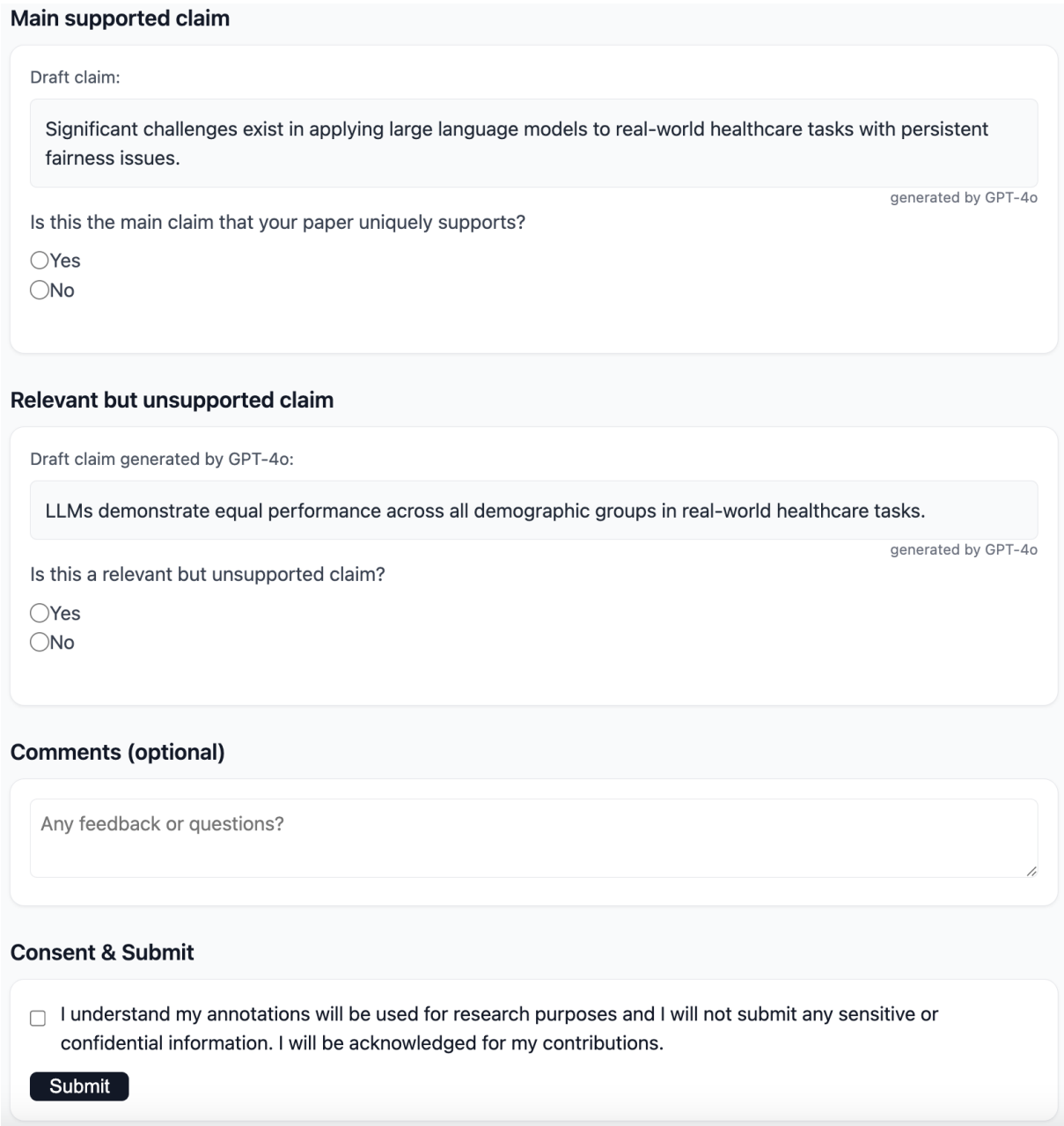

Figure 6: An example of the scientific claims author annotation interface (cont.). Authors can either accept the proposed claim or edit it in the edit boxes.

Model	Preservation	Distortion	Loss	Acquisition	Distortion	Loss	Projection	Loss
SYNTHETIC CLAIMS	86.3	4.1	9.6	38.9	28.3	32.8	24.1	41.3
AUTHOR-ANNOTATED CLAIMS	89.1	3.2	7.7	34.2	27.5	38.3	22.6	37.2

Table 6: We evaluate Standard Instruction-tuning on LLaMA-8B using our claim judgment task with synthetic and author-annotated claims in *Computer Science* respectively. The results demonstrate no statistically significant difference between model performance on synthetic versus author-annotated claims, which validates the effectiveness of our synthetic claim generation approach.

Prompt: Claim Judgment Task - Claim Verification (Prior and New Knowledge)

System Prompt
You are an AI research assistant designed to provide accurate, evidence-based responses.

User Prompt
claim: {claim}
Can every detail in the given claim be substantiated by the paper {title}?

Prompt: Claim Judgment Task - Claim Classification (Future Knowledge)

System Prompt
You are an AI research assistant designed to provide accurate, evidence-based responses.

User Prompt
claim: {claim}
Is the claim correct?

Table 7: Prompt templates for Claim Judgment Task.

Prompt: Claim Generation Task - Prior and New Knowledge

System Prompt
You are an AI research assistant designed to provide accurate, evidence-based responses.

User Prompt
State the main scientific claim made in the paper {title}. A scientific claim is a atomic verifiable statements expressing a finding about one aspect of a scientific entity or process, which can be verified against a single source.

Prompt: Claim Generation Task - Future Knowledge

System Prompt
You are an AI research assistant designed to provide accurate, evidence-based responses.

User Prompt
State a scientific claim about {subject}. A scientific claim is a atomic verifiable statements expressing a finding about one aspect of a scientific entity or process, which can be verified against a single source.

Table 8: Prompt templates for Claim Generation Task.

D Metrics

Table 9 provides the detailed mathematical definitions of knowledge preservation, knowledge acquisition, and knowledge projection, as well as `distortion` and `loss`.

E Experiment Details

We use learning rate of 2×10^{-4} for model optimization. Experiments are performed on a cluster with 4 A100 GPUs each with 40 GB memory.

F Change in Factual Accuracy without Model Confidence

We also report the results of different knowledge update methods on the factual accuracy of prior knowledge in Table 10, without taking model confidence into account.

G Scientific LLMs

Rather than relying solely on off-the-shelf LLMs, researchers also use scientific LLMs (Zhang et al., 2024b), LLMs specifically trained or adapted for science. In this work, we also experiment with HoneyBee (Song et al., 2023), a llama-based model fine-tuned for *Materials Science* domain using high-quality, relevant textual data from the open literature. As we find that the performance of scientific knowledge updates in *Materials Science* is significantly lower than other domains (Section 4), we wonder if applying on a specialized scientific LLM could help. As shown in Table 11, scientific LLMs show no significant improvement compared to off-the-shelf LLMs of similar sizes, highlighting the unique challenges involved in updating scientific knowledge in terms of preservation, acquisition, and projection.

H Cross-domain Analysis

This section provides details on cross-domain analysis discussed in Section 4.

H.1 Citation Count

See Table 12 for the average citation counts of papers in our dataset across ten scientific domains, collected from Semantic Scholar (Ammar et al., 2018).

H.2 Domain-specific Tokens

We randomly retrieve 1,000 conference or journal papers published between October 2022 and September 2023 for each of the ten domains. From these abstracts, we extract the 100 least frequently occurring tokens using the LLAMA3.1 tokenizer. Stop words, punctuation, and numbers are removed from the list. The full list of tokens is provided in Table 13 and Table 14.

H.3 Occurrence in Pretraining Corpus

Refer to Table 15 for the average occurrence of domain-specific tokens in the Dolma-v1.7 pretraining corpus (Soldaini et al., 2024), collected using Infini-gram (Liu et al., 2024a).

H.4 Correlation

Figure 7 presents the correlation between Preservation and pretraining occurrence, Acquisition and average citation count, as well as Projection and pretraining occurrence.

I Discussion and Justification

I.1 Paper Titles in Task Formulation

Considering the cases that different papers may present conflicting or opposing claims, for example, one paper may report that drinking water is beneficial for health, while another may argue that excessive water intake can be harmful under certain conditions, we include paper titles in the task setting when evaluating both prior and newly introduced knowledge. This design makes the task more challenging: models must not only understand or memorize individual claims, but also associate claims with their sources and distinguish between potentially conflicting findings across papers.

I.2 Knowledge Projection

We use claim classification and claim generation as a practical way to operationalize and quantify aspects of future scientific knowledge. We view it as a tractable and meaningful approximation that allows controlled evaluation. Our goal is to assess whether a model can use its internalized knowledge to form a reasonable sense of potential future discoveries. The ability to extrapolate from existing knowledge, recognize emerging patterns, and generalize to future findings is an important aspect of forecasting in scientific contexts. Our evaluation measures whether models can anticipate or correctly judge claims that reflect subsequent scientific developments, which we consider a meaningful and measurable facet of forecasting ability.

Knowledge Preservation $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{prior}}) = \textit{correct} \mid g(LM, p^i_{\text{prior}}) = \textit{correct}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}{\sum_i \mathbb{I}(g(LM, p^i_{\text{prior}}) = \textit{correct}, g(LM, p^i_{\text{new}}) = \textit{unknown})}$
`distortion` in Preservation $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{prior}}) = \textit{incorrect} \mid g(LM, p^i_{\text{prior}}) = \textit{correct}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}{\sum_i \mathbb{I}(g(LM, p^i_{\text{prior}}) = \textit{correct}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}$
`loss` in Preservation $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{prior}}) = \textit{unknown} \mid g(LM, p^i_{\text{prior}}) = \textit{correct}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}{\sum_i \mathbb{I}(g(LM, p^i_{\text{prior}}) = \textit{correct}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}$
Knowledge Acquisition $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{new}}) = \textit{correct} \mid g(LM, p^i_{\text{new}}) = \textit{unknown})}{\sum_i \mathbb{I}(g(LM, p^i_{\text{new}}) = \textit{unknown})}$
`distortion` in Acquisition $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{new}}) = \textit{incorrect} \mid g(LM, p^i_{\text{new}}) = \textit{unknown})}{\sum_i \mathbb{I}(g(LM, p^i_{\text{new}}) = \textit{unknown})}$
`loss` in Acquisition $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{new}}) = \textit{unknown} \mid g(LM, p^i_{\text{new}}) = \textit{unknown})}{\sum_i \mathbb{I}(g(LM, p^i_{\text{new}}) = \textit{unknown})}$
Knowledge Projection $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{future}}) = \textit{correct} \mid g(LM, p^i_{\text{future}}) = \textit{unknown}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}{\sum_i \mathbb{I}(g(LM, p^i_{\text{future}}) = \textit{unknown}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}$
`loss` in Projection $= \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}^{\text{test}}_{\text{new}})}, p^i_{\text{future}}) = \textit{unknown} \mid g(LM, p^i_{\text{future}}) = \textit{unknown}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}{\sum_i \mathbb{I}(g(LM, p^i_{\text{future}}) = \textit{unknown}, g(LM, p^i_{\text{new}}) = \textit{unknown}))}$

Table 9: Detailed formulations of evaluation metrics introduced in Section 2.3. Note that p^i_{prior} and p^i_{future} are considered only if p^i_{new} is *unknown* to the model before knowledge updates, as otherwise no new scientific knowledge is introduced.

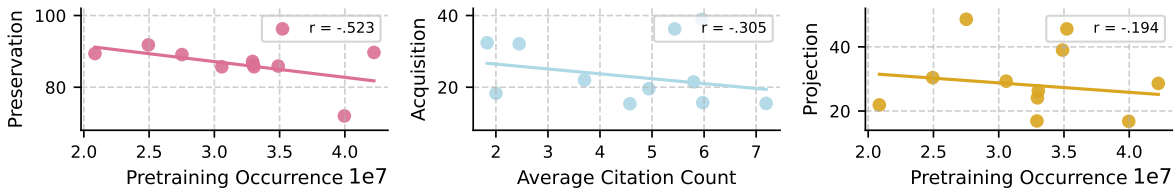

Figure 7: The correlation between Preservation, Acquisition, Projection and average citation count, as well as pretraining occurrence. The closer the points are to the best-fit line, the stronger the correlation.

We define Knowledge Projection as the percentage of claims in $\mathcal{P}_{\text{future}}$ that transition from *unknown* to *correct* after the knowledge update. Two conditions must be satisfied:

- The claim is unknown before the update.
- The claim is correct with high confidence after the update.

In other words, prior to the knowledge update, the model is unable to answer the claim with sufficient confidence (regardless of whether its low-confidence response happens to be correct or incorrect). After the update, the model answers correctly and with high confidence. The only variable that changes between these two responses is the knowledge update itself. This controlled setup isolates the effect of the introduced knowledge trajectory. Therefore, any systematic shift from "unknown" to "correct with high confidence" can be attributed to the knowledge update, rather than to domain-wide priors or generic scientific templates. While some stochastic variation is inevitable in generative models, the controlled before–after comparison provides direct evidence that the projection is conditioned on the introduced knowledge. If the model were merely relying on field-level expectations, we would not expect such transition.

I.3 Other Models

We additionally evaluate two more models: OLMo-2-7B-Instruct and Llama-3.2-3B-Instruct, using standard instruction-tuning and inference settings,

Method	Claim Judgment Task				Claim Generation Task			
	Accurate to Accurate	Accurate to Inaccurate	Inaccurate to Accurate	Inaccurate to Inaccurate	Accurate to Accurate	Accurate to Inaccurate	Inaccurate to Accurate	Inaccurate to Inaccurate
LLAMA3.1-8B-INSTRUCT								
CNT PRETRAIN	58.3	5.2	1.7	34.8	34.3	19.9	9.6	36.1
INST TUNE	60.9	2.6	1.7	34.8	44.6	9.6	12.7	33.1
PRE INST TUNE	37.4	26.1	30.4	6.1	40.4	13.9	11.4	34.3
INFER	40.9	22.6	12.2	24.3	13.3	41.0	4.8	41.0
INST TUNE + INFER	46.1	17.4	1.7	34.8	12.0	42.2	2.4	43.4
OLMO2-32B-INSTRUCT								
CNT PRETRAIN	52.9	1.6	4.2	41.3	59.3	12.6	8.4	19.8
INST TUNE	51.3	3.2	6.9	38.6	61.7	10.2	9.0	19.2
PRE INST TUNE	51.9	2.6	3.7	41.8	60.5	11.4	9.9	18.3
INFER	54.0	0.5	0.0	45.5	30.8	41.0	5.1	23.1
INST TUNE + INFER	53.4	1.1	0.5	45.0	29.9	41.9	6.9	21.3

Table 10: Changes in the factual accuracy of prior knowledge in the domain of *computer science* under different models and settings.

Model	Preservation	Distortion	Loss	Acquisition	Distortion	Loss	Projection	Loss
LLAMA3.1-8B-INSTRUCT	**72.0**	11.0	**17.0**	15.5	**13.4**	71.7	**16.8**	75.8
OLMO2-7B	60.3	16.9	22.8	11.9	36.0	52.1	15.6	76.0
HONEYBEE-7B	62.6	**0.0**	37.4	**18.2**	42.8	**39.0**	15.1	**69.0**

Table 11: Performance of Standard Instruction-tuning on off-the-shelf and scientific LLMs in the claim judgment task within the domain of *Materials Science*. Best results in **bold** and second best in underline. HoneyBee is a materials science model fine-tuned on LLaMa-7B.

Domain	Citation Count
Computer Science	5.957
Medicine	4.575
Biology	5.799
Materials Science	7.192
Psychology	3.702
Business	2.447
Political Science	1.832
Environmental Science	5.973
Agricultural and Food Sciences	4.939
Education	2.002

Table 12: The average citation count of 1,000 conference or journal papers published between October 2022 and September 2023 across different domains.

respectively. Results (Table 16) on these additional smaller models further support our conclusion: smaller models generally require training, and inference-based methods alone are often insufficient due to their limited in-context learning capabilities.

I.4 Confidence Estimation Methods

In our framework, we do not predefine a single method for confidence estimation. Instead, we leave the choice of confidence measurement to users or practitioners who adopt the framework, so they can select methods that best fit their models and evaluation settings.

I.5 Abstract instead of Full paper

We use abstracts as the primary source of scientific content for the following reasons:

- **Copyright and data access.** Using the full text of papers can introduce practical constraints related to data availability and copyright restrictions, which limit large-scale, systematic use of full articles.

Domain	Specialized Tokens
Computer Science	[Background, chunks, mined, keywords, -res, ourced, Million, logistic, LR, encoder, coder, isolate, solved, imperfect, realized, abrupt, transmitted, connects, ended, tan, imoto, waveform, coefficients, -current, Self, -driving, navigation, drivers, orientation, camera, installed, videos, combinations, Next, Track, substantially, Thanks, mil, king, Trad, itionally, EB, -from, -more, including, dairy, deviations, sequent, lact, Est, herit, splitting, Rem, Mi, Together, encompass, setup, concluding, seaborn, matplotlib, Num, Py, Log, Literary, properly, client, -server, Client, Server, -Agent, Rapid, oring, planner, inverse, kin, ematic, fourth, Transfer, HTTP, send, Wireless, Control, missions, envi, nets, -agent, Path, -aware, preservation, extends, intermediate]
Medicine	[logic, outputs, AY, applic, subt, ropical, underscores, gam, publicly, overlapping, warrant, intimate, website, Domestic, Violence, DV, item, DV, observation, attainment, jun, foregoing, divorce, offspring, focused, uns, aturated, UF, palm, Animals, aily, Spatial, lost, -Jan, uary, Identification, Matrix, Laser, Ion, Time, Flight, rometer, MAL, Possible, encountering, opportun, Admission, inertia, RAP, charge, iny, -fe, alan, mem, brane, reversed, PAR, carbohydrate, -chain, aur, -en, rich, chicken, iated, recipients, misconception, abandonment, sustaining, optim, -ag, Enhanced, -k, Da, property, aiding, TJ, Adopt, -condition, polarization, Moh, -Tr, optic, interf, amil, arth, ref, dup, sister, ismus, disclosed, fe]
Biology	[Qu, QS, attracts, basics, realization, solving, oriented, priorities, productive, oking, subt, timely, priority, ials, uns, aturated, Di, Twenty, palm, Animals, aily, -k, aiding, ulcer, rebound, TJ, Adopt, ada, Br, voltage, Nav, hurdle, arr, hyth, mic, :c, :p, Nav, exponentially, impose, UG, specialised, Laur, Material, -response, iaux, -comp, -death, Cock, ayne, olated, -rate, visceral, Unexpected, dc, Sp, pm, Nos, Statement, Kid, okes, emergency, monitored, urine, elo, album, -sk, ewed, Highly, inherently, paths, quasi, Americans, idi, opathic, ATIC, III, restrictive, Character, omorphic, stature, -height, Binding, slow, cognition, BF, doubled, unexpectedly, Well, come, Council, Horizon]
Materials Science	[Even, afford, lacks, Rh, unnecessary, seem, believed, restrictive, transistor, NR, ampl, -terminal, stand, gate, COM, vertically, dissertation, satisfy, gien, Regular, convinced, rows, satin, stitch, ext, skipping, stitches, spent, row, Regression, duce, choose, passes, -cons, istent, VP, -mult, VP, supplemented, strengthened, deflect, meticulous, seems, Hamilton, energetic, warp, shr, mesh, widening, inferior, Spacer, twisting, earlier, oogeneous, alter, retained, vil, abundance, ol, clin, partition, Applying, minute, iles, Tit, -visible, follow, -first, law, Ev, Add, itive, Manufacturing, yx, ylon, chopped, Fil, Fabric, manners, inevitable, Increase, Rotary, sty, ABS, isot, chips, dispens, conforms, Cross, VR, incomplete]
Psychology	[supporters, Turkish, Federation, league, season, mekte, aca, reside, deploy, February, undergone, referral, .Pre, vious, Using, partner, .Actor, .Al, though, careers, disabling, hab, ilitation, vo, rehabilit, ROT, Reality, pencil, Compar, ventional, uo, -exec, expanding, Council, arts, theatrical, Documentation, encode, professor, Yuri, Kon, stant, ovich, orn, twenties, ungen, Menschen, affen, Ap, ensured, Wall, Thought, Anyway, -long, Sol, wind, -al, gorithm, izable, Simon, scientists, .L, Rub, Brush, insky, Ya, onom, contempor, Kor, la, Login, .F, Spi, rid, Lap, isto, pol, sk, .N, Sav, ols, outlines, specializing, Today, attracted, squares, phys, immense, scan, ancers, dancers, Oper, recognizing]

Table 13: Specialized tokens by domains.

- **Context window limitations.** Incorporating full papers into the context window is challenging, particularly for the model sizes considered in this work. We also did not rely on closed commercial models for this analysis, as their knowledge cutoffs are not transparent and they do not allow training or controlled updating, which is necessary for our experimental setting.

- **Consistency with prior work.** Our approach follows prior research in scientific domains that also uses abstracts as the primary source of scientific content (Wadden et al., 2020; Newman et al., 2024).

I.6 Grouping Scientific Knowledge with Citations

Citation links are the signal we use to group knowledge and we want to argue that scientific progress is usually not a strict continuation of a single, narrowly defined research topic. Using the example of transformers, given a specific topic, what evolves most frequently and rapidly is often not a direct extension of the core architecture itself, but related ideas, such as new applications of the architecture, analyses of its properties, or methodological variations centered around the same general area. Direct follow-up work that proposes entirely new architectures certainly exists, but it tends to be less common than these broader, related developments. The broader developments are often the actual driving force behind subsequent direct advances.

Domain	Specialized Tokens
Business	[universities, shaped, intents, Planned, variance, cur, ricula, colleges, hiding, thinks, Mixed, edir, esign, Ped, est, rian, busiest, worship, Aut, ad, plan, Sketch, stone, lamps, disabilities, night, shade, trees, trash, cans, benches, ender, neutrality, pose, proves, Cycling, territory, initially, loc, quali, quant, -line, Content, ardin, -art, supervised, ervised, Random, Boost, IW, AL, CH, AG, PH, tactics, Analy, engaging, Among, tips, istrict, awi, Boston, Consulting, regulator, backward, penetration, automobiles, -side, Rating, internal, insurance, distribution, Observation, interpreted, bands, Merch, andise, band, tok, po, plain, publish, stories, inders, keepers, consent, Customers, annoyed, technology, ynamic, breakdown, Revenue, GRA, Tam]
Political Science	[Mayor, alignments, reputation, cular, tapping, -period, Deputy, Chair, chairman, upcoming, Glob, unexpected, bur, sts, -demand, ding, negligence, -ray, ultrasound, oxygen, cylinders, bribery, coll, usion, cov, -care, India, rebuild, aped, reproductive, Pregnancy, Assessment, Monitoring, merged, -unit, corresponds, predicted, Models, Medicaid, uninsured, constr, lethal, essential, inaccurate, hypothetical, underestimate, innocent, ale, assass, massac, adopts, rig, idity, Usage, Use, Sig, Received, affili, omics, Highlands, Ranch, Colorado, Ang, lia, Norfolk, Economics, Biology, eos, Cor, respond, Andrew, Page, prosper, chaotic, breaks, Reports, Officials, slap, scr, ulous, collusion, receives, update, unify, Method, -trans, subordinate, ordination, prescribed, abandon, status, liquid, ields]
Environmental Science	[Ach, arya, Narendra, Technology, Kum, anj, Ay, hya, .P, ban, horizontally, Sm, breaks, combust, ibles, -contained, breathing, charger, differentiated, gar, mist, charged, firefighters, etal, attractive, afford, hollow, template, lacks, alloy, aceous, giving, mo, ieties, Associated, flatt, aling, current, illustrates, Growing, arms, easing, offsets, margins, sentinel, Gui, Woody, Native, Increase, ensured, Gas, economical, ENT, uction, pression, -dis, charge, formula, Fresh, someone, wants, easiest, acronym, Add, Assessment, Alternative, opted, executed, adversely, effected, Jas, Percent, retain, igated, executor, affairs, ochrome, ringing, anch, Geo, -grid, PL, Net, rein, forc, -ing, Mon, omantic, trace, retained, diamond, vil, syn, analog]
Agricultural and Food Sciences	[Background, arms, easing, -offs, offsets, sentinel, Gui, Version, Woody, Native, Imp, Increase, Trees, expense, Across, que, stration, Ins, bodily, Large, Blue, elle, Wood, Color, guaranteed, components, vit, dispersion, mixer, completeness, sustaining, urgently, scheduling, intric, sovereignty, expenditures, spending, excessively, pleasing, purple, Jerusalem, Hel, thus, Cal, brom, igh, yer, Fiber, Analyzer, -An, kom, zap, mango, Sit, aja, Sap, vene, rys, llum, lance, andra, -J, reserves, subsets, categorized, identical, fairness, CNN, impressive, showcasing, Bihar, consequ, odule, attrib, inferred, unavoidable, worrying, conscient, happens, prescribed, impossible, shed, anticipation, CCC, cricket, WF, blends, CCC, Purchase, atisfied, usted, assert, predictors, Leipzig, Actual, fuels]
Education	[Jur, udence, Enough, Weak, magnitude, =a, +b, RTL, param, etric, Plans, Infrastructure, Super, Japanese, Young, intents, Planned, squares, legitimate, recognized, launch, Br, song, gains, core, ivism, Ecology, Human, entity, instantly, environmentally, Ec, teeth, minimized, .Result, .Con, waves, alyze, Evaluate, WAR, PER, IOD, ropy, yz, hev, Regional, Archive, rad, martial, informational, histor, resist, battlefield, acting, aters, Ukrain, protect, fitting, super, asks, otic, Lim, Tang, gam, overlapping, ineffective, Liter, Connected, compat, Evalu, SET, unten, redesign, -made, checked, weighted, Messenger, emails, iber, rooted, emancip, deficit, implicitly, depr, overly, applic, -created, Ps, LD, omencl, etiquette, ingu]

Table 14: Specialized tokens by domains. (Continued)

Domain	Average Token Occurrence
Computer Science	32966797
Medicine	33036396
Biology	30569548
Materials Science	39959970
Psychology	34891007
Business	42227384
Political Science	24943232
Environmental Science	32928017
Agricultural and Food Sciences	20853024
Education	27514910

Table 15: The average number of occurrence of domain-specific tokens (as identified in Table 13 and Table 14) in the Dolma-v1.7 (Soldaini et al., 2024) pretraining corpus.

Method	Preservation	Distortion	Loss	Acquisition	Distortion	Loss	Projection	Loss
				LLAMA-3.2-3B-INSTRUCT				
INST TUNE	56.2	21.0	22.8	29.5	29.6	40.9	23.7	30.2
INFER	51.3	35.6	13.1	30.0	31.3	38.7	25.2	18.8
				OLMO-2-7B-INSTRUCT				
INST TUNE	72.0	8.9	19.1	37.6	31.6	30.8	25.5	40.2
INFER	62.2	21.9	15.9	35.2	41.0	23.8	29.3	24.0

Table 16: Results on additional models under claim judgment task.

Our goal is to mimic real scientific progress in the wild, which is often more scattered and heterogeneous than a sequence of tightly connected, linear continuations. Accordingly, when we update a model with new scientific progress, we aim to evaluate to what extent the model can preserve prior knowledge, acquire new knowledge, and project forward. This projection does not necessarily have to be a direct continuation of a specific topic; it may also involve generating or reasoning about work that is broadly related within the surrounding research area. Closely related knowledge is indeed more likely to be impacted, but broader knowledge may also be reshaped, forgotten, blurred, or recontextualized as scientific understanding evolves. Importantly, we do not discard directly related knowledge. Such work is naturally included within the subset of cited papers and therefore remains part of our evaluation. Our approach extends the scope to additionally measure influence on more generally related work, aiming for a more comprehensive assessment of knowledge evolution.

I.7 Scalability

Our contribution is primarily the methodology and evaluation framework, rather than a fixed dataset.

We have validated the effectiveness of our data curation process, particularly the claim generation methods, which makes the framework adaptable to new domains or to models with different temporal knowledge cutoffs. In practice, adapting the framework only requires collecting relevant papers and sampling new claims, rather than redesigning the evaluation from scratch.

Moreover, if the difference between knowledge cutoff dates is not substantial, users or practitioners can also reuse appropriate subsets of the existing dataset, which further reduces reconstruction costs and improves scalability.