SCIENCEMETER: Tracking Scientific Knowledge Updates in Language Models

Yike Wang$^\heartsuit$ **Shangbin Feng**$^\heartsuit$ **Yulia Tsvetkov**$^\heartsuit$ **Hannaneh Hajishirzi**$^{\heartsuit\spadesuit}$
$^\heartsuit$University of Washington $\spadesuit$Allen Institute for Artificial Intelligence
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are increasingly used to support scientific research, but their knowledge of scientific advancements can quickly become outdated.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 of 15,444 scientific papers and 30,888 scientific claims from ten domains including *medicine, biology, materials science,* and *computer science*. We evaluate five representative knowledge update approaches including training- and inference-time methods. With extensive experiments, we 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. Inference-based methods work for larger models, whereas smaller models require training to achieve comparable performance. Cross-domain analysis reveals that performance on these objectives is correlated. Even when applying on specialized scientific LLMs, existing knowledge update methods fail to achieve these objectives collectively, underscoring that developing robust scientific knowledge update mechanisms is both crucial and challenging.

 Code and Data   github.com/yikee/ScienceMeter

1 Introduction

LLMs are being widely used to aid scientific research [30, 34, 17, 44, 22], with the potential to enable even greater future discoveries [2, 47]. However, due to the rapid pace of scientific advancements [26] and the static nature of pre-trained LLMs [7], 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., via continual pre-training [14], instruction-tuning [59], or retrieval-augmented generation [46], it is not clear 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
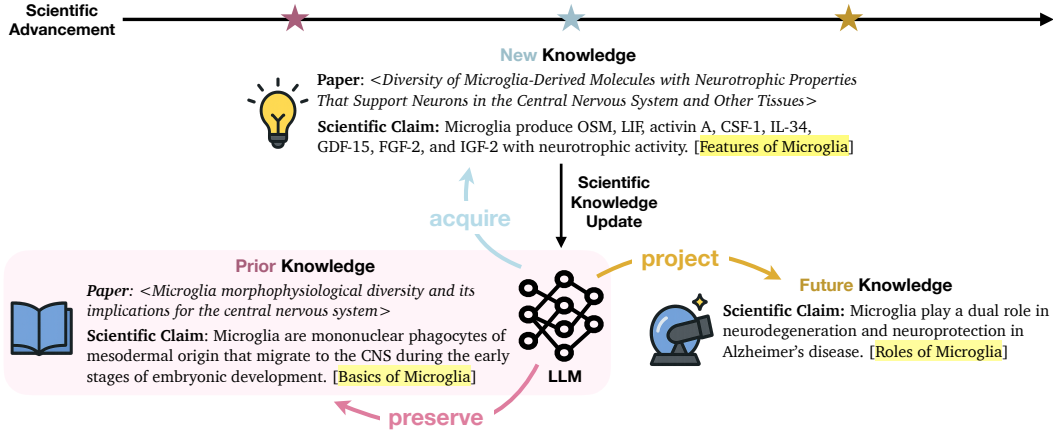
Figure 1: We propose an evaluation framework, SCIENCEMETER, along with novel metrics to quantify the reliability and usefulness of scientific knowledge updates in LLMs: preservation of existing scientific claims and their linkage to existing literature, acquisition of new scientific claims from emerging research, and projection of future scientific claims. For example, when we update an LLM with a new paper introducing the effective identification of 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 effectively use its parametric knowledge to extrapolate future knowledge on Microglia, such as potential roles of Microglia in the Alzheimer's disease.

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 advanced reasoning, hypothesis generation [44], and the formulation of novel ideas [47]—key future usages of AI for science.

Inspired by SciFact [54], 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 [53, 60], we argue that scientific claims are more appropriate and meaningful units of knowledge in scientific contexts, as they better capture the core insights and implications of research beyond isolated numerical values.

In SCIENCEMETER, we curate a large-scale, multi-domain dataset encompassing 15,444 research papers and 30,888 scientific claims across 10 rapidly evolving scientific fields, including *medicine, biology, materials science,* and *computer science*. 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 based on publication dates. To evaluate scientific knowledge, we focus on two tasks: *claim judgment*, and *claim generation*. To better reflect the rigor of the scientific domain, our evaluation methodology emphasizes both *factual accuracy* and model's *confidence*. 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. 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 specialized, domain-adapted scientific LLMs struggle to balance these objectives, underscoring persistent challenges in updating scientific knowledge in LLMs.
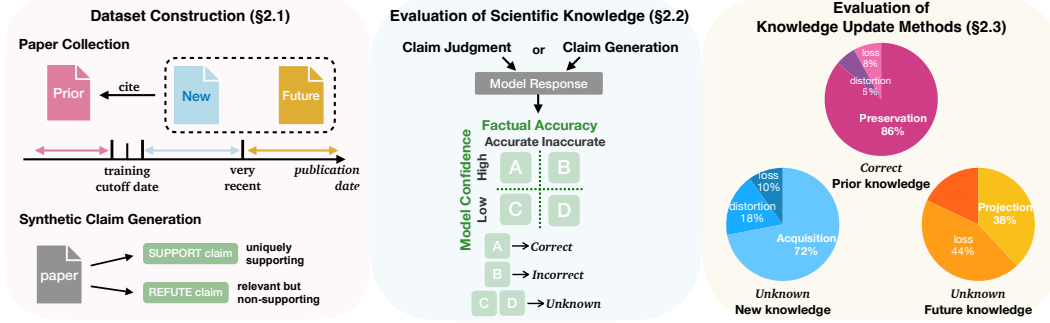
Figure 2: An overview of SCIENCEMETER: (1) We curate chronologically organized datasets of scientific papers and claims across 10 rapidly evolving domains; (2) define claim judgment and generation tasks to evaluate scientific knowledge, incorporating both factual accuracy and model confidence; and (3) introduce metrics for evaluating scientific knowledge updates.

2 The SCIENCEMETER Evaluation Framework

To systematically evaluate scientific knowledge updates in LLMs, SCIENCEMETER integrates three core components: (1) a carefully curated dataset consisting of 15,444 scientific papers and 30,888 scientific claims (§2.1); (2) evaluation of 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 illustrated in Figure 2.

2.1 Dataset Construction

Paper Collection

We identify 10 rapidly evolving scientific domains: *Computer Science*, *Medicine*, *Biology*, *Materials Science*, *Psychology*, *Business*, *Political Science*, *Environmental Science*, *Agricultural and Food Sciences*, and *Education*. LLMs are increasingly integrated into scientific research across these domains, so it is crucial to assess whether LLMs can continuously contribute to these domains through knowledge updates.

For each domain, we retrieve 1,000 journal or conference papers (excluding review or survey papers) published at least three months before the knowledge cutoff date of the given model using the Semantic Scholar API [4]. This three-month window accounts for potential discrepancies between a paper's online availability and its official publication date, ensuring a more accurate representation of the knowledge. To obtain more recent knowledge on the subjects relevant to each paper, we perform an additional query to the Semantic Scholar API, retrieving 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. The specific cutoff dates used for all models examined in this study are detailed in Appendix C. 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.

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

Synthetic Claim Generation and Expert Validation We 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 confirms that at least 80% of the generated claims strictly comply with the specified rule, while over 95% broadly align with our expectations, demonstrating the effectiveness of our synthetic

claim generation approach. To further validate our methodology, we collect a set of author-annotated claims and conduct additional experiments. The results consistently support the efficacy of our approach, indicating that synthetic claims achieve comparable quality to those annotated by human experts while offering significant scalability benefits. Further details are provided in Appendix C.

2.2 Evaluation of Scientific Knowledge

Now we want to evaluate a model's scientific knowledge on the papers collected in Section 2.1, 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" scientific 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 any specific 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 $y(\hat{c}, t) = \text{SUPPORT}$.

Task Evaluation

To better reflect the rigor of the scientific domain, 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 prior, new, and future 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. 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 scientific claims associated with $\mathcal{P}_{\text{prior}}$ 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 scientific claims associated with $\mathcal{P}_{\text{new}}$ 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 scientific claims associated with $\mathcal{P}_{\text{future}}$ 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 E. 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, Knowledge Acquisition, and Knowledge Projection.

3 Experiment Settings and Results

In this section, we evaluate five knowledge update methods, covering training, inference, or a combination of both. Experiments on both standard and frontier models, along with three confidence measurement approaches, demonstrate the challenge of developing a such 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 [13], and OLMo2-32B-Instruct [39] 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 [40].

3.2 Factual Accuracy and Model Confidence Measurement Methods

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) = $ 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 [10, 12], 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 responses for each version, and classify the model as confident if all responses converged on the same final binary answer.

- **Linguistic Confidence** Given only the model's response, we prompt GPT-4O [40] 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 [54, 38], we use

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	<u>85.0</u>	5.5	<u>9.5</u>	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	<u>28.3</u>	32.8	24.1	41.3	**72.2**	**17.8**	**10.0**	56.1	38.2	<u>5.7</u>	**29.4**	64.7
PRE INST TUNE	59.0	38.3	**2.7**	**64.2**	**26.8**	<u>9.0</u>	<u>44.9</u>	48.2	<u>63.3</u>	23.3	13.3	56.1	37.4	6.5	11.8	64.7
INFER	68.6	17.8	13.6	<u>43.2</u>	50.8	**6.0**	**48.3**	<u>13.7</u>	14.4	62.2	23.3	**84.4**	**8.4**	7.3	<u>23.5</u>	**5.9**
INST TUNE + INFER	69.9	19.2	10.9	41.8	43.3	15.0	<u>44.9</u>	**6.8**	12.2	58.9	28.9	<u>76.0</u>	<u>11.5</u>	12.6	17.6	<u>17.6</u>
OLMO2-32B-INSTRUCT																
CNT PRETRAIN	89.4	**0.0**	10.6	18.7	40.7	<u>40.7</u>	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	<u>84.2</u>	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	<u>79.3</u>	9.9	10.8	**37.6**	**13.7**
INST TUNE + INFER	<u>96.1</u>	0.9	2.9	<u>46.6</u>	<u>6.8</u>	46.7	<u>33.3</u>	<u>18.7</u>	41.7	57.1	1.3	**80.5**	<u>8.7</u>	10.8	<u>30.4</u>	26.4

Table 2: Performance of knowledge update methods in the domain of *Computer Science*. Best results in **bold** and second best in <u>underline</u>. Performance are color-coded per category: Preservation, Acquisition, Projection. 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.

abstracts of papers in $\mathcal{P}_{new}$ 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 [18] 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).** $\mathcal{P}_{new}^{test}$ is introduced through autoregressive training [14], 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 $\mathcal{P}_{new}^{train}$ and $\mathcal{P}_{new}^{test}$, and then fine-tuned [59] 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. [24] introduces a new method that exposes LLMs to QA pairs before continued pre-training on documents. Specifically, the model is instruction-tuned on training QA along with $\mathcal{P}_{new}^{train}$ prior to autoregressively trained on $\mathcal{P}_{new}^{test}$.

Inference (INFER). The success of in-context learning [8] highlights the potential for introducing new knowledge at inference time, offering a more cost-efficient approach. Many existing knowledge augmentation methods, including retrieval-augmented generation [46], search engines [43], and multi-LLM collaborations [10, 11], leverage this strategy to provide additional information. In our setting, we add corresponding paper p_{new} in $\mathcal{P}_{new}^{test}$ to the prompt text and $g(LM_{f(\mathcal{P}_{new}^{test})}, p) = g(LM|p_{new}, p)$.

Training + Inference (INST TUNE + INFER). Following Tang et al. [53], 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.

3.4 Results

No knowledge update method can simultaneously achieve all three objectives. As shown in Table 2, 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. Overall, standard instruction-tuning and inference methods remain as the strongest methods across five. Enabling models to project future knowledge presents a new challenge for knowledge update. As LLMs become increasingly integrated into scientific workflows, especially tasks 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 scientific advancements.
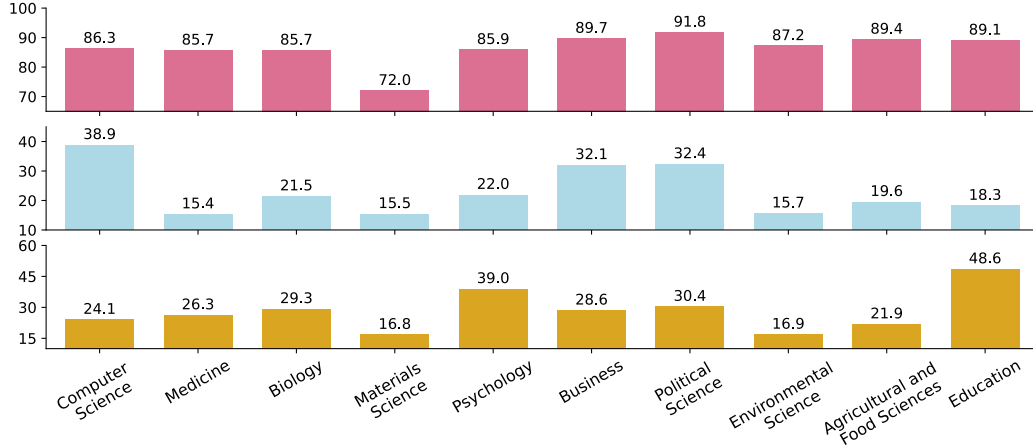
Figure 3: Performance of Standard Instruction-tuning on LLAMA3.1-8B in the claim judgment task. Performance are color-coded per category: Preservation, Acquisition, Projection. The performance in *Materials Science* and *Environmental Science* is poor across all three objectives, whereas *Political Science* and *Education* show relatively strong performance in all three.

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 [8, 6, 37], 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 [16, 45, 28], 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 component to compensate for their limited ability to generalize and disambiguate new knowledge in context.

In the Claim Generation task, `distortion` is significantly greater than `loss`. Specifically, in the more challenging Claim 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 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 Analysis

In this section, we perform additional analyses across different scientific domains. The results reveal that performance on the three objectives is correlated; Preservation and Projection exhibit strong dependence on domain volatility, whereas the availability of domain knowledge in the pretraining corpus demonstrates only marginal influence. We also evaluate scientific LLMs in the worst-performing domain (i.e., *Materials Science*) and find that even applying on domain-adapted models struggle to achieve all three objectives, underscoring persistent challenges in updating scientific knowledge.

4.1 Cross-domain Analysis

In this section, we further break down performance by scientific domain and analyze potential factors that may influence the preservation, acquisition, and projection of scientific knowledge. As shown in
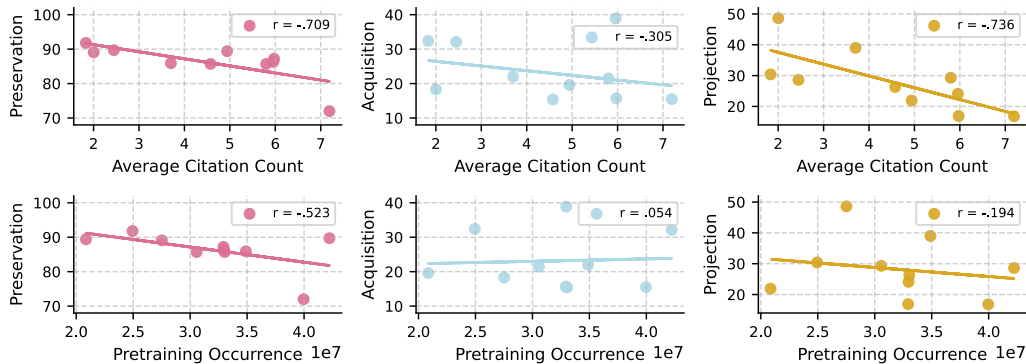
Figure 4: The correlation between Preservation, Acquisition, Projection and average citation count, as well as pretraining occurrence. The strength of the correlation is reflected in how closely the data points cluster around the best-fit line.

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* and *Education* exhibit relatively strong performance across all three objectives.

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

First, the nature of the domain, specifically the stability or volatility of knowledge within that domain. In some domains, such as *Political Science*, knowledge is more stable, with long-established theories and principles that evolve slowly over time. In contrast, other domains, such as *Materials Science*, may experience more volatility, with frequent breakthroughs or shifting paradigms that rapidly change the state of knowledge. Knowledge preservation and projection may be more challenging in domains with higher volatility compared to those with greater stability. To test this hypothesis, we randomly retrieve 1,000 conference or journal papers published between October 2022 and September 2023 in each domain, and calculate the average citation count for these papers (Appendix G), 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 [31] to count the occurrence of these tokens in the pretraining corpus Dolma-v1.7 [48], and use the average occurrence of domain-specific tokens as a proxy for the availability of domain knowledge in the pretraining data. A complete list of specialized tokens and average occurrences across domains is provided in Appendix G.

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 coefficient [42] between average Citation Count and Knowledge Preservation is -0.709, while its correlation with Knowledge Projection is -0.736, both indicating a significant relationship. Highly dynamic domains with frequent updates may lead to more knowledge conflicts [58], making preservation and projection particularly challenging. In contrast, the correlation between pretraining availability and the robustness of scientific knowledge updates is relatively weak, indicating that pretraining alone may have a limited impact on how well models adapt to evolving scientific information.

4.2 Scientific LLMs

Rather than relying solely on off-the-shelf LLMs, researchers also use scientific LLMs [62], LLMs specifically trained or adapted for science. In this work, we also experiment with HoneyBee [49], 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

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 3: 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.

Materials Science is significantly lower than other domains (Section 4.1), we wonder if applying on a specialized scientific LLM could help. As shown in Table 3, 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.

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 [30, 34, 17, 44, 22, 2, 47, 50]. Researchers utilize off-the-shelf LLMs [3], domain-specific scientific LLMs [62], or LLMs augmented with external resources [5] to assist in scientific research. Studies show that current LLMs can be useful across various stages of the research cycle [34], including literature review [17, 1], hypothesis proposing [44], idea generation [47], and experiment planning and implementation [22, 19]. However, to the best of our knowledge, we are the first to explore whether LLMs can effectively stay up to date with evolving scientific fields while remaining reliable and useful. Specifically, we examine whether LLMs can 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 [29, 51, 55] in three key aspects. First, most prior work primarily regards the effective incorporation of new information as the only objective [20, 41, 24, 61, 53, 60, 63, 21], with few studies also considering the preservation of old knowledge [24, 61]. However, they rely on generic benchmarks such as Natural Questions [27] and CommonsenseQA [52], 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 [44], and the creation of novel ideas [47], which are the key future applications of AI in science. Second, existing approaches heavily rely on Wikipedia as a data source and assess knowledge at the factoid level (e.g., names, locations) [20, 41, 24, 61, 53, 60, 63, 21], 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 [63, 60, 21], aiming to align knowledge to specific timestamps, such as associating a president with a particular year, our goal, in scientific domains, is to align scientific claims with scientific literature.

Furthermore, we distinguish our evaluation of knowledge preservation from Catastrophic Forgetting (CF) in Continual Learning (CL) [25, 20, 9], as our setting involves multiple training stages and methods. Another relevant line of work is knowledge editing [36, 57, 35, 62, 33, 56, 23, 32, 15], 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.

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.

Acknowledgments

References

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

[2] Sangzin Ahn. 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*, 2024.

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

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

[5] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024.

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

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

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

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

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

[11] 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. When one llm drools, multi-llm collaboration rules. *arXiv preprint arXiv:2403.17852*, 2024.

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

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

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

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

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

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

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

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

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

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

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

[23] Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, et al. Learning to edit: Aligning llms with knowledge editing. *arXiv preprint arXiv:2402.11905*, 2024.

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

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

[26] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

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

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

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

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

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

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

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

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

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

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

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

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

[39] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

[40] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

[42] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

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

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

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

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

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

[48] 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, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*, 2024.

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

[50] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

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

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

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

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

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

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

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

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

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

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

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

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

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

A 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/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 [20] 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.

B 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.

C Dataset Details

C.1 Date Cutoffs

Table 4 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.

C.2 Synthetic Claims

To generate synthetic claims for each paper, we prompt GPT-4O with the instructions detailed in Table 5. We explored various alternative methods 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).

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

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 5: Prompt templates for synthetic claim generation.

We further conduct an expert evaluation of our synthetic claims. We invite two PhD students in *Computer Science* and two PhD students in *Biology*. Each student is assigned 30 papers within their respective domain 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 6, showing that at least 80% of claims strictly adhere to the rule, while more than 95% broadly meet our expectations. While the results demonstrate the effectiveness of our synthetic claims, there is still room for improvement, so we collect author-annotated claims in Section C.3.

	Computer Science		Biology	
	SUPPORT	REFUTE	SUPPORT	REFUTE
Uniquely Supported	83.3%	0.0%	80.0%	0.0%
Broadly Supported	13.3%	16.7%	15.0%	11.7%
Not Supported	3.3%	83.3%	5.0%	88.3%

Table 6: Expert evaluation results on synthetic claims, averaged across two experts per domain.

C.3 Author-annotated Claims

We argue that the original authors of research papers possess the most appropriate expertise for claim annotation. Under budget constraints, we conducted a randomized survey of 50 computer science researchers, requesting annotations of claims from their own publications. This process yielded 284 scientific claims (142 SUPPORT and 142 REFUTE claims) derived from 142 papers spanning various publication dates. While these papers do not necessarily share citation relationships, we consider them conceptually related as they all belong to the AI subfield of Computer Science.

Our evaluation using LLaMA-8B on this author-annotated dataset (Table 7) 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.

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	89.3	3.7	7.0	33.3	26.5	40.2	20.9	43.0

Table 7: 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.

D   Claim Judgment and Generation Tasks

We present the prompts used for the claim judgment and generation tasks in Table 8 and Table 9.

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 8: Prompt templates for Claim Judgment Task.

E   Metrics

Table 10 provides the detailed mathematical definitions of knowledge preservation, knowledge acquisition, and knowledge projection, as well as `distortion` and `loss`.

<div style="border:1px solid;">

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.

</div>

<div style="border:1px solid;">

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.

</div>

Table 9: Prompt templates for Claim Generation Task.

$$\textbf{Knowledge Preservation} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{prior}}^i) = correct \mid g(LM, p_{\text{prior}}^i) = correct, g(LM, p_{\text{new}}^i) = unknown))}{\sum_i \mathbb{I}(g(LM, p_{\text{prior}}^i) = correct, g(LM, p_{\text{new}}^i) = unknown)}$$

$$\texttt{distortion in Preservation} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{prior}}^i) = incorrect \mid g(LM, p_{\text{prior}}^i) = correct, g(LM, p_{\text{new}}^i) = unknown))}{\sum_i \mathbb{I}(g(LM, p_{\text{prior}}^i) = correct, g(LM, p_{\text{new}}^i) = unknown))}$$

$$\texttt{loss in Preservation} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{prior}}^i) = unknown \mid g(LM, p_{\text{prior}}^i) = correct, g(LM, p_{\text{new}}^i) = unknown))}{\sum_i \mathbb{I}(g(LM, p_{\text{prior}}^i) = correct, g(LM, p_{\text{new}}^i) = unknown))}$$

$$\textbf{Knowledge Acquisition} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{new}}^i) = correct \mid g(LM, p_{\text{new}}^i) = unknown)}{\sum_i \mathbb{I}(g(LM, p_{\text{new}}^i) = unknown)}$$

$$\texttt{distortion in Acquisition} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{new}}^i) = incorrect \mid g(LM, p_{\text{new}}^i) = unknown)}{\sum_i \mathbb{I}(g(LM, p_{\text{new}}^i) = unknown)}$$

$$\texttt{loss in Acquisition} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{new}}^i) = unknown \mid g(LM, p_{\text{new}}^i) = unknown)}{\sum_i \mathbb{I}(g(LM, p_{\text{new}}^i) = unknown)}$$

$$\textbf{Knowledge Projection} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{future}}^i) = correct \mid g(LM, p_{\text{future}}^i) = unknown, g(LM, p_{\text{new}}^i) = unknown))}{\sum_i \mathbb{I}(g(LM, p_{\text{future}}^i) = unknown, g(LM, p_{\text{new}}^i) = unknown))}$$

$$\texttt{loss in Projection} = \frac{\sum_i \mathbb{I}(g(LM_{f(\mathcal{P}_{\text{new}}^{\text{test}})}, p_{\text{future}}^i) = unknown \mid g(LM, p_{\text{future}}^i) = unknown, g(LM, p_{\text{new}}^i) = unknown))}{\sum_i \mathbb{I}(g(LM, p_{\text{future}}^i) = unknown, g(LM, p_{\text{new}}^i) = unknown))}$$

Table 10: Detailed formulations of evaluation metrics introduced in Section 2.3. Note that p_{prior}^i and p_{future}^i are considered only if p_{new}^i is *unknown* to the model before knowledge updates, as otherwise no new scientific knowledge is introduced.

F Experiment Details

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

G Cross-domain Analysis

This section provides details on cross-domain analysis discussed in Section 4.1.

G.1 Citation Count

See Table 11 for the average citation count across ten scientific domains.

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

G.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 12 and Table 13.

G.3 Occurrence in Pretraining Corpus

Refer to Table 14 for the average occurrence of domain-specific tokens in the pretraining data.

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, wave-form, 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, To-gether, 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, preserva-tion, extends, intermediate]
Medicine	[logic, outputs, AY, applic, subt, ropical, underscores, gam, publicly, overlapping, warrant, intimate, website, Domestic, Violence, DV, item, DV, observation, attainment, jun, forego-ing, 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, car-bohydrate, -chain, aur, -en, rich, chicken, iated, recipients, misconception, abandonment, sustaining, optim, -ag, Enhanced, -k, Da, property, aiding, TJ, Adopt, -condition, polariza-tion, 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 12: Specialized tokens by domains.

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 13: 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 14: The average number of occurrence of domain-specific tokens (as identified in Table 12 and Table 13) in the Dolma-v1.7 [48] pretraining corpus.