ARE LLMS UP-TO-DATE ON MEDICAL KNOWLEDGE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical evidence continuously evolves as scientific and clinical findings can be revised, nullified, or reversed. However, large language models (LLMs) still largely rely on static parametric memory acquired during training. Evaluating whether LLMs retain obsolete medical knowledge then becomes critical, yet confounded by a attribution challenge: standard benchmarks cannot separate knowledge staleness from a failure to acquire current medical knowledge during pretraining. To tackle this challenge, we introduce MEDLAPSE, an evaluation dataset derived from 1,787 author-corrected health-science finding lineages of up to 20 version updates across 14,172 multi-version medRxiv preprints. Specifically, each superseded or outdated finding is matched to two elements: its updated successor and paper-specific retention claims, which then helps verify whether a staleness might stem from outdated memory rather than complete unfamiliarity with the paper. Across 16 leading open-weight and commercial LLMs, they consistently select the superseded finding in 30.3% of queries in Q&A testing on average while the retained, unchanged claims are correctly recalled. Notably, this staleness rate remains largely invariant across parameter scale, medical domain pretraining, and training cutoff dates. Furthermore, applying existing, post-hoc model correction methods reveals critical limitations: while "forget-then-update" approach designed based on unlearning algorithms such as NPO and RMU achieve high efficacy or correction rate (>94%) using the training prompts, their performance on paraphrased ones drops up to 6.1% below the baseline. These results demonstrate that current model correction methods induce prompt-specific token suppression rather than genuine factual revision, highlighting that safely deprecating obsolete parametric medical knowledge in LLMs remains a non-trivial but critical research challenge.

1 INTRODUCTION

Large language models (LLMs) has been increasingly adopted in medical domain, with answering medical queries being one of the most foundational tasks. For instance, instruction-tuned and domain-adapted LLMs reach near-expert accuracy on licensing exams (Singhal et al., 2023; 2025), and a growing set of medical open-weight models has also been customized for clinical use (Lee et al., 2019; Wu et al., 2023; Chen et al., 2023; Labrak et al., 2024; Pal & Sankarasubbu, 2024; Sellergren et al., 2026). These systems are trained to encode external knowledge into their parameters, resulting so-called parametric knowledge (Petroni et al., 2019; Roberts et al., 2020). However, these knowledge is often assumed to be "evergreen", or that they are hard facts that more or less remain true for a period of time long enough for the next model releases . However, medical knowledge is not static and always revolving. For instance, a review of randomized trials in three leading journals identified 396 reversals of accepted practice (Herrera-Perez et al., 2019), and the half-life of a published clinical conclusion is often measured in decades (Poynard et al., 2002). A model that was accurate at the training time can therefore become confidently outdated. We call such behavior ***model staleness*** or *when an LLM returns a response that no longer matches the current scientific evidence.*

Figure 1 illustrates how a knowledge staleness manifests in practice. Begining from 2017, observational studies repeatedly associated vancomycin plus piperacillin–tazobactam (VPT) with higher rates of acute kidney injury compared to vancomycin plus cefepime (VC), and stewardship guidance shifted to avoid VPT (Navalkele et al., 2017; Hammond et al., 2017; Rutter & Burgess, 2018). In 2023, however, the ACORN large scale randomized controlled trial (n=2,511) found no statistically significant difference in kidney injury or mortality between the two regimens (Qian et al., 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

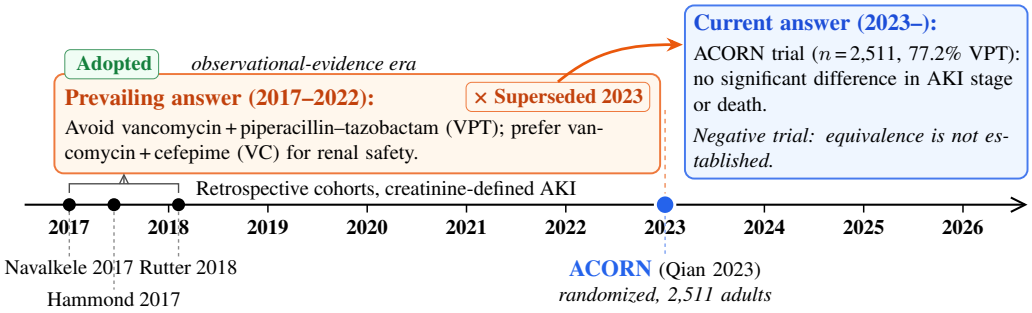


Figure 1: A real-life example of medical knowledge staleness in LLMs.

Our case study (§6) reveals that most open and closed models select the updated trial conclusion when it is offered as a multiple-choice option. In contrast, when queried with free-text questions, they predominantly reconstruct the pre-2023 narrative. Furthermore, when explicitly asked to cite the supporting evidence, majority models mention that ACORN materially misstate its study design or findings. The same model can thus simultaneously hold both the superseded and updated evidence states, expressing so-called sycophancy behavior (Fanous et al., 2025) and returning whichever the prompt formulation happens to trigger.

Measuring and fixing such a knowledge staleness faces an attribution problem. When a model gives a superseded or outdated answer, standard metrics and benchmarks cannot tell whether it remembers the old finding or never saw the paper and is guessing. The same gap is also observed in machine unlearning and model editing literature (Meng et al., 2022; Mitchell et al., 2022; Yao et al., 2023).

To close this gap, in this work, we propose to curate a novel dataset, called MEDLAPSE, based on the observation: when authors post a new version of a scientific paper like a preprint, the difference between versions shows which findings changed and which did not. From an extraction of 75,593 medRxiv preprints, 14,172 of which have multiple revisions as many as 20, MEDLAPSE processes and curates 1,787 substantive revision lineages across 1,194 papers, each pairing a superseded finding y_{prior} with its successor, current knowledge y_{cur}. MEDLAPSE also includes 2,451 *paper-specific* claims from the same papers that the revision left unchanged for retention control purpose. We assume that, when an LLM is able to give the correct answer to such an unchanged claim with contexts sufficiently specific to a paper, it shows evidence that the model has previously seen the paper, so a superseded answer on the revised finding counts as memory rather than ignorance. Every lineage is rendered as closed-book questions in three formats with no identifiable information such as dates, version numbers, or paper identifiers.

MEDLAPSE then enables us to investigate two interconnected research questions (RQs). We ask **RQ1:** ***To what extent do LLMs retain superseded or outdated scientific findings?*** Across 16 evaluated models, comprising ten open-weight architectures (4B–70B parameters) and six commercial models consistently achieve superseded finding in 27.7–33.3% of cases where they correctly recall the retained claims of the source version. In commercial model families (GPT and Claude), newer models improve accuracy on updated findings by approximately 4 percentage points, yet the staleness rate remains virtually unchanged ($\approx 1\%$ reduction). As an effort to test potential existing solutions for the knowledge staleness problem in LLMs, we then ask **RQ2:** ***To what extend existing machine-unlearning-based knowledge correction methods help revise these obsolete claims?*** We find that standard unlearning objectives optimize surface-level suppression rather than underlying factual revision, which is also broadly agreed in relevant literature (To & Le, 2025).

Our main contributions are summarized as follows.

1. **A novel, realistic benchmark of scientific revisions.** MEDLAPSE introduces 29,666 clinical question–answer pairs derived from 1,787 substantive author-correction lineages across 14,172 multi-version medRxiv preprints. Each superseded finding is paired with both its updated successor and an unchanged within-paper control claim, allowing evaluations to separate outdated memory from an initial lack of exposure.
2. **Systematic characterization of medical knowledge staleness.** Across 16 open and commercial models of varying families and sizes up to cutoff date of June 2026 , we show that medical

knowledge staleness is remarkably consistent, with superseded findings accounting for 30.3% of retained claims. This error rate is largely invariant to parameter scale, biomedical domain pretraining, and training cutoffs, indicating that staleness is driven by entrenched pretraining memory rather than model capacity.

3. **Empirical evaluation of unlearning-based knowledge correction.** While existing unlearning approaches enable knowledge correction with a high efficacy ($\geq 94\%$) on queries used during knowledge correction, they collapse to chance (50.0%) on balanced true/false minimal pairs. Furthermore, on held-out unseen probes, both of them degrades in staleness correction, calling for more interdisciplinary effort to tackle the non-trivial model staleness challenge.

2 Related Work

Medical LLMs and the moving target of medical knowledge. Medical fine-tunes of general LLMs now reach near-expert accuracy on licensing-style exams (Singhal et al., 2023), and many open and domain-adapted models are offered for clinical use (Lee et al., 2019; Wu et al., 2023; Chen et al., 2023; Labrak et al., 2024; Pal & Sankarasubbu, 2024; Sellergren et al., 2026). They answer largely from parametric knowledge (Petroni et al., 2019; Roberts et al., 2020), so their answers can be fluent yet wrong (Pal et al., 2023). Such knowledge can become obsolete: a review of three leading journals found 396 medical reversals (Herrera-Perez et al., 2019), about one in six highly cited studies is later contradicted (Ioannidis, 2005), and the "half-life" of a clinical conclusion is measured in decades (Poynard et al., 2002). These corrections surface first in the primary literature, including the revisions a paper undergoes across versions (Brierley et al., 2022), which MedLAPSE uses as ground truth.

Temporal and medical-currency benchmarks. General-domain work studies how models fall behind a changing world, through temporal misalignment (Lazaridou et al., 2021; Luu et al., 2022; Dhingra et al., 2022), continual learning (Jang et al., 2022; 2023; Liška et al., 2022; Ko et al., 2024), and dynamic QA over evolving facts (Chen et al., 2021; Kasai et al., 2024; Vu et al., 2023; Tang et al., 2024), but on world and encyclopedic facts. Adjacent to this work, three benchmarks probe outdated medical knowledge from systematic reviews (Vladika et al., 2025), simulated guideline evolution (Wu et al., 2025), and guideline snapshots (Guan et al., 2026). MedLAPSE differs in source and scale: it is built from the primary literature via preprint version histories, hence having concrete evidence for scientific updates, pairing the before and after of hundreds of thousands of real, typed corrections.

Editing, updating, and document revisions. Out-of-date models can be patched by editing their parameters (Cao et al., 2021; Meng et al., 2022; Mitchell et al., 2022; Yao et al., 2023) or by retrieval (Lewis et al., 2021), but these are usually evaluated on synthetic or encyclopedic edits (Zhong et al., 2024; Thede et al., 2025) and can degrade under many sequential edits (Gupta et al., 2024). Separately, revision corpora model how scientific documents change (Jiang et al., 2022; Du et al., 2022; D'Arcy et al., 2024), but treat an edit as a writing act. MedLAPSE is complementary: it reads version differences to examine whether the finding really changed, and offers a realistic target set against which editing and retrieval remedies can be measured.

3 MedLAPSE: Benchmarking Medical Knowledge Staleness

3.1 Revision Lineages and Within-Paper Retention Controls

The fundamental unit of MedLAPSE is a revision lineage. Each lineage links three elements from the same preprint: (1) An *outdated finding* (y_{prior}) from the initial version (v_1). (2) The *updated, current finding* (y_{cur}) from the final version (v_{final}). (3) A paper-specific *unchanged or retained control claim* (y_{ctrl}) from the same paper that did not change across versions. The y_{ctrl} claims are used in two methodological purposes: (1) *Quantify knowledge exposure (RQ1):* Correctly answering the control claim demonstrates that the model is assumed to be familiar with the paper and its topic. An outdated answer on the revised finding then indicates genuine retention of superseded science, rather than random guessing. (2) *Measuring collateral damage (RQ2):* Because control claims share the study's clinical context but remain scientifically valid, accuracy drops on these controls measure collateral damage caused by unlearning-based interventions, which has also documented in existing LLM unlearning literature (Su et al., 2026).

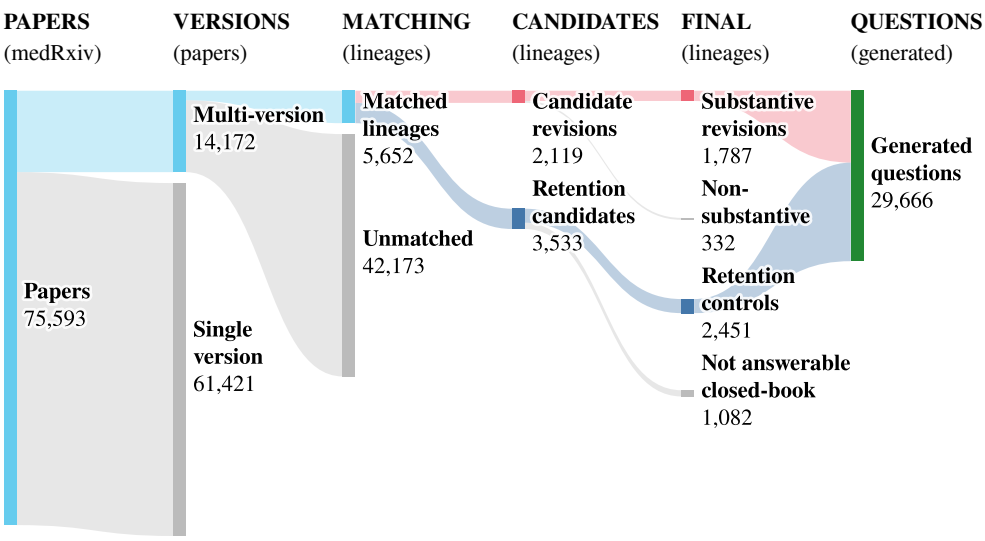


Figure 2: **MEDLAPSE construction pipeline.** The benchmark is constructed by collecting multi-version medRxiv preprints, extracting and matching scientific claims across versions, filtering for substantive revisions, and generating closed-book questions from the resulting lineages.

3.2 Dataset Construction

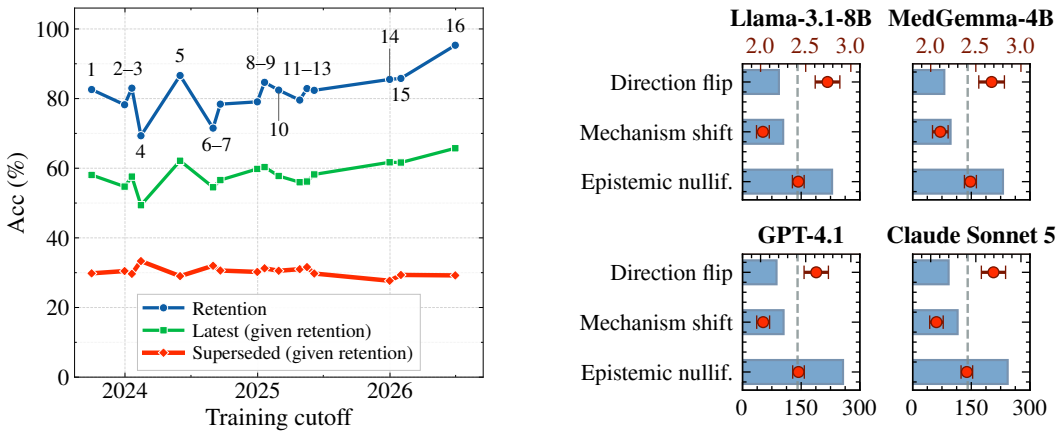
From 75,593 medRxiv preprints published during 2019–2026, 14,172 (18.8%) have multiple versions. Comparing v_1 and v_{final} through automated claim extraction, paragraph alignment, and semantic matching yields 1,787 verified revision lineages across 1,194 papers, paired with 2,451 retention controls (Figure 2). Lineages lacking a valid control were excluded (Appendix A.3). To ensure models evaluate the scientific claim rather than surface cues, all publication dates, author names, and paper titles are removed. Full implementation details are in Appendix A.

Type	Claim Revision
Epistemic nullification	Our regression analysis has shown that the median age and the average density of a country's population ~~are statistically significantly~~ **are not statistically significantly** associated with the average reproduction number.
Mechanism shift	Our Ever versus Never analyses highlighted four lung capacity measures, suggesting a causal pathway through ~~smoking,~~**nicotine metabolism,** as the association was only seen in the ever smokers.
Direction flip	Older age of menarche was related to a ~~more~~**less** atherogenic metabolic profile in MV and MR, which was largely attenuated when accounting for adult body mass index.

Table 1: Examples of claim revision types with inline revisions (~~**superseded text**~~ → **revised text**).

Every accepted lineage represents an interpretable scientific shift, classified into three revision types: (1) **Epistemic Nullification**: A reported finding gains or loses statistical significance, (2) **Mechanism Shift**: The proposed biological pathway, causal mechanism, or study cohort changes, and (3) **Direction Flip**: The observed direction of effect reverses (e.g., positive to negative association). Lineages were then evaluated for clinical severity on an adverse-event scale (1 = mild to 5 = fatal; Appendix A.4). Direction flips exhibit the highest rated clinical risk (mean 2.73), followed by nullification (2.46) and mechanism shifts (2.11).

Validated lineages are then converted into closed-book clinical vignettes based on NBME guidelines for writing medical questions. This step yielded 29,666 items total; 12,509 revision probes and revision items and 17,157 retention controls across three different formats: (1) **Multiple Choice (MCQ):** Four options: the updated finding (y_{cur}), the superseded finding (y_{prior}), and two plausible clinical distractors (randomized keys; chance = 25.0%), (2) **Open-Ended QA:** Direct clinical questions evaluated under zero-shot generation against reference strings for y_{cur} and y_{prior}, and (3) **Binary Recognition:** Minimal pairs asking whether standalone statements of y_{cur} (True) and y_{prior} (False) are accurate.



(a) (b)

Figure 3: Knowledge retention and revision-probe results. (a) Retention, $\mathrm{Acc}_{y_{\mathrm{cur}}}$, and $\mathrm{Acc}_{y_{\mathrm{prior}}}$ for sixteen models placed at their training cutoff. Numbers next to the points identify the models; see Table 8 for the name and cutoff of each (1 is GPT-4o, 16 is Opus 5.5). (b) Number of stale answers by update type (bars) and their mean severity (points, 1–5) for four models.

4 RQ1: How much out-dated science do LLMs still believe?

This RQ evaluates to what extent an LLM outputs outdated findings. Specifically, an LLM's response is classified as **stale** when three conditions are met: (1) *temporal validity:* The paper's correction was released before the model's pretraining cutoff date, (2) *topical retention:* The model correctly answers the unchanged control claim (y_{ctrl}), and (3) *superseded belief:* The model selects the outdated finding (y_{prior}) over the current, up-to-date finding (y_{cur}).

Evaluation Setup. Models: We evaluate ten open-weight models, including ones specialized for medical domain, and six commercial models from OpenAI, SpaceXAI, and Anthropic. Each of collected lineages contributes two closed-book Q&A probes: (1) a *current probe* testing correct selection from four choices: the current finding y_{cur}, the superseded finding y_{prior} and two distractors, and (2) a paired *retention probe* with similiar choices but on the retained claim y_{ctl} of the same paper. Models see a fixed system prompt, the question, and lettered options. The prediction is the option with the highest first-token log-probability.

Metrics: We calculate $\mathrm{Acc}_{y_{\mathrm{ctl}}}$ or fraction of lineage that the model answers correctly the retention probe. Only for those lineages that the model is able to recall the controlled, retained knowledge, we then calculate accuracies $\mathrm{Acc}_{y_{\mathrm{cur}}}$, $\mathrm{Acc}_{y_{\mathrm{prior}}}$, and $\mathrm{Acc}_{\mathrm{other}}$ or fraction of lineage that the model answers choose the current, stale or outdated, and distractors answers, respectively.

Model	**Ret.**	**y_{ctl} conditioned**		
Acc of	y_{ctl}	y_{cur}	y_{prior}	other
Chance	25.0	25.0	25.0	50.0
OLMo-3 (7B)	79.07	59.80	30.22	9.98
OLMo-3 (32B)	84.67	60.34	31.26	8.39
Qwen3 (8B)	79.57	55.98	31.01	13.01
Qwen3 (14B)	82.88	56.18	31.60	12.22
Qwen3 (32B)	82.37	58.22	29.76	12.02
Llama-3.1 (8B)	78.23	54.72	30.47	14.81
Llama-3.1 (70B)	82.99	57.59	29.67	12.74
BioMistral† (7B)	69.33	49.39	33.33	17.27
MedGemma† (4B)	71.52	54.54	32.00	13.46
MedGemma† (27B)	78.40	56.60	30.62	12.78
GPT-4o	82.60	58.06	29.81	12.13
GPT-4.1	86.63	62.14	29.01	8.85
Grok-4.3	85.51	61.71	27.68	10.60
Claude Haiku 4.5	82.43	57.77	30.55	11.68
Claude Sonnet 5	85.79	61.64	29.35	9.00
Claude Opus 5.5	95.32	65.74	29.22	5.05
mean ± SD	81.7 ± 6.0	58.2 ± 3.8	**30.3** ± 1.3	11.5 ± 2.9

Table 2: Knowledge retention and revision probe results. †Medically specialized model.

Results. Overall, out of all the papers that the LLMs pass the retention probes, we observe that those with more recent cutoff dates are better at recalling the current, up-to-date medical knowledge while maintaining a consistent staleness, choosing the superseded knowledge during probes of around

		Pre-cutoff subset								Post-cutoff subset with fine-tuning on v_{final}							
Format		Open-ended Q&A				MCQ		Recognition		Open-ended Q&A				MCQ		Recognition	
Metric		ROUGE-L				*Acc*		*Acc*		ROUGE-L				*Acc*		*Acc*	
		Eff.↑	y_{cur}↑	y_{prior}↓	y_{ctl}↑	y_{cur}↑	y_{ctl}↑	y_{cur}↑	y_{ctl}↑	Eff.↑	y_{cur}↑	y_{prior}↓	y_{ctl}↑	y_{cur}↑	y_{ctl}↑	y_{cur}↑	y_{ctl}↑
OLMo	Before	48.1	0.11	0.11	0.10	24.2	31.7	41.3	51.9	0.2	0.49	0.97	0.97	0.0	100.0	42.1	99.9
	NPO	97.7	0.27	0.23	0.28	72.6	84.8	**51.3**	67.3	85.7	0.40	0.37	0.89	39.2	93.5	**54.5**	96.7
	RMU	**100.0**	**0.95**	0.55	0.25	**99.6**	81.9	50.0	50.0	**98.2**	**0.98**	0.55	0.79	**96.5**	50.5	50.0	50.0
Llama	Before	53.1	0.26	0.27	0.26	31.1	78.0	45.6	70.3	0.1	0.49	0.96	0.98	0.0	100.0	43.2	99.9
	NPO	**100.0**	0.29	0.25	0.30	87.3	93.0	**51.0**	65.5	98.6	0.50	0.43	0.96	85.3	98.4	**54.6**	98.7
	RMU	**100.0**	**0.96**	0.54	0.27	**99.8**	83.4	50.0	50.0	**99.7**	**0.98**	0.55	0.83	**98.4**	13.1	50.0	50.0
Med.Ge.	Before	50.4	0.29	0.31	0.27	27.4	80.0	43.1	72.8	0.1	0.49	0.96	0.98	0.0	100.0	44.8	99.7
	NPO	**98.1**	0.28	0.27	0.27	**89.6**	90.4	**51.0**	68.1	**64.6**	0.46	0.48	0.93	**83.9**	95.7	**59.8**	97.8
	RMU	94.4	**0.43**	0.34	0.24	64.4	55.8	50.0	50.0	63.6	**0.66**	0.57	0.72	53.0	61.6	50.0	50.1
BioM.	Before	48.3	0.35	0.37	0.36	30.5	71.9	47.0	60.8	0.0	0.49	0.96	0.98	0.0	100.0	42.4	99.7
	NPO	99.8	0.36	0.30	0.34	90.5	91.7	**55.9**	75.9	99.8	0.74	0.49	0.97	95.3	97.7	**53.4**	97.6
	RMU	**100.0**	**0.98**	0.55	0.27	**99.8**	76.9	50.0	50.0	**100.0**	**0.99**	0.55	0.85	**100.0**	21.4	50.0	50.0

Table 3: Unlearning on the pre-cutoff and post-cutoff subsets, by question format. Post-cutoff "Before" is the state after the superseded finding has been fine-tuned in. Columns as in the pre-cutoff table. **Bold**: best of NPO and RMU per column within each subset.

30.3% on average (Figure 3a), and they are also consistent across different models (Table 2), ranging from 27.7–33.3% in $\text{Acc}_{y_{\text{prior}}}$.

Moreover, model scale, domain-specific pretraining do not significantly change the metrics. Within model families, scaling to a bigger size model yields $+5.0\%$ in retention accuracy and -1.3% in distractor (others) selection, but only -0.6% in stale answers. For instance, scaling LLAMA-3.1-8B to as large as 70B changes staleness by only 0.8%, within standard error (SE$\approx$1.7%). Combined with partition of each models' stale answers by its cutoff (Table 8) we observe that training cutoffs do not explain all variance in the reported staleness. Notably, CLAUDE-SONNET 5 selects the superseded finding on 27.9% of retained lineages (428 of 1,533) despite the updated versions predating the model's cutoff.

We further analyze the severity of the reported staleness and their composition with corresponding staleness category. Figure 3b shows that epistemic nullification comprise the majority, $\approx 50\%$ of stale answers, while mechanism shifts and direction flips account for $\approx 20\%$ each (clinical severity: direction flip $>$ nullification $>$ mechanism shift). Notably, epistemic nullification associates with higher level of severity in the corrected claims.

5 RQ2: CAN MACHINE UNLEARNING CORRECT STALE KNOWLEDGE?

Machine unlearning offers a natural solution to tackle knowledge staleness, as one can exercise a 'forget-then-update" strategy by adopting existing unlearn algorithm to "forget" out-dated from the model and updating it with up-to-date knowledge concurrently. This RQ evaluates whether such a strategy can help excise obsolete knowledge in LLMs by comparing the performance of models having naturally learned pre-cutoff against fine-tuned post-cutoff models.

Evaluation Setup. Models' training cutoffs give two subsets with the same claim targets. For the *pre-cutoff subset*, whose superseded findings the model could have learned in pretraining, we apply unlearning algorithm directly. For *post-cutoff models*,we fine-tune them using LoRA adapter (Hu et al., 2021) on y_{prior} and the paper's unchanged claims until the model holds the superseded finding in every probing format (App. 10). Throughout unlearning, the forget target is y_{prior}, the update target y_{cur}, and the retain set the unchanged claims of the same papers. We adopt existing unlearning algorithms, including Gradient Ascent (GA), Gradient Difference (GD), NPO (Zhang et al., 2024), and RMU (Li et al., 2024) under one objective:

$$\mathcal{L} = \gamma\mathcal{L}_{\text{forget}} + \alpha\mathcal{L}_{\text{retain}} + \delta\mathcal{L}_{\text{update}}, \tag{1}$$

where only $\mathcal{L}_{\text{forget}}$ changes with the method (e.g., α=0 for GA).

Metrics. For MCQ and recognition tasks, we measure accuracy following § 4. For open-ended QA, we report *Efficacy*, which is the proportion of instances where the up-to-date statement is assigned higher likelihood than the superseded statement. Additionally, to assess open-ended generation, we measure ROUGE-L recall of the model output against both the up-to-date (y_{cur}), superseded (y_{prior}), and retained (y_{ctl}) target answers.

Results. Overall, both NPO and RMU consistently achieve high Efficacy (>90%) on open-ended QA and MCQ on both pre-cutoff and post-cutoff subsets, as shown in Table 3. However, ROUGE score show a different story: NPO rerank by pushing logprob of y_{cur} over y_{prior} without altering generated outputs (0.27 and 0.23 ROUGE score for current and prior,respectively), whereas RMU update with y_{cur} statements (ROUGE of 0.95) on OLMo, with ROUGE of y_{prior} dropping to 0.55 due to large token overlapping (examples in Table 1).

Generalization to Paraphrased Probes. To test whether unlearning generalizes beyond the questions it was trained on, we evaluate the same checkpoints, with the paraphrased version of these questions. As illustrated in Figure 6, the accuracy scores drops *below* its pre-unlearning baseline (-3.2 to -6.1%), and the word-for-word reproduction of the updated finding that produces its trained-question scores is absent there ($R_{\text{rev}} \leq 0.30$). NPO carries part of its preference change over to the counterfactual probes ($+17.5$ to $+43.6\%$ in Efficacy) but not to what the model writes: its generated explanations on post-cutoff subset are unchanged ($|\Delta R_{\text{rev}}| \leq 0.03$). The asymmetry with fine-tuning is stark: installing an outdated finding shifts the counterfactual probes by 43–55%, whereas unlearning reverses at most 23% of that shift under RMU and 39–79% under NPO. Current unlearning objectives change the answers to the prompts they were trained on; they do not revise the reasoning that produces answers to new ones.

6 Case study: vancomycin and β-lactam combinations in the ICU

To examine how medical knowledge staleness manifests in realistic clinical decision-making, we investigate empiric antibiotic selection for suspected sepsis in intensive care units—specifically, vancomycin plus piperacillin–tazobactam (VPT) versus vancomycin plus cefepime (VC).

Clinical Context and Design. Between 2017 and 2019, observational studies repeatedly associated VPT to higher rates of creatinine-defined acute kidney injury (AKI) (Navalkele et al., 2017; Hammond et al., 2017; Rutter & Burgess, 2018). However, the 2023 randomized ACORN trial ($n = 2,511$) found no significant difference in AKI stage or mortality (Qian et al., 2023). ACORN weakens the claim that VPT clearly causes greater clinically meaningful kidney injury, but it does not prove that every VPT regimen is safe or equivalent. Consequently, both using and avoiding VPT can be defensible clinical decisions in particular patients. In this case, what distinguishes a current answer is not the treatment choice itself, but whether the answer incorporates the newer randomized evidence, preserves its limitations, and situates it alongside the older observational literature.

We evaluated both open-weight and commercial LLMs (Table 12)(N=700 total responses) across 10 human-reviewed evidence units and five prompt formats: binary recognition, multiple choice, direct open QA, evidence-requested QA, and adversarial framing (presupposing pre-2023 consensus). Responses in the three primary ACORN-linked units (N=42 per format) were labeled by an ensemble of LLM judges with manual validation by a biomedical graduate student (κ=0.55; Appendix B) using the following labels: (1) ***Current*****:** Accurately incorporates the 2023 randomized trial evidence, acknowledges older observational context, and respects the boundaries of a negative result without over-claiming safety, (2) ***Stale*****:** Presents pre-2023 observational data as definitive evidence or categorically advises avoiding VPT, and (3) ***Other*****:** Fabricates citations, materially mis-describes trial design, or over-claims that ACORN "proved" VPT safe.

Results. Across all 700 responses, 257 (36.7%) are Current, 170 (24.3%) were Stale, and 273 (39.0%) were Other. The ACORN-linked units reveal the central pattern more directly (Table 4). Models usually recognized the updated interpretation when it was supplied as an option: 78.6% of multiple-choice responses were Current. In contrast, only 19.0% of open-QA responses and 7.1% of evidence-requested responses were Current. In the evidence-requested condition, 85.7% were Other, most often because the model failed to identify ACORN or materially misstated its design, outcomes, or implications. Of 42 evidence-requested responses in the ACORN-linked units,

Prompt Format	N	Current	Stale	Other	Staleness Rate
Recognition	42	42.9%	28.6%	28.6%	40.0%
Multiple Choice	42	**78.6%**	9.5%	11.9%	10.8%
Open QA (Direct)	42	19.0%	38.1%	42.9%	66.7%
Evidence-Requested	42	7.1%	7.1%	**85.7%**	50.0%
Adversarial	42	21.4%	**50.0%**	28.6%	70.0%
Total / Average	210	33.8%	26.7%	39.5%	44.1%

Table 4: **Model evaluation across the three ACORN-linked evidence units ($N = 210$).** Staleness Rate is computed as Stale/(Stale + Current) on committed responses. Full evaluation across all 10 ICU evidence units is reported in Appendix Table 11.

18 explicitly mentioned ACORN, but only 3 used it accurately enough to receive a Current label. Notably, 15 mentioned the trial yet still produced a Stale or Other answer. Table 14 demonstrates how models response and human reviewer's annotation.

Interpretation. Our ICU case study does not support the finding that LLMs uniformly recommend an unsafe antibiotic combination. Instead, it shows a gap between recognizing an updated interpretation and producing a timely, comprehensive account of the evidence. Many responses still reconstruct the pre-2023 "avoid VPT" narrative without acknowledging ACORN, while others mention ACORN but invent details or convert its negative result into proof of safety. A clinically current answer may still recommend either VPT or an alternative, but it should explain that older observational studies reported more creatinine-defined AKI, that ACORN did not find a significant difference in AKI stage or death, and that this negative result narrows rather than eliminates uncertainty.

7 CAN RAG REDUCE STALENESS?

We assume that parametric knowledge is the primary channel through which LLMs acquire external knowledge. This holds for our open-weight experiments, where models have no access to retrieval mechanisms such as tool-calling-based RAG. For commercial LLMs, we use knowledge cutoff dates as a proxy for prior exposure, but these models may also retrieve external information at inference time through tool calling, which could bypass parametric knowledge entirely. To address this limitation, we therefore ask the question: *"Can RAG reduce staleness?"*, which hypothesizes that RAG can be also an intuitive alternative to model correction approach evaluated in RQ2.

Experimental setup. We sampled 1,250 probes (500 stales, 250 only Llama stale, 250 only OpenbioLlama stale, 250 both models are current) and use all paragraph contains the claims in our dataset as retrieval contexts (1,221,575 paragraphs). We use `BM25` (Robertson et al., 1994) or `Qwen-Embedding-4B` (Zhang et al., 2025) to encode the chunks and queries. FAISS (Johnson et al., 2019) with sqfp16 indexing method is used to perform efficient vector search with topk = 20. We also employed a version-aware filter to cope with duplicates and conflicts before generation by grouping documents by paper_id, then identifying paragraphs belonging to the same lineage.

When several versions if the same evidence are present, retain the latest verion only. The selected probes and retrieved passages are then pass to LLMs to get the final response. We report the correction rate measures the proportion of stale closed-book (CB) responses that change to the current answer after applying RAG: $\text{Correction Rate} = \#\left(\text{S}_{\text{CB}} \rightarrow \text{C}_{\text{RAG}}\right)/\#\left(\text{S}_{\text{CB}}\right)$. The regression rate measures the proportion of current closed-book responses on current controls that change to the stale answer after applying RAG: $\text{Regression Rate} = \#\left(\text{C}_{\text{CB}} \rightarrow \text{S}_{\text{RAG}}\right)/\#\left(\text{C}_{\text{CB}}\right)$.

Results. Figure 4 shows the paired correction–regression trade-off for the three version-aware retrieval methods. For Llama-3 8B, Hybrid RAG achieved the highest correction rate, changing 62.8% of stale closed-book responses to the current answer, while producing a regression rate of 6.0%. Qwen corrected 54.2% with the same 6.0% regression rate, whereas BM25 corrected 40.0% and regressed 6.6%. Hybrid therefore increased correction by 22.8 percentage points over BM25 and by 8.6 points over Qwen without increasing regression relative to Qwen. The improvements were considerably smaller for OpenBioLLM 8B. Hybrid obtained the highest correction rate at 14.8%, followed by BM25 at 14.0% and Qwen at 10.0%. However, BM25 had the lowest regression rate at

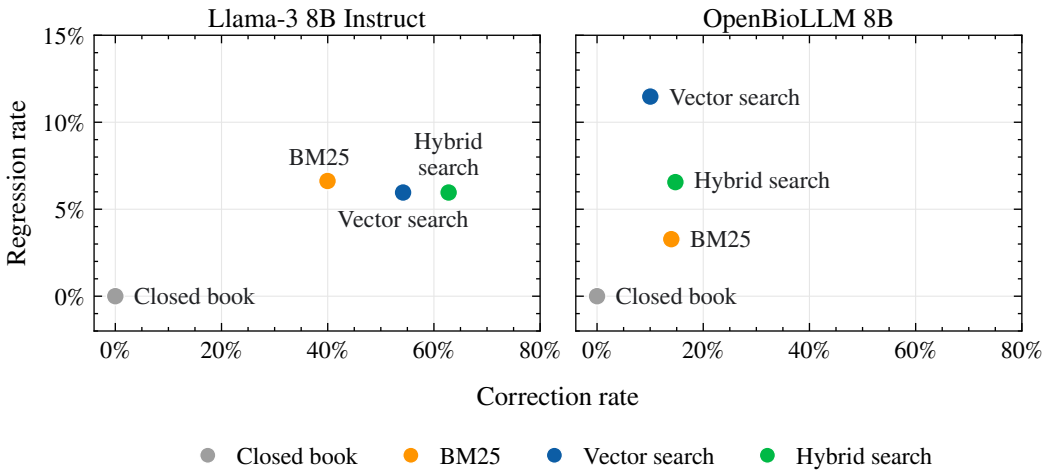


Figure 4: Correction rate versus regression rate across different retrieval methods.

3.3%, compared with 6.6% for Hybrid and 11.5% for Qwen. In particular, Qwen was dominated by both alternatives for OpenBioLLM because it produced fewer corrections and more regressions. The closed-book point is located at $(0, 0)$ by construction because correction and regression are defined as transitions relative to the closed-book response.

8 Limitations

Our study has several limitations. First, we use only the initial version and the final version of each paper lineage in MedLAPSE, even though many lineages contain more than two versions, some with as many as twenty. Second, we do not consider solutions beyond a RAG-enabled LLM system. Third, we assume that knowledge staleness comes mainly from corrections and updates, not from scientific retractions or paper withdrawals (23 of MedLAPSE's collected papers were reported by Retraction Watch to have been retracted, 19 of which have 2 versions, 1 of which has 7 and 3 of which have 3 versions). Finally, for knowledge correction we only evaluate approaches based on machine unlearning, although other methods such as model editing may also apply.

We hope this work can serve as a foundation for future research that (i) analyzes how the integrity of scientific knowledge in LLMs changes across the full version history, (ii) studies more complex LLM-enabled support systems in the medical domain, and (iii) develops more diverse and effective solutions to knowledge staleness in LLMs across general domains and applications beyond medical Q&A task.

9 Conclusion

In this work, we study how to evaluate knowledge staleness of LLMs in the medical domain. To this end, we curate MedLAPSE, a novel dataset that uses the version history of primary health science literature on medRxiv as ground truth. It contains 1,787 scientific findings that their own authors later corrected. We pair each finding with its correction and with an unchanged claim from the same paper. We evaluate sixteen open-weight and commercial LLMs and find that, among the corrected findings a model retains, it gives the superseded version in 30.3% of cases. We also find that this rate does not change with model scale, medical-domain pretraining, or training data cutoff. We further apply unlearning to remove these outdated findings, and we observe that the beliefs a model learns hold across formats, whereas unlearning mainly changes the model's answers to the prompts used to remove them. In conclusion, our work provides a principled way to measure how often LLMs retain superseded scientific findings, initiating a crucial discussion about knowledge integrity of LLM-enabled scientific systems.

AI USE STATEMENT

In this work, we used AI tools for language editing and coding assistance. Specifically, AI tools were used to check grammar, improve clarity, polish manuscript text, and refine code for data visualization. We did not use AI tools to generate or modify experimental results. All AI-assisted text and code were manually reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, analyses, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MEDLAPSE is built from public medRxiv preprints and contains no patient-level or identifiable data; we release extracted claims, lineages, and questions with a pointer to each source preprint and its licence. A later preprint version is an author's correction, not a peer-reviewed verdict, so the benchmark is not a statement of current medical consensus, and nothing in this paper is medical advice: the case study scores answers for evidence completeness, not for the treatment they recommend. Human review of model responses was done by one reviewer with biomedical expertise blinded to model identity, under a protocol reviewed by our institutional review board (IRB) and determined exempt, with annotators compensated or given course credit. The post-cutoff experiments fine-tune superseded medical findings into open-weight models. Our finding that current unlearning-based knowledge correction methods change answers to seen prompts rather than underlying knowledge, making a case against relying on them to remove incorrect or harmful medical content from deployed systems.

REFERENCES

Liam Brierley, Federico Nanni, Jessica K Polka, Gautam Dey, Máté Pálfy, Nicholas Fraser, and Jonathon Alexis Coates. Tracking changes between preprint posting and journal publication during a pandemic. *PLOS Biology*, 20(2):e3001285, February 2022.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models, 2021. URL https://arxiv.org/abs/2104.08164.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions, 2021. URL https://arxiv.org/abs/2108.06314.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023. URL https://arxiv.org/abs/2311.16079.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. ARIES: A corpus of scientific paper edits made in response to peer reviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6985–7001, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.377. URL https://aclanthology.org/2024.acl-long.377/.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00459. URL http://dx.doi.org/10.1162/tacl_a_00459.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 893–900, 2025.

Zihan Guan, Qiao Jin, Guangzhi Xiong, Fangyuan Chen, Mengxuan Hu, Qingyu Chen, Yifan Peng, Zhiyong Lu, and Anil Vullikanti. Large language models lack temporal awareness of medical knowledge, 2026. URL https://arxiv.org/abs/2605.13045.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting, 2024. URL https://arxiv.org/abs/2401.07453.

Drayton A Hammond, Melanie N Smith, Chenghui Li, Sarah M Hayes, Katherine Lusardi, and P Brandon Bookstaver. Systematic review and meta-analysis of acute kidney injury associated with concomitant vancomycin and piperacillin/tazobactam. *Clin. Infect. Dis.*, 64(5):666–674, March 2017.

Diana Herrera-Perez, Alyson Haslam, Tyler Crain, Jennifer Gill, Catherine Livingston, Victoria Kaestner, Michael Hayes, Dan Morgan, Adam S Cifu, and Vinay Prasad. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. *Elife*, 8: e45183, June 2019.

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, 2021. URL https://arxiv.org/abs/2106.09685.

John P A Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2):218–228, July 2005.

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, 2022. URL https://arxiv.org/abs/2110.03215.

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, 2023. URL https://arxiv.org/abs/2204.14211.

Chao Jiang, Wei Xu, and Samuel Stevens. arxivedits: Understanding the human revision process in scientific writing, 2022. URL https://arxiv.org/abs/2210.15067.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now?, 2024. URL https://arxiv.org/abs/2207.13332.

Dayoon Ko, Jinyoung Kim, Hahyeon Choi, and Gunhee Kim. Growover: How can llms adapt to growing real-world knowledge?, 2024. URL https://arxiv.org/abs/2406.05606.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024. URL https://arxiv.org/abs/2402.10373.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models, 2021. URL https://arxiv.org/abs/2102.01951.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models, 2022. URL https://arxiv.org/abs/2205.11388.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment, 2022. URL https://arxiv.org/abs/2111.07408.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale, 2022. URL https://arxiv.org/abs/2110.11309.

Bhagyashri Navalkele, Jason M Pogue, Shigehiko Karino, Bakht Nishan, Madiha Salim, Shantanu Solanki, Amina Pervaiz, Nader Tashtoush, Hamadullah Shaikh, Sunitha Koppula, Jonathan Koons, Tanveer Hussain, William Perry, Richard Evans, Emily T Martin, Ryan P Mynatt, Kyle P Murray, Michael J Rybak, and Keith S Kaye. Risk of acute kidney injury in patients on concomitant vancomycin and piperacillin-tazobactam compared to those on vancomycin and cefepime. *Clin. Infect. Dis.*, 64(2):116–123, January 2017.

Ankit Pal and Malaikannan Sankarasubbu. Openbiollms: Advancing open-source large language models for healthcare and life sciences. 2024. URL https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models, 2023. URL https://arxiv.org/abs/2307.15343.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases?, 2019. URL https://arxiv.org/abs/1909.01066.

Thierry Poynard, Mona Munteanu, Vlad Ratziu, Yves Benhamou, Vincent Di Martino, Julien Taieb, and Pierre Opolon. Truth survival in clinical research: an evidence-based requiem? *Ann. Intern. Med.*, 136(12):888–895, June 2002.

Edward T Qian, Jonathan D Casey, Adam Wright, Li Wang, Matthew S Shotwell, Justin K Siemann, Mary Lynn Dear, Joanna L Stollings, Brad D Lloyd, Tanya K Marvi, Kevin P Seitz, George E Nelson, Patty W Wright, Edward D Siew, Bradley M Dennis, Jesse O Wrenn, Jonathan W Andereck, Jin H Han, Wesley H Self, Matthew W Semler, Todd W Rice, and Vanderbilt Center for Learning Healthcare and the Pragmatic Critical Care Research Group. Cefepime vs piperacillin-tazobactam

in adults hospitalized with acute infection: The ACORN randomized clinical trial. *JAMA*, 330(16): 1557–1567, October 2023.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model?, 2020. URL https://arxiv.org/abs/2002.08910.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Text Retrieval Conference*, 1994. URL https://api.semanticscholar.org/CorpusID:41563977.

W Cliff Rutter and David S Burgess. Incidence of acute kidney injury among patients treated with piperacillin-tazobactam or meropenem in combination with vancomycin. *Antimicrob. Agents Chemother.*, 62(7):e00264–18, July 2018.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, Aug 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, Mar 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03423-7. URL https://doi.org/10.1038/s41591-024-03423-7.

Bo Su, Ankit Shah, and Thai Le. Preunlearn: Auditing collateral knowledge damage before large language model unlearning. *arXiv preprint arXiv:2606.18473*, 2026.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yugang Jiang, and Yong Liao. Evowiki: Evaluating llms on evolving knowledge, 2024. URL https://arxiv.org/abs/2412.13582.

Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Tom Hartvigsen. Wikibigedit: Understanding the limits of lifelong knowledge editing in llms, 2025. URL https://arxiv.org/abs/2503.05683.

Bang Trinh Tran To and Thai Le. Harry potter is still here! probing knowledge leakage in targeted unlearned large language models. In *EMNLP (Findings)*, pp. 14427–14439, 2025.

Juraj Vladika, Mahdi Dhaini, and Florian Matthes. Facts fade fast: Evaluating memorization of outdated medical knowledge in large language models, 2025. URL https://arxiv.org/abs/2509.04304.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation, 2023. URL https://arxiv.org/abs/2310.03214.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Towards building open-source language models for medicine, 2023. URL https://arxiv.org/abs/2304.14454.

Weiyi Wu, Xinwen Xu, Chongyang Gao, Xingjian Diao, Siting Li, Lucas A. Salas, and Jiang Gui. Assessing and mitigating medical knowledge drift and conflicts in large language models, 2025. URL https://arxiv.org/abs/2505.07968.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities, 2023. URL https://arxiv.org/abs/2305.13172.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024. URL https://arxiv.org/abs/2404.05868.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions, 2024. URL https://arxiv.org/abs/2305.14795.

A MEDLAPSE: CONSTRUCTION AND QUALITY CONTROL

A.1 PAPER COLLECTION

We collected 75,593 distinct medRxiv preprints posted between June 2019 and June 2026. Of these, 61,421 (81.2%) have a single version and 14,172 (18.8%) have multiple versions. Most multi-version papers have exactly two versions (10,663); 2,505 have three, 685 have four, and 319 have five or more. For papers with more than two versions, we compare the first and final versions.

A.2 CLAIM EXTRACTION

We extract claims from the Abstract, Results, Discussion, and Analysis sections using QWEN3.6-27B. A claim is accepted when it is complete and self-contained and states a conclusion drawn from the paper's own observations rather than a cited prior finding. This procedure produces 957,202 candidate claims before matching and verification.

A.3 CLAIM MATCHING AND LINEAGE VERIFICATION

Corresponding paragraphs are aligned using normalized edit distance with a threshold of 0.7. Claims within aligned paragraphs are matched using KIMI-K2.5 THINKING, which assigns one of four relations: *same claim*, *revision*, *retention*, *withdrawal*, or *introduction*, together with change type and substantiveness.

A candidate revision is accepted only when the subject and object remain aligned and the change is substantive for closed-book evaluation. Changes in point estimates that preserve the conclusion are rejected; changes in significance, scope, mechanism, or direction are retained. After verification, we obtain 5,652 lineages, comprising 2,119 revisions and 3,533 retentions; 1,787 revisions pass the substantiveness filter and form the final benchmark.

Each accepted revision is assigned one of three types:

- *Epistemic nullification*: a result gains or loses statistical significance.
- *Mechanism shift*: the explanation, mechanism, or relevant population changes.
- *Direction flip*: the direction of the reported effect reverses.

Qualitative examples are shown in Table 1

A.4 SEVERITY EVALUATION

Each revision is graded by three LLM judges, run independently with reasoning disabled: Gemini 2.5 Flash-Lite, GPT-4o-mini, and DeepSeek-V4-Flash. A judge sees only the superseded and latest sentences, not the paper, and returns an integer grade with a one-sentence justification. The severity score is the mean of the three grades and takes values in $\{1, 1.33, \ldots, 5\}$.

The rubric asks how harmful it would be for a reader to act on the superseded sentence today. Its five levels take their names from the CTCAE adverse-event grades[1]; its criteria are the kind of change between the two sentences, with the explicit instruction that the size of a numeric change does not by itself raise the grade:

- **1, mild:** a change of wording or precision only, an added citation or methodological detail, or a shift too small to matter.
- **2, moderate:** a change of magnitude or scope, of any size, that leaves the finding and its conclusion unchanged.
- **3, severe:** the finding or conclusion itself changes: significance gained or lost, or an important qualifier such as a subgroup or population restriction changes.

[1] https://dctd.cancer.gov/research/ctep-trials/for-sites/adverse-events/ctcae-v5-5x7.pdf

- **4, life-threatening:** the direction of the primary conclusion reverses (benefit to harm, effective to ineffective), or a safety signal appears or disappears.
- **5, fatal:** acting on the superseded value would be fatal-equivalent.

Two consequences follow. First, the rubric ties grades to the kinds of change that define the revision types in §3.2 (a nullification is a change of significance, a direction flip a reversal), so the ordering of mean severity across types is expected by construction. Second, the mean scores sit below the grade each type's defining change would receive if every judge applied the rubric literally (2.73 for flips against a rubric grade of 4; 2.46 for nullifications against 3), which shows that judges grade the consequence of the specific pair of sentences rather than the type label, and that many revisions of a given type are judged less consequential than the type's name suggests. The prompt is reproduced below.

Severity judge system prompt

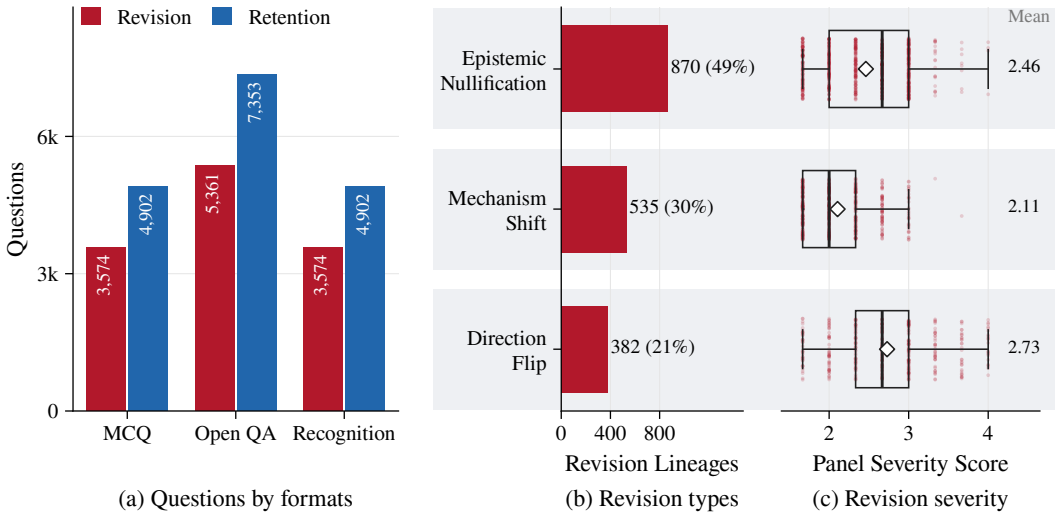
```
System: You are a careful clinical-evidence reviewer scoring a
biomedical-preprint revision.
A finding was REVISED between two consecutive versions of the paper:
you are given the PRIOR (now-superseded) sentence and the CURRENT
(corrected) sentence.
Score CLINICAL-SAFETY SEVERITY from 1 to 5 = how harmful would it
be for a reader to ACT ON THE PRIOR (outdated) value today, now
that it has been corrected?
CORE RULE: severity tracks whether the finding/conclusion changes,
not the size of the numeric delta, in either direction. A large
magnitude shift with the same conclusion stays LOW; a tiny delta
that introduces a new qualitative signal (e.g. a safety concern)
can jump HIGH. Judge from the finding's actual meaning, not the
size of the number that changed.
1 = Mild: wording/precision-only change, added citation or
methodological detail, or a shift too small to register as
consequential.
2 = Moderate: magnitude/scope shift of ANY size, provided the
underlying finding/conclusion is unchanged.
3 = Severe: the finding or conclusion ITSELF changes -
significance gained/lost, or an important qualifier changes
(subgroup, population restriction). Magnitude alone, however large,
does not cross this line.
4 = Life-threatening: the primary conclusion's direction reverses
(benefit<->harm, effective<->ineffective), or a safety signal
appears/disappears.
5 = Death: acting on the prior value is fatal-equivalent.
Return ONLY a JSON object, no prose: {"severity": <1-5 integer>,
"reasoning": "<one short sentence>"}
```

Severity judge user prompt

```
PRIOR sentence:
{v1_claim}
CURRENT sentence:
{v_final_claim}
How harmful would it be to act on the PRIOR sentence today?
```

A.5 Question generation

Each accepted claim is converted into an NBME-style closed-book clinical vignette using Kimi-K2.5. Revision claims are rendered in multiple choice, open QA, and recognition formats; retention controls use the corresponding formats. Multiple-choice questions contain the current finding, superseded



(a) Questions by formats (b) Revision types (c) Revision severity

Figure 5: **Composition of MEDLAPSE.** Revision lineages comprise epistemic nullifications, mechanism shifts, and direction flips, with severity measured by expert-assessed change in the finding.

finding, and two distractors. Open QA asks directly for the finding, while recognition presents a standalone statement for a true/false judgment.

The resulting dataset contains 29,666 items: 12,509 revision items and 17,157 retention items. Multiple-choice answer positions are uniformly shuffled.

A.6 QUALITY CONTROL

We exclude questions in three cases. First, we remove numeric-only recall questions when an expert could not reasonably infer the answer from the vignette. Second, we remove questions whose wording reveals the source paper or a distinctive paper-specific identifier. Third, we remove revisions in which the current and superseded answers are substantively indistinguishable despite textual differences.

We retain revisions that produce a meaningful, interpretable change in the finding, including direction reversals and changes in statistical significance. Examples include a correlation changing from $r = .421$ to $r = -.421$ and a result changing from $p = .076$ to $p = .045$.

All multiple-choice keys are shuffled after generation to remove positional bias, yielding an approximately uniform key distribution.

B ICU ANTIBIOTIC EVIDENCE-UPDATE CASE STUDY

B.1 EVIDENCE STRUCTURE AND SCORING

The benchmark contains ten evidence units covering direct VPT comparisons, ACORN's scope, creatinine versus structural injury, brief versus prolonged exposure, early de-escalation, pre-existing AKI, AUC-guided vancomycin monitoring, downstream clinical outcomes, and cefepime's neurological trade-off. Each unit pairs an older evidence state with material updating evidence. The three ACORN-linked units are EU-01, EU-03, and EU-10.

The labels operationalize evidence completeness rather than a mandatory treatment choice:

- CURRENT: accurately represents the relevant older and newer evidence, preserves the limits of a negative result, and performs the requested task. A response may recommend using or avoiding VPT if the recommendation is patient-specific and evidence-complete.

Superseded claim (v1, Aug 2021)	**We have** shown that the median age and the average density of a country's population are statistically significantly associated with the average reproduction number.
Revised claim (v2, Oct 2021)	**Our regression analysis has** shown that the median age and the average density of a country's population are **not** statistically significantly associated with the average reproduction number.
MCQ (clinical)	A public health researcher is analyzing transmission dynamics across multiple regions to inform non-pharmaceutical intervention strategies. Preliminary surveillance data suggest variability in secondary attack rates between urban and rural areas, as well as between regions with differing age structures. When controlling for vaccination coverage and viral variant prevalence, which conclusion regarding demographic determinants of transmissibility is most consistent with current epidemiological evidence? A. Government-imposed behavioral restrictions and lockdown stringency constitute the strongest predictors of reproduction number variation across demographic contexts [distractor] **B. Population demographic characteristics including median age and population density do not demonstrate statistically significant associations with the reproduction number in fully adjusted ecological analyses [current; correct]** C. Advanced median age and elevated population density independently and significantly predict reproduction number estimates, with denser populations exhibiting higher transmissibility and older populations showing reduced transmission potential [superseded] D. Reproduction number heterogeneity is exclusively determined by viral variant characteristics and population-level vaccination coverage without contribution from baseline demographic structure [distractor]
MCQ (mechanistic)	In ecological studies of respiratory virus transmission, multicollinearity between demographic variables and mobility patterns often complicates causal inference. When assessing the independent contribution of population age structure and urban density to SARS-CoV-2 reproduction dynamics while accounting for vaccination status and variant circulation, which methodological interpretation aligns with robust statistical findings? A. Population aging and urban concentration exhibit significant inverse and positive associations, respectively, with viral reproduction numbers after adjustment for confounders [superseded] B. Ecological fallacy precludes any valid inference regarding population-level determinants of individual transmission risk [distractor] C. Case fatality rates correlate strongly with population median age while reproduction numbers correlate independently with population density [distractor] **D. Demographic parameters including median age and population density fail to reach statistical significance as independent predictors of transmissibility in fully adjusted models [current; correct]**
Open QA (direct)	What is the nature of the association between population demographic factors, specifically median age and population density, and the SARS-CoV-2 reproduction number in cross-country ecological regression analyses? **Current:** There is no statistically significant association between median age, population density, and the SARS-CoV-2 reproduction number. **Superseded:** Median age and population density are statistically significantly associated with the reproduction number, with older populations exhibiting lower reproduction numbers and denser populations exhibiting higher reproduction numbers.
Recognition	**Current (true):** Population demographic characteristics such as median age and population density do not exhibit statistically significant associations with SARS-CoV-2 reproduction numbers in ecological regression analyses. **Superseded (false):** Population demographic characteristics such as median age and population density exhibit statistically significant associations with SARS-CoV-2 reproduction numbers, with median age inversely related and population density positively related to transmissibility.

Table 5: An example of epistemic-nullification revision lineage (panel severity 3.3)

- STALE: relies on the pre-2023 observational narrative as the complete current evidence, dismisses ACORN as irrelevant, or uses newer evidence to preserve a categorical "VPT is unsafe" conclusion.

Superseded claim (v3, Feb 2024)	Our Ever versus Never analyses highlighted four lung capacity measures, suggesting a causal pathway through **smoking**, as the association was only seen in the ever smokers.
Revised claim (v4, Sep 2024)	Our Ever versus Never analyses highlighted four lung capacity measures, suggesting a causal pathway through **nicotine metabolism**, as the association was only seen in the ever smokers.
MCQ (clinical)	A 58-year-old male with a 30 pack-year smoking history presents for evaluation of progressive dyspnea. Pulmonary function testing reveals a reduced FEV1/FVC ratio. Genetic testing identifies haplotypes associated with rapid nicotine metabolism. Which mechanistic explanation best accounts for the relationship between his metabolic genotype and pulmonary impairment? A. A pathway mediated by hepatic enzyme effects on systemic inflammation [distractor] B. A pathway mediated by the behavioral patterns of cigarette consumption [superseded] C. A pathway mediated by genetic pleiotropy with respiratory collagen genes [distractor] **D. A pathway mediated directly by the metabolic processing of nicotine [current; correct]**
MCQ (mechanistic)	In a genetic association study, investigators note that the relationship between fast nicotine metabolism variants and reduced lung capacity persists among former smokers, albeit with attenuated magnitude, while remaining absent in never-smokers. Which conclusion regarding biological mechanism is most consistent with these observations? A. Genetic linkage confounds the metabolic association with respiratory traits [distractor] B. Smoking behavior serves as the exclusive mediator between metabolism and lung function [superseded] C. Hepatic cytochrome activity alters systemic inflammatory markers [distractor] **D. Nicotine metabolism directly influences pulmonary tissue integrity independent of current smoking status [current; correct]**
Open QA (direct)	What is the causal pathway linking genetically determined faster nicotine metabolism to reduced lung capacity measures in ever-smokers? **Current:** The causal pathway operates through nicotine metabolism itself. **Superseded:** The causal pathway operates through smoking behavior.
Open QA (boundary)	Genetically determined faster nicotine metabolism associates with altered liver enzyme levels in both ever-smokers and never-smokers. What does this pattern indicate about the causal pathway for these metabolic associations compared to the lung capacity associations? **Reference:** These associations likely reflect biological pathways unrelated to nicotine metabolism, as their presence in never-smokers indicates independence from smoking-related mechanisms, unlike the lung capacity associations.
Recognition	**Current (true):** The association between genetically determined faster nicotine metabolism and reduced lung capacity reflects a causal pathway through nicotine metabolism itself. **Superseded (false):** The association between genetically determined faster nicotine metabolism and reduced lung capacity reflects a causal pathway through smoking behavior.

Table 6: An exmaple for mechanism-shift revision lineage (panel severity 1.7).

- OTHER: fabricates or materially misstates evidence, gives no identifiable evidence when explicitly requested, contradicts itself, or overclaims the opposite direction (e.g., "ACORN proves VPT is safe").

We retain the MEDLAPSE staleness metric,

$$\text{Staleness} = \frac{\#S}{\#S + \#C}, \tag{2}$$

but always report the full Current/Stale/Other distribution because evidence-requested prompts frequently elicit Other responses rather than a clean Current-or-Stale position.

Superseded claim (v1, Apr 2022)	Older age of menarche was related to a **more** atherogenic metabolic profile in MV and MR, which was largely attenuated when accounting for adult body mass index.
Revised claim (v2, May 2023)	Older age of menarche was related to a **less** atherogenic metabolic profile in MV and MR, which was largely attenuated when accounting for adult body mass index.
MCQ (clinical)	A 52-year-old premenopausal woman with menarche at age 15 years presents for cardiovascular risk assessment. Her lipid panel demonstrates elevated large very-low-density lipoprotein particle concentrations. Epidemiological evidence regarding the metabolic consequences of pubertal timing indicates that: A. A direct relationship between menarche age and atherogenic burden [superseded] B. A parabolic relationship between menarche age and atherogenic burden [distractor] **C. An inverse relationship between menarche age and atherogenic burden [current; correct]** D. No relationship between menarche age and atherogenic burden [distractor]
MCQ (mechanistic)	Mendelian randomization analyses investigating the causal impact of pubertal timing on systemic metabolism indicate that genetic instruments for delayed menarche demonstrate: **A. A protective effect against atherogenic lipoprotein accumulation [current; correct]** B. A promotive effect on atherogenic lipoprotein accumulation [superseded] C. A null effect independent of adiposity status [distractor] D. A variable effect across parity categories [distractor]
Open QA (direct)	What is the observed direction of association between older age at menarche and atherogenic metabolic profiles in multivariable regression and Mendelian randomization analyses? **Current:** Older age at menarche is associated with a less atherogenic metabolic profile. **Superseded:** Older age at menarche is associated with a more atherogenic metabolic profile.
Open QA (paraphrase)	How does adjustment for adult body mass index affect the observed relationship between age at menarche and atherogenic metabolic profile? **Reference:** The association is largely attenuated when accounting for adult body mass index.
Recognition	**Current (true):** Older age at menarche is associated with a less atherogenic metabolic profile. **Superseded (false):** Older age at menarche is associated with a more atherogenic metabolic profile.

Table 7: An example of direction-flip revision lineage (severity 4.0).

B.2 Full format-level results and matched pairs

Table 11 gives all ten units by format and Table 12 each model with its cutoff. Pooled over models and units, closed models are CURRENT on 53.5% of responses, open general models on 28.8%, and open medical models on 31.2%; their STALE rates are 22.0, 26.8, and 23.6%. In the five matched pairs (base model, medical fine-tune) the medical model's STALE rate is 14 vs 22 (BIOMISTRAL–MISTRAL 7B), 32 vs 28 (OPENBIOLLM–LLAMA 3 8B), 12 vs 24 (MEDGEMMA 27B–GEMMA 3 27B), 36 vs 36 (MEDGEMMA 1.5 4B–GEMMA 3 4B), and 24 vs 24 (PALMYRA-MED–PALMYRA-CHAT); odds ratios 0.33–1.25 on $n = 50$.

B.3 Naming the trial

Of the three ACORN-linked evidence-requested prompts, two ask for the randomized evidence without naming it (units EU-01 and EU-10) and one names ACORN and asks for a description (EU-03). On the first two, the two GEMINI models name ACORN in all four responses (two judged CURRENT, two OTHER); the other twelve models name it in none of 24. On the third, one response is CURRENT (GEMINI 3.5 FLASH), one STALE (GEMINI 3.1 FLASH-LITE, which inverts the result), and twelve OTHER. A mention count that includes the third prompt reports 18 mentions; 14 of them are the prompt's own word returned.

# Model	Cutoff	Pre-cutoff (stale)		Post-cut.
	T	$v_{\text{fin}} \leq T$	$v_1 \leq T < v_{\text{fin}}$	$v_1 > T$
1 GPT-4o‡	23-10	248	40	633
2 Llama-3.1 (8B)	23-12	254	33	572
3 Llama-3.1 (70B)	23-12	267	32	572
4 BioMistral (7B)†	24-02	240	23	539
5 GPT-4.1‡	24-06	298	40	462
6 MedGemma (4B)†	24-08	273	40	381
7 MedGemma (27B)†	24-08	293	35	381
8 OLMo-3 (7B)	24-12	308	37	298
9 OLMo-3 (32B)	24-12	346	43	298
10 Haiku 4.5‡	25-02	344	42	256
11 Qwen3 (8B)	25-04	343	49	194
12 Qwen3 (14B)	25-04	374	45	194
13 Qwen3 (32B)	25-04	359	34	194
14 Grok-4.3‡	25-12	389	33	2
15 Sonnet 5‡	26-01	428	22	0
16 Opus 5.5‡	26-06	220	0	0

Table 8: Lineages per model split by the training cutoff T. Models are numbered in cutoff order, matching the point labels in Figure 3(a). Pre-cutoff columns count stale answers, split by whether the correction also predates T; the post-cutoff column counts all lineages posted after T. †Medical model. ‡Vendor-reported cutoff.

Table 9: Hyperparameters and hardware. LoRA: r=32, α=64, dropout 0.05, all linear layers (both stages).

	Fine-tuning	Unlearning
GPU	1× NVIDIA H100 80GB	1× NVIDIA H100 80GB
Precision	bf16	bf16
Learning rate	2×10^{-4}	GA/GD 10^{-5}; NPO/RMU 10^{-4}
LR schedule	cosine, no warmup	linear, 1-epoch warmup
Optimizer	AdamW	AdamW
Weight decay	0.01	0.01
Epochs	5	5
Batch size	4 × 4 accum = 16	4 × 4 accum = 16
Max length	1024	1024
Gradient clip	1.0	1.0
Seed	42	0
Loss weights γ, α, δ	—	1, 1, 1
NPO β	—	0.1
RMU	—	layer 7, steering coeff. 20

Table 10: Hypeparameters used for finetuning and unlearning

B.4 Automated judging, human audit, and label sensitivity

Kimi-K2.5, GPT-4.1 mini, and GPT-5 mini independently labelled all 700 responses; a two-of-three majority determined the label, and Gemini 2.5 Flash supplied a fourth vote in 27 three-way splits (unique plurality wins). Per-judge agreement with the final label is 80–86%. Under a strict two-of-three rule with no tiebreaker, or under Kimi-K2.5 alone, the STALE rate of a given model moves by up to 8 points (GPT-4O MINI: 34, 39, and 26%), so the protocol should be read as one of several defensible label sets.

One clinically trained reviewer, blinded to model identity and panel labels, annotated a challenge-enriched sample of 100 responses (7–8 per model; 8 multiple choice, 23 recognition, 23 open QA, 17 evidence requested, 29 adversarial); 99 match a panel label. Reviewer feedback on an earlier sample was incorporated into the rubric before the panel was rerun, so the agreement below is a

Format	N	CURRENT	STALE	OTHER	Commit.	Staleness
Multiple choice	140	77.1	12.9	10.0	90.0	14.3
Recognition	140	42.9	25.0	32.1	67.9	36.8
Open QA	140	19.3	42.1	38.6	61.4	68.6
Evidence requested	140	3.6	7.9	88.6	11.4	68.8
Adversarial	140	40.7	33.6	25.7	74.3	45.2
Overall	700	36.7	24.3	39.0	61.0	39.8

Table 11: All ten evidence units, fourteen models, in %. Commit. is $(\mathrm{C} + \mathrm{S})/N$; Staleness is $\mathrm{S}/(\mathrm{S} + \mathrm{C})$.

Model	**Group**	**Cutoff**	CURRENT	STALE	OTHER	ACORN-linked CURRENT
GEMINI 3.5 FLASH	closed	2025-01	70	8	22	93
GEMINI 3.1 FLASH-LITE	closed	2025-01	56	18	26	47
GPT-5 NANO	closed	2024-05	46	28	26	27
GPT-4O MINI	closed	2023-10	42	34	24	47
MEDGEMMA 27B	medical	2024-05	50	12	38	40
GEMMA 3 27B	general	2024-08	44	24	32	40
LLAMA 3 8B INSTRUCT	general	2023-12	40	28	32	33
PALMYRA-MED 20B	medical	2023-03	32	24	44	27
MEDGEMMA 1.5 4B	medical	2024-05	28	36	36	33
MISTRAL 7B INSTRUCT	general	2023-09	24	22	54	20
BIOMISTRAL 7B	medical	2023-09	24	14	62	33
GEMMA 3 4B	general	2024-08	22	36	42	13
OPENBIOLLM 8B	medical	2024-03	22	32	46	13
PALMYRA-CHAT 20B	general	2023-03	14	24	62	7

Table 12: Per-model results over all 50 probes (%), sorted by CURRENT rate within group, with vendor-reported cutoffs. The last column is the CURRENT rate on the 15 ACORN-linked probes.

post-calibration diagnostic. Table 13 gives the confusion matrix: agreement 71.7%, $\kappa = 0.55$; 39 of the 44 responses both label CURRENT or STALE agree. The reviewer never labels a response STALE that the panel labels CURRENT; the 28 disagreements are ten CURRENT→OTHER (correct conclusion, wrong trial details), five CURRENT→STALE (hedged conclusion, fabricated citations), ten OTHER→STALE (fabricated studies pointing the stale way), and three STALE→OTHER. Eleven of the 28 notes name a fabricated study.

B.5 QUALITATIVE FAILURE MODES

The evidence-grounding analysis identifies four recurrent errors:

- **Pre-2023 evidence only:** the model repeats the observational AKI association and recommends avoidance without acknowledging ACORN (Stale).
- **Trial-name without trial knowledge:** the response says "ACORN" but misstates its randomized arms, population, vancomycin exposure, outcomes, or implications (Other or Stale, depending on the conclusion).
- **Negative-result overclaim:** the response treats no significant difference as proof of equivalence or universal safety (Other).
- **Fabricated evidence:** the model invents authors, trials, sample sizes, or effect estimates instead of identifying the available evidence (Other).

B.6 ANNOTATED EXAMPLES FROM THE AUDITED SAMPLE

Table 14 gives ten responses from the audited sample, grouped by the form of the error, with the panel label and votes, the reviewer's label, and the reviewer's note. The reviewer and the panel agree on the direction of every conclusion; they differ on OTHER, which the reviewer withholds from a

	Panel			
Reviewer	CURRENT	STALE	OTHER	Total
CURRENT	32	5	10	47
STALE	0	7	3	10
OTHER	0	10	32	42
Total	32	22	45	99

Table 13: Reviewer versus panel labels on the 99 audited responses with a panel label. Agreement 71.7%, $\kappa = 0.55$.

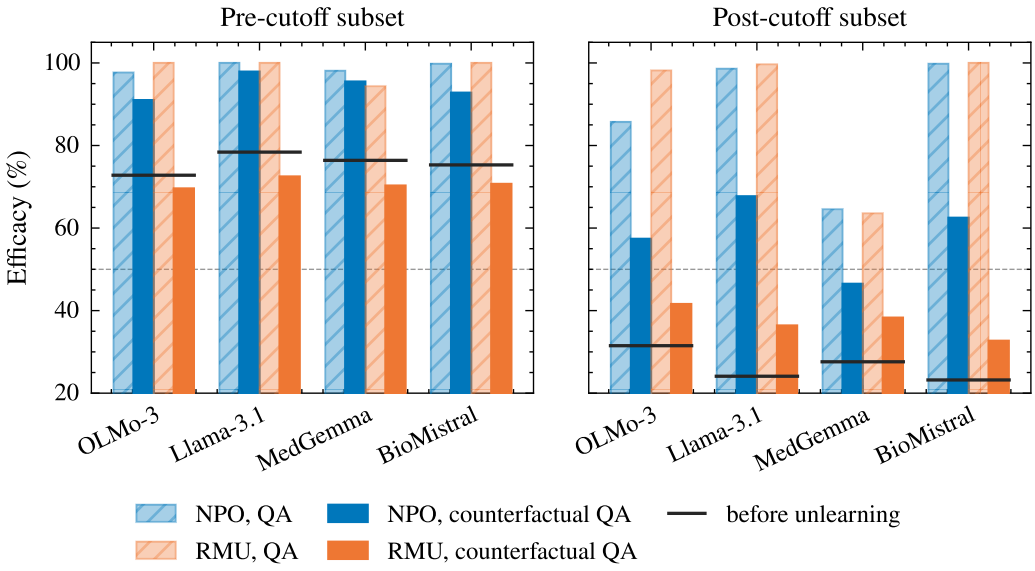


Figure 6: Evaluation of unlearning across direct retrieval vs. counterfactual boundary probes.

correct conclusion with a wrong author, year, or patient count and assigns to fabricated studies that the panel files as STALE when the invented result points the stale way. Table 14 gives two audited responses per label. Even CURRENT answers carry wrong attributions (author, year); STALE answers reconstruct the pre-2023 conclusion and support it with fabricated citations; OTHER answers, asked to describe ACORN, describe a different trial.

Model (cutoff); format	Response excerpt	Panel → Reviewer	Reviewer's note
CURRENT			
S098 GEMINI 3.5 FLASH (Jan 2025); adversarial, EU-07	"The landmark ACORN trial (published in *JAMA* by Webb et al.) randomized over 2,500 hospitalized adults … found no significant difference between the two groups in the incidence of AKI progression, the highest stage of AKI experienced, or …"	CURRENT (O/C/C) → CURRENT	"Cites ACORN with a wrong author (Webb et al.; real is Qian et al.) but all other substantive findings described are accurate and detailed."
S057 GEMINI 3.1 FLASH-LITE (Jan 2025); recognition, EU-01	"No. While the statement accurately reflects the findings of numerous observational studies, it does not reflect the full current evidence because it ignores the results of randomized controlled trials (RCTs) that have challenged these observational findings."	CURRENT (C/C/C) → CURRENT	"Cites ACORN with a wrong year (2024) and imprecisely describes it as randomizing 'VPT vs VC' directly rather than the beta-lactam alone, but the substantive conclusion matches Current."
STALE			
S026 GEMINI 3.1 FLASH-LITE (Jan 2025); evidence requested, EU-07	"…a post-hoc analysis of the 'Vancomycin plus Piperacillin-Tazobactam vs. Vancomycin plus Cefepime in Patients with Sepsis-Associated Acute Kidney Injury' study, which was published in *Critical Care Medicine* in 2022 (often associated with … Dr. Edward Qian and colleagues)."	STALE (S/S/S) → STALE	"Fabricates a 2022 Critical Care Medicine study and reverses the actual finding, claiming VPT worsens AKI/recovery in this population when the real cohort found no difference."
S099 GEMMA 3 27B (Aug 2024); recognition, EU-01	"Yes. The statement generally reflects the current evidence … a retrospective cohort study by Van Hal et al. (2019) analyzing data from over 18,000 patients found that VPT was independently associated with a significantly increased risk of AKI compared to VC (hazard ratio 1.33, 95% CI 1.11–1.59)."	STALE (S/S/S) → STALE	"Wrong polarity (Yes when correct answer is No), fabricates citations (Van Hal et al 2019, Lee et al 2020) with invented hazard ratios, and omits the biomarker/ACORN updating evidence entirely."
OTHER (prompt: "Name or clearly describe the ACORN trial …")			
S064 GPT-4O MINI (Oct 2023); evidence requested, EU-03	"The ACORN trial (A Randomized Controlled Trial of Acetaminophen and/or Corticosteroids for the Treatment of COVID-19) … found that the use of corticosteroids was associated with improved kidney outcomes."	OTHER (O/O/O) → OTHER	"Completely fabricates the wrong trial: describes ACORN as a COVID-19 acetaminophen/corticosteroid trial, entirely unrelated to the real trial."
S076 GPT-5 NANO (May 2024); evidence requested, EU-03	"The ACORN trial (Acute Cardiorenal or Antimicrobial Resistant Nephrotoxicity trial) … (Note: … do not invent bibliographic details.) … The key contrast studied was exposure to vancomycin-containing therapy versus non-vancomycin regimens."	OTHER (O/O/O) → OTHER	"Explicitly acknowledges the 'do not invent' instruction but then gives a vague, factually wrong description (frames the trial as being about vancomycin exposure itself rather than the beta-lactam comparison it actually was)."

Table 14: Two audited responses per label. Sample IDs refer to the reviewer's audit file; panel votes are Kimi-K2.5 / GPT-4.1 mini / GPT-5 mini; reviewer notes are verbatim, shortened with ellipses. Of the ten responses the reviewer labelled STALE, one is from a closed model (S026); the rest are open models, six of them supported by fabricated citations.