Facts Fade Fast: Evaluating Memorization of Outdated Medical Knowledge in Large Language Models

Juraj Vladika, Mahdi Dhaini, Florian Matthes
Technical University of Munich, Germany
School of Computation, Information and Technology
Department of Computer Science
`{juraj.vladika, mahdi.dhaini, matthes}@tum.de`

Abstract

The growing capabilities of Large Language Models (LLMs) show significant potential to enhance healthcare by assisting medical researchers and physicians. However, their reliance on static training data is a major risk when medical recommendations evolve with new research and developments. When LLMs memorize outdated medical knowledge, they can provide harmful advice or fail at clinical reasoning tasks. To investigate this problem, we introduce two novel question-answering (QA) datasets derived from systematic reviews: *MedRevQA* (16,501 QA pairs covering general biomedical knowledge) and *MedChangeQA* (a subset of 512 QA pairs where medical consensus has changed over time). Our evaluation of eight prominent LLMs on the datasets reveals consistent reliance on outdated knowledge across all models. We additionally analyze the influence of obsolete pre-training data and training strategies to explain this phenomenon and propose future directions for mitigation, laying the groundwork for developing more current and reliable medical AI systems.

1 Introduction

The advent of pre-trained Large Language Models (LLMs) has revolutionized the field of Natural Language Processing (NLP) (Naveed et al., 2025). One of their most promising application domains is healthcare, where they hold the potential to democratize access to health services and assist in crucial clinical workflows (Thirunavukarasu et al., 2023; Ayers et al., 2023; Liu et al., 2025).

LLMs are trained to predict the next token on massive amounts of text data, which results in deeply encoding a lot of knowledge in their weights (Dhingra et al., 2022; Chang et al., 2024). Recent studies suggest that LLMs encode clinical knowledge effectively (Singhal et al., 2023; Zhang et al., 2025), by being trained on medical texts like patient records and clinical guidelines. The model's ability to recall specific facts from this data is often referred to as *memorization* (Carlini et al., 2022).

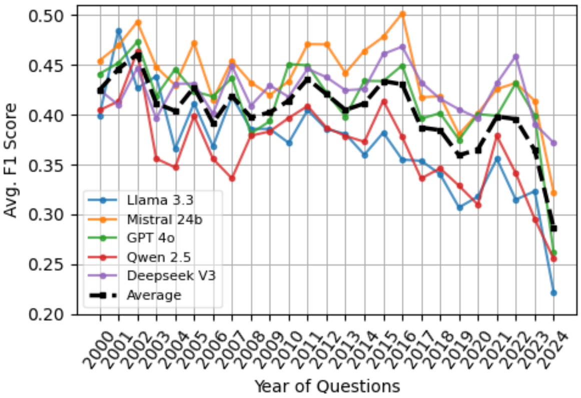

Figure 1: Average F1 scores of five LLMs on medical questions originating from different years in our dataset. The performance decline as questions get more recent points to stronger memorization of older knowledge.

World knowledge quickly evolves in dynamic domains like entertainment or politics. However, this also happens with scientific knowledge. In medicine, new high-quality evidence constantly emerges, rendering previous recommendations obsolete (Hodder et al., 2024). Consequently, the knowledge memorized by an LLM at its training time can become outdated, as they struggle to keep up with the evolving world knowledge (Zhang et al., 2023). This is a major safety concern, as it can lead to LLMs providing consumers with incorrect health advice (Li et al., 2023; Ong et al., 2024) or fail in clinical settings when using flawed facts in their reasoning (Hager et al., 2024). Even when augmented with retrieved up-to-date information, LLMs can reject it and resort to internal knowledge in so-called *knowledge conflicts* (Xu et al., 2024).

While recent work has explored the memorization of outdated knowledge in the encyclopedic domain (Vu et al., 2024; ChenghaoZhu et al., 2025), the temporal decay of *medical* knowledge has been less explored. To address this critical gap, we con-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9161–9174
November 4-9, 2025 ©2025 Association for Computational Linguistics

struct new datasets and use them to evaluate memorization of outdated medical knowledge:

- We introduce **MedRevQA**, a new dataset of 16,501 QA pairs from medical systematic reviews; and **MedChangeQA**, a subset of 512 pairs where answers have changed over time.
- We benchmark eight LLMs on our datasets, demonstrating that all models exhibit memorization of outdated medical information.
- We provide in-depth analysis, including tracing outdated knowledge to the training data, and discuss promising mitigation strategies.

The datasets and code are publicly available.[1]

2 Related Work

NLP has a wide array of applications to the biomedical field (Wang et al., 2023) and LLMs have shown great potential in various medical tasks and clinical applications (Thirunavukarasu et al., 2023). A popular task within NLP for healthcare is biomedical question answering (BQA) (Jin et al., 2022; Nentidis et al., 2024). BQA is seen as a good proxy for evaluating how well LLMs encode and recall medical knowledge (Subramanian et al., 2024; Singhal et al., 2023, 2025) – therefore, we use it as our main task. The most similar QA dataset in construction is MedREQAL (Vladika et al., 2024b), but we majorly expand the scope and the purpose.

Recent work has explored how to measure memorized training examples in LLMs (Jagielski et al., 2023; Maini et al., 2024; Kassem et al., 2025). Similarly, temporal QA datasets have been constructed to investigate quickly changing knowledge, mostly focusing on the general, encyclopedic domain (Kasai et al., 2023; Vu et al., 2024; Li et al., 2024).

To the best of our knowledge, we introduce the first QA dataset focusing on knowledge change specifically for the *medical* domain and the first investigation of how much outdated medical knowledge popular LLMs encode.

3 Dataset

Systematic Literature Reviews (SLRs). Our dateset is constructed from medical SLRs, studies which aim to bring evidence together to answer a pre-defined research question. This involves the identification of primary research relevant to the question, the critical appraisal of this research, and the synthesis of the findings (Kolaski et al., 2023). SLRs are considered the highest quality evidence in the medical "hierarchy of evidence" (Wallace et al., 2022). We use SLRs to construct a QA dataset because their clear structure and strict criteria used for decisions make them a well-suited proxy for evaluating the state of encoded medical knowledge. We use the SLRs from Cochrane Collaboration (Cumpston et al., 2022), which is the most well-known international organization specializing in the construction of SLRs. Many Cochrane SLRs are updated as new evidence for a question arises.

Question: Does long-term antibiotic use help prevent recurrent urinary tract infections in children?

Conclusion: Long-term antibiotics may reduce the risk of repeat symptomatic UTI in children who have had one or more previous UTIs but the benefit may be small and must be considered together with the increased risk of microbial resistance. (...) [Williams, 2019]

Verdict: **Supported**

Question: Does long-term antibiotic use help prevent recurrent urinary tract infections in children?

Conclusion: Large, properly randomised, double blinded studies are needed to determine the efficacy of long-term antibiotics for the prevention of UTI in susceptible children. (...) [Williams, 2011]

Verdict: **Not Enough Information**

Table 1: Example of two instances from our dataset, showing how the verdict changed through time as new, higher quality evidence was discovered.

Dataset Construction. PubMed, the largest database of medical publications (White, 2020), contains all the Cochrane systematic review abstracts from 2000 to 2024 (until January 2024, when we scraped). We built a Python scraping script using BeautifulSoup and scraped all the Cochrane SLR abstracts.[2] Every SLR in the dataset consists of the same sections: *Background, Objectives, Search methods, Selection criteria, Data collection and analysis, Main results,* and *Authors' conclusions*. Our final QA dataset consists of **questions** and **labels**. We used *gpt-4o-mini-2024-07-18* to semi-automatically construct the questions and labels, by providing it with the full SLR abstract.

Questions were derived from the *Objectives* section, by rewriting them to the interrogative form. Labels originate from the *Authors' conclusions* sec-

---

[1] `https://github.com/jvladika/MedChange`

[2] `https://pubmed.ncbi.nlm.nih.gov/?term=%22Cochrane+Database+syst+rev%22%5BJournal%5D`

tion – the LLM selects one of the labels Supported, Refuted, or Not Enough Information (NEI), as the final **label**. These labels align with common labels in other medical QA and fact-checking datasets (Glockner et al., 2024b; Vladika et al., 2024a). In total, this dataset has 16,501 QA pairs, spanning virtually all medical disciplines and covering a wide array of important biomedical questions for benchmarking. We call this dataset **MedRevQA**.

Changed Knowledge. Our dataset consists of 16,501 SLR records. Out of those, 12,122 are unique SLRs that have never had an update. The remaining 4379 SLRs constitute 1535 groups (with a minimum of 2 SLRs in a group, a maximum of 9, and a mean of 2.85) that researched the same question. This means there are 1535 research questions that have had multiple SLR iterations written about them. Out of 1535, **512** have had a verdict change over time, meaning that the authors changed the conclusion of the investigated research question in a follow-up SLR study, when they acquired updated evidence from research. This follows findings from medical research studies that have shown how 20 to 30% of Cochrane reviews change their conclusions throughout time (Hughes et al., 2012; Babić et al., 2022). We consolidate these questions with changed verdicts into the **MedChangeQA** dataset and collect all their verdicts through different iterations. MedChangeQA has **questions**, **latest label**, and (the most recent) **outdated label** for those studies where the label changed.

Annotation Quality. Two annotators, one who is our in-house physician from the university clinic and another an author with a background in biomedical engineering, evaluated a random subset of 100 examples for the correctness of generated questions and verdicts. They found 95% of questions and 92% of labels to be correct. We deem this to be relatively high, since even the human annotation process is imperfect (Klie et al., 2023), with errors due to incorrect problem understanding or loss of concentration. A common source of label errors was conflating Refuted and NEI labels. On the other hand, all 512 labels in MedChangeQA were manually checked and corrected by the two annotators. Therefore, MedRevQA has silver labels, while MedChangeQA has gold labels.

Dataset Description. In total, MedRevQA has 16,501 questions, of which 6499 are *Supported*, 3124 are *Refuted*, and for 6878 there is *Not enough information*. In MedChangeQA, for the 512 questions with changed verdicts, the newest labels have a 221/131/160 ratio for S/R/NEI, and the outdated labels are at 152/123/237 for S/R/NEI, showing how the most common change is from not having enough information to becoming supported or refuted by relevant research. Still, some questions can go from support/refute to inconclusive findings with more research (see an example in Table 8).

4 Experimental Setup

The experiments consisted of instructing the LLMs to predict one of the three labels (S/R/NEI), given the medical question as input. No additional context was provided, as the goal is to evaluate their internal knowledge and memorization. The models also explained their output. For evaluation, we extract the predicted label and compare it to ground-truth labels from the dataset, using the (macro-averaged) precision, recall, and F1 scores.

We test multiple LLMs, starting with **GPT-4o** (*2024-08-06*), as the most popular proprietary LLM. We also benchmark four open-weights LLMs: **Mistral 24B**, **Llama 3.3 (70B)**, **Qwen 2.5 (7B)**, **DeepSeek-V3 (685B)**; and finally the fully open-source **OLMo 2 (13B)**. See Appendix B for a summary of public info on their pre-training data.

To compare general-purpose LLMs to domain-specific ones, we also benchmark the performance of **PMC-LLaMa 13B** (Wu et al., 2023), an extension of Llama 2, and **BioMistral 7B** (Labrak et al., 2024), an extension of Mistral-v0.2; both further pre-trained on biomedical research papers.

All prompts can be found in Table 6. GPT 4o was prompted through the OpenAI API. The four general-purpose models were prompted via the API of Together AI. Two biomedical LLMs were run locally (in an 8-bit quantized version) on one Nvidia V100 GPU with 16 GB VRAM, for two computation hours. The token limit was set to 512 and the temperature to 0 to maximize deterministic outputs.

5 Results

Experiment Rounds. We first test (a) the full dataset, MedRevQA. We also did two experiments on MedChangeQA, first with (b) outdated labels as ground truth, then with (c) latest labels as ground truth. We use the difference between the scores of (b) and (c) as a proxy to show the extent of outdated medical knowledge in LLMs. Final results are systematized in Table 2, measured by macro P, R,

		(a) **Full Dataset** (16.5k)			**Changed Knowledge Dataset** (512)							
	Release Date				(b) **Outdated Lab.** ↓			(c) **Latest Labels** ↑			**F1 diff.**	**Outdated Answers**
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**		
GPT-4o	2024-05-13	52.6	45.1	42.9	45.5	38.9	34.1	35.2	34.5	31.1	-3.0	39.4%
Mistral 24B	2025-01-30	50.6	**46.3**	**45.7**	38.2	37.6	33.9	36.9	35.5	33.7	-0.2	38.7%
Llama 3.3 70B	2024-12-06	52.7	45.9	39.3	38.9	36.6	**26.7**	**42.8**	**39.3**	34.1	**+7.4**	32.2%
Qwen 2.5 7B	2024-09-19	46.4	42.3	38.7	42.6	37.1	30.8	27.1	30.8	26.0	**-4.8**	35.4%
Deepseek V3	2024-12-26	**56.2**	46.2	43.8	43.2	38.6	33.9	40.2	35.1	32.2	-1.7	40.6%
OLMo 2 13B	2024-11-24	43.5	42.5	37.9	**36.2**	**35.3**	29.3	35.5	35.7	33.2	+2.9	32.0%
PMC-Lm 13B	2023-08-28	39.5	37.6	36.5	41.9	39.8	35.9	34.5	34.3	33.1	-2.8	37.3%
BioMistral 7B	2024-02-19	41.2	41.5	40.9	36.8	37.2	36.3	35.4	35.5	**35.3**	-1.5	37.1%

Table 2: Final results of eight LLMs, measured by macro Precision (P), Recall (R), and F1. Experiments include (a) the full dataset; and the changed knowledge dataset, using (b) outdated labels and (c) latest labels as ground truth, respectively. The final column is the percentage of answers in (b) where an outdated label was predicted.

and F1. The last column shows the percentage of answers in the 3rd experiment (c) where the outdated label was predicted (and not the correct latest label or an incorrect label altogether).

Performance. On the full dataset, Mistral exhibited the best R and F1, showing it has the best overview of the overall medical knowledge landscape. Precision was the highest in Deepseek-V3. Nevertheless, none of the models has a very high performance, pointing to the challenging nature of MedRevQA as a general biomedical QA testbed.

When it comes to outdated knowledge, Llama 3.3 had the highest degree of the latest knowledge as compared to the outdated labels (+7.4), while OLMo also had a positive difference (+2.9). Mistral showed an almost identical performance, while GPT, Qwen, DeepSeek, and PMC-Llama all struggled. Qwen was also the smallest and least capable model, which could explain low scores in general and low awareness of recent knowledge.

An example of outdated and incorrect knowledge is shown in Table 7 in Appendix. Additionally, Figure 1/4 shows how the average F1 across LLMs on questions from different years on MedRevQA declines in more recent years, as all post-2016 average scores are lower than any beforehand. A similar drop in LLM performance on more recent medical questions was found by Park et al. (2025).

6 Discussion and Analysis

Pre-training Data. Most popular LLMs do not fully disclose their pre-training data, making it difficult to assert if concrete medical studies were memorized. Still, recent studies demonstrated empirically the presence of memorized medical datasets (Gallifant et al., 2024; Yang et al., 2024). We also saw a tendency of models to explicitly mention specific studies in their answers, including Cochrane reviews, many of which were decade-old (see Table 8), thus displaying outdated memorized knowledge (see Table 4). We outline pre-training corpora of used LLMs in Appendix B, and for the fully open OLMo, we show the presence of all used SLRs in its pre-training corpus, with earlier ones being more prevalent (Figure 2).

Inspection of OLMo. OLMo 2 (OLMo et al., 2025) is trained on the Dolma corpus (Soldaini et al., 2024), a fully open dataset containing around 3 trillion tokens. It contains the peS2o (Soldaini and Lo, 2023) and S2ORC (Lo et al., 2020) corpora that constitute the academic knowledge base Semantic Scholar. This database also indexes all of Cochrane's systematic literature reviews.[3] Therefore, we can with high certainty say that the OLMo models have seen Cochrane's SLRs during its pre-training. Other than in the two academic corpora, there is a wide presence of these SLRs in other parts of the dataset, especially various online websites found in Common Crawl (Dodge et al., 2021).

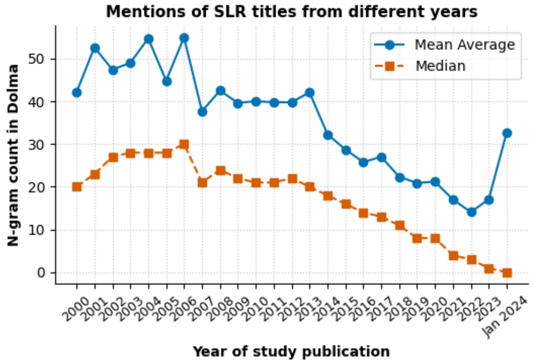

Figure 2: N-gram counts per the year of studies in Dolma, the pre-training corpus of OLMo.

We used Infini-gram (Liu et al., 2024), an n-

---

[3] `https://www.semanticscholar.org/venue?name=Cochrane%20Database%20of%20Systematic%20Reviews`

gram language model that can be used to query Dolma and other pre-training corpora,[4] to inspect the presence of Cochrane's SLRs. Searching for "*Cochrane Database of Systematic Reviews*" (the exact journal name, case-sensitive) returns 144,493 hits for Dolma v1.7 (used for OLMo 2). Additionally, we queried the title of each of the SLR studies found in MedRevQA and report on the mean and median amount of n-gram counts per year in Figure 2. The mean and median almost steadily decrease over years, meaning that the most mentioned and discussed studies are the earliest ones since they have had more time to spread throughout the web. The higher frequency of mentions can lead to to stronger encoding of outdated knowledge in LLMs.

Mentions of Specific Studies In Table 4, we show the number of mentions of some common terms referring to specific medical studies (such as *systematic review*, *meta-analysis*, *journal*), across all LLM answers on MedRevQA questions. This shows how models tend to cite specific studies when providing some of their answers, which is useful for source attribution, but becomes problematic when the referred studies are outdated and deprecated. It is notable how GPT overwhelmingly resorted to using general phrasing such as "*studies have shown a positive effect...*" without specifying what studies exactly it is referring to. This likely comes from its final alignment and preference-learning phase, where a particular answering style is learned.

Potential Explanations. We hypothesize some reasons for the presence of strongly encoded outdated knowledge. Firstly, older scientific findings have been around for a longer time and have already permeated the Internet, news, follow-up studies, and are present in pre-training corpora. Additionally, scientific findings are often misrepresented online (Glockner et al., 2024a; Wührl et al., 2024), so faulty medical knowledge could get encoded.

Secondly, LLM memorization rate has been correlated in past work with various training parameters, such as learning rate (Tirumala et al., 2022), model size (Biderman et al., 2023), or frequency of appearance in training data (Carlini et al., 2023). Therefore, it is possible that Llama had the highest data quality and more weight during training put on more recent text, leading to less outdated knowledge. Finally, the cutoff of all models is 2023, and the vast majority of "latest labels" are from before 2023 (see Figure 3). Cutoff could explain the drop in 2023/2024 (Figure 1) but not earlier years.

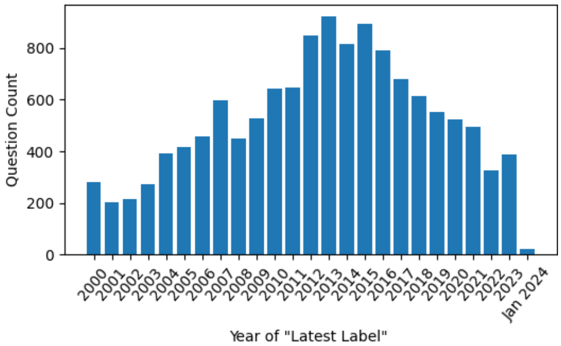

Figure 3: Distribution of the year of "Latest Label" – the label of the most recent study for a given question

Future Directions. One way to overcome outdated knowledge is using retrieval-augmented generation (RAG). We show in Appendix A how a simple retrieval strategy (with the first PubMed result added to prompt) can already bring improvements. Still, LLMs can hallucinate extra information even in RAG settings (Adlakha et al., 2024) or not follow the provided references (Liu et al., 2023). Therefore, advanced RAG techniques are needed (Yu et al., 2024), including filtering and re-ranking of retrieved evidence by recency and source quality. Future work could also investigate more nuanced labels or long-form answer generation with explanations. Other promising directions for mitigation of memorized knowledge include: resolving knowledge conflicts (Wang et al., 2024b), machine *unlearning* (Yao et al., 2024; Gao et al., 2025), knowledge editing (Wang et al., 2024a; Jiang et al., 2024), and continual learning (Shi et al., 2024).

7 Conclusion

We introduce two new QA datasets constructed from rigorous biomedical SLRs for benchmarking the general biomedical knowledge of LLMs. With a subset of 512 questions where the answer (verdict) changed over the years, we showcase how eight popular LLMs fare better on older medical knowledge and encode a considerable amount of outdated knowledge labels, which can hinder their usability in healthcare settings, including helping physicians, researchers, and patients. We outline future work directions and hope our datasets will serve as a challenging testbed for tackling the LLM memorization of outdated knowledge.

---

[4] `https://huggingface.co/spaces/liujch1998/infini-gram`

Limitations

Most of the MedRevQA dataset introduced in this study was constructed semi-automatically, by scraping the content and using an LLM to generate the question and label. It is possible that some of the generated questions and labels are imperfect. Our manual analysis of 100 randomly selected instances showed that the performance is 92–95% correct, for labels and questions, respectively. We considered this to be a good enough performance, considering that even human annotation is not always perfect. MedRevQA should be interpreted as having silver labels and used as such.

We use the difference in F1 scores between the predicted labels when using "outdated labels" and "latest labels" as ground truth, as a proxy for evaluating the degree of encoded outdated medical knowledge. This is not a perfect measure since it is possible that the LLM predicted an incorrect label due to some logical error or misinterpreting the question. Still, our manual inspection of a large number of generated labels and explanations showed that outdated references were indeed the most common explanation for the label misprediction, and models often referred to old SLRs and meta-analyses, dating back many years.

Given the quickly evolving landscape of new LLMs, some of the LLMs we evaluate can get outdated and deprecated quickly. Additionally, we do not test all the relevant models since some were not available or computationally too expensive for us to run. Due to a lack of resources, our study also lacks a deeper human evaluation of the generated model labels with medical experts, which could have given a more rigorous evaluation.

Ethics Statement

The work presented in this study focuses on the sensitive fields of healthcare and medical AI. We predict answers to questions in a zero-shot setting to uncover their internal encoded medical knowledge for research purposes, but this is not suitable for end users or patients. Some responses can include inaccuracies and misleading medical advice, which should be critically evaluated and verified with reliable sources or medical professionals.

The original text of Cochrane's systematic literature review abstracts belongs to the Cochrane Collaboration. We will release only the generated questions and label pairs under an open license, and respect the copyright of Cochrane's work.

Acknowledgements

This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS17049 Software Campus 2.0 (TU München). We would like to thank the anonymous reviewers for their insightful comments and suggestions to our paper.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596.

Andrija Babić, Tina Poklepović Peričić, Dawid Pieper, and Livia Puljak. 2022. When is the evidence conclusive? analysis of systematic reviews for which cochrane declared that conclusions will not change with further studies. *Research Synthesis Methods*, 13(4):478–488.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Gregory Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668.

ChenghaoZhu ChenghaoZhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang.

2025. Is your LLM outdated? a deep look at temporal generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7433–7457, Albuquerque, New Mexico. Association for Computational Linguistics.

Miranda S Cumpston, Joanne E McKenzie, Vivian A Welch, and Sue E Brennan. 2022. Strengthening systematic reviews in public health: guidance in the cochrane handbook for systematic reviews of interventions. *Journal of Public Health*, 44(4):e588–e592.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack Gallifant, Shan Chen, Pedro José Ferreira Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. 2024. Language models are surprisingly fragile to drug names in biomedical benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12448–12465, Miami, Florida, USA. Association for Computational Linguistics.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. 2025. On large language model continual unlearning. In *The Thirteenth International Conference on Learning Representations*.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024a. Missci: Reconstructing fallacies in misrepresented science. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4372–4405.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024b. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Rebecca K Hodder, Joshua P Vogel, Luke Wolfenden, and Tari Turner. 2024. Living systematic reviews and living guidelines to maintain the currency of public health guidelines. *American journal of public health*, 114(1):21–26.

EG Hughes, M van Wely, and CM Farquhar. 2012. Cochrane reviews in perspective: the importance of appropriate conclusions and timing of publication.

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. 2023. Measuring forgetting of memorized training examples. In *The Eleventh International Conference on Learning Representations*.

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, et al. 2024. Learning to edit: Aligning llms with knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4705.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Aly M. Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. 2025. ALPACA AGAINST VICUNA: Using LLMs to uncover memorization of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8296–8321, Albuquerque, New Mexico. Association for Computational Linguistics.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Kat Kolaski, Lynne Romeiser Logan, and John PA Ioannidis. 2023. Guidance to best tools and practices for systematic reviews. *Journal of Pediatric Rehabilitation Medicine*, 16(2):241–273.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Latesteval: addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, et al. 2025. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. LLM dataset inference: Did you train on my dataset? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima-López, Eulàlia Farré-Maduell, Martin Krallinger, Natalia Loukachevitch, Vera Davydova, Elena Tutubalina, and Georgios Paliouras. 2024. Overview of bioasq 2024: the twelfth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–27. Springer.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J. Butte, Nigam H. Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, and Daniel Shu Wei Ting. 2024. Medical ethics of large language models in medicine. *NEJM AI*, 1(7):AIra2400038.

Yein Park, Chanwoong Yoon, Jungwoo Park, Donghyeon Lee, Minbyul Jeong, and Jaewoo Kang. 2025. Chroknowledge: Unveiling chronological knowledge of language models in multiple domains. In *The Thirteenth International Conference on Learning Representations*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Luca Soldaini and Kyle Lo. 2023. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI. ODC-By, `https://github.com/allenai/pes2o`.

Anand Subramanian, Viktor Schlegel, Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Vijay Prakash Dwivedi, and Stefan Winkler. 2024. M-QALM: A benchmark to assess clinical reading comprehension and knowledge recall in large language models via question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4002–4042, Bangkok, Thailand. Association for Computational Linguistics.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024a. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024b. MedREQAL: Examining medical knowledge recall of large language models via question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14459–14469, Bangkok, Thailand. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Sowdhamini S Wallace, Gal Barak, Grace Truong, and Michelle W Parker. 2022. Hierarchy of evidence within the medical literature. *Hospital Pediatrics*, 12(8):745–750.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pretrained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

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

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024b. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Amelie Wührl, Dustin Wright, Roman Klinger, and Isabelle Augenstein. 2024. Understanding fine-grained distortions in reports of scientific findings. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6175–6191.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Xinyu Yang, Zichen Wen, Wenjie Qu, Zhaorun Chen, Zhiying Xiang, Beidi Chen, and Huaxiu Yao. 2024. Memorization and privacy risks in domain-specific large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Yuanshun Yao, Xiaojun Xu, and YangLiu. 2024. Large language model unlearning. In *Advances in Neural Information Processing Systems*, volume 37, pages 105425–105475. Curran Associates, Inc.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

Ziheng Zhang, Zhenxi Lin, Yefeng Zheng, and Xian Wu. 2025. How much medical knowledge do llms have? an evaluation of medical knowledge coverage for llms. In *Proceedings of the ACM on Web Conference 2025*, pages 5330–5341.

A Simple RAG Improvement

Common ways to address the outdated knowledge with *knowledge editing* include continual learning methods and external search augmentation. We do a simple experiment using a retrieval-augmented method. For each of the 512 questions in MedChangeQA, we query the PubMed API and take the abstract of the Top 1 result, and append it to the main prompt as an additional context. Results are shown in Table 3. This improves the F1 scores by a margin of 3–16 and partially closes the gap, but still leaves a lot of room for improvement. This serves as a simple demonstration of one way to address the outdated knowledge – future work could focus on retrieving more documents, using structured and focused search queries (like searching for SLRs), semantic search, graph RAG, learning to re-rank and avoid conflicts, etc. Additionally, methods of continual learning and fine-tuning could be used, with MedChangeQA serving as a testbed to measure the rate of success of the proposed techniques.

	P	**R**	**F1**	**Improv. F1**
GPT 4o	43.4	40.2	39.8	+8.7
Mistral	47.5	41.5	39.6	+5.9
Llama 3.3	44.1	42.3	38.8	+4.7
Qwen	43.3	43.7	42.2	+16.2
Deepseek	40.7	39.3	35.4	+3.2

Table 3: Performance improvements with the abstract of the top PubMed paper included in prompt.

B Pre-Training Data of LLMs

This section provides an overview of what is publicly known about pre-training data for the used LLMs, as reported in their technical reports or official documentation:

- **Llama 3.3**: Pretrained on approximately 15 trillion tokens of data sourced from publicly available online sources. The exact composition and breakdown of the dataset are not detailed, but Meta emphasizes that the data is "a new mix" of public internet data. The data cutoff for pretraining is December 2023.
- **Mistral 24B** The official technical report and available documentation do not provide explicit details about the pre-training corpus for Mistral 24B. However, Mistral's models are generally known to be trained on large-scale, diverse datasets, often including filtered web data, code, and other standard sources, but no specifics are publicly disclosed for the 24B version in the sources provided.
- **GPT 4o** It was trained on data up to October 2023, sourced from a "wide variety of materials," including: (a) publicly available data (web pages, ML datasets, common crawls), (b) proprietary data (obtained via data partnerships, e.g., paywalled content, archives, metadata), (c) key dataset components (web data, code and math data, multimodal data). The dataset underwent safety filtering to remove harmful content, personal information, and explicit material. OpenAI does not provide a detailed breakdown of dataset proportions or specific sources.
- **Qwen 2.5** It was trained on up to 18 trillion tokens of data. The dataset is described as "large-scale" and "high-quality," but the technical report does not specify the exact sources. The data is designed to provide a strong foundation for common sense, expert knowledge, and reasoning. Qwen 2.5 also supports multilingual capabilities across more than 29 languages.
- **DeepSeek V3** It was trained on 14.8 trillion tokens of "diverse and high-quality" data. The dataset construction focused on: an increased ratio of mathematical and programming samples, multilingual coverage, and a data processing pipeline optimized for diversity and minimal redundancy. The technical report does not provide a granular breakdown of data sources but highlights the focus on math, code, and multilingual content.
- **OLMo 2** is trained on the Dolma corpus (Soldaini et al., 2024), a fully open dataset containing around 3 trillion tokens. This is a high-level breakdown of the composition of the pre-training corpus: Common Crawl (2,479 billion tokens), GitHub (411 billion tokens), Reddit (89 billion tokens), Semantic Scholar (70 billion tokens), Project Gutenberg (6.0 billion tokens), Wikipedia and Wikibooks (4.3 billion tokens).

- **PMC-Llama** and **BioMistral** use the base models of Llama and Mistral as described before, but were then further pre-trained on abstracts of biomedical publications from PubMed and full publications from PubMed Central. As described in our paper, PubMed contains all the abstracts of Cochrane systematic reviews, which means all the Cochrane SLRs from our dataset were a part of training.

C Mentions of Common Terms

Table 4 show the count of how many times were some common terms mentioned in answers of the eight evaluated LLMs to 16,501 questions from the MedRevQA dataset. These terms include those that signal a mention of specific studies used as a reference for the answer, such as *systematic review*, *meta-analysis*, *journal*, or *Cochrane* since that is the publishing organization, from which our questions and labels originate. We also included the generic term *studies*, which is often mentioned in those answers that do not refer to specific studies but only give a general statement such as "*Many studies have shown that...*". The use of this generic term was especially common in GPT-4o, which mentioned specific studies the least. The two biomedical LLMs, BioMistral and PMC-Llama, which were further pre-trained on full texts of biomedical publications from PubMed Central, also tended to cite specific studies the most.

D Prompts and Examples

This Appendix section provides additional material for the study, including the model prompts in full length (Tables 5 and 6) and example questions and model answers (Tables 7 and 8). Figure 4 shows a larger version of the plot of the average F1 score for tested LLMs on questions over the years.

	Llama 3.3	**Mistral**	**GPT-4o**	**Qwen 2.5**	**Deepseek**	**OLMo 2**	**BioMistral**	**PMC-L**
"Cochrane"	51	783	2	629	901	283	2067	2344
"systematic review"	221	1664	623	3194	3046	531	3990	4956
"meta-analysis"	844	3511	714	4180	2776	981	4253	4618
"journal"	53	689	7	4620	448	196	574	624
"studies"	7024	12419	13493	12516	13421	6615	7598	9720

Table 4: Number of answers (out of 16,501) from each tested LLM where the respective terms were mentioned. This shows the tendency to refer to and cite relevant medical studies that were memorized during pre-training. Two biomedical models, which were pre-trained on biomedical publications, also refer to specific studies the most.

Use Case	**Prompt Content**
Question & Label generation	`SYSTEM`: You're a helpful assistant. Your task is to help with generating questions and labels in the medical and clinical domain. `USER` You will be given an excerpt of an abstract of a clinical systematic review. Based on the given background, objectives, and author's conclusions, generate only ONE SINGLE question, answerable with yes/no/uncertain, that sums up the main medical objective that was investigated. Please keep the question short and general and use the "Objectives" section to construct the question. The question should be about a general medical hypothesis, not about this specific review. Afterwards, please also give a label for the author's conclusions. The label tries to answer the objective by looking at the conclusion. The label may be ONLY from one of the following three: (1) SUPPORTED; (2) REFUTED; (3) NOT ENOUGH INFORMATION. Do not try to make up a new label. Please only select the third label if not enough evidence was found to reach a verdict, not if the certainty of the conclusion is low! Please aim to predict "SUPPORTED" or "REFUTED" even if the certainty of these conclusions by the authors is low or moderate. Please structure the output in two lines, as: QUESTION: (question) LABEL: (label) The documents begins now.

Table 5: Overview of applied prompts for data generation and annotation.

Model	**Prompt**
PMC-LLaMa	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Based on your best current knowledge, please answer the following medical question. If you think there is not enough evidence to answer, then say so. Please answer the question with "SUPPORTED" or "REFUTED" or "NOT ENOUGH INFORMATION". Briefly explain your answer. ### Input: {question} ### Response:
BioMistral	<s>[INST] Based on your knowledge, please answer this clinical question only with SUPPORTED (if the question is supported by the clinical research) or REFUTED (if the hypothesis is refuted by the current clinical research) or NOT ENOUGH INFORMATION (if there is insufficient evidence for the question in current research). Please give your output in form of LABEL: (label) . Briefly explain your answer. QUESTION: {question} [/INST]
Mistral 24B, Llama 3.3, GPT-4o, Qwen 2.5, DeepSeek-V3	`SYSTEM` You are an AI assistant helping answer clinical and medical questions based on your best knowledge. `AGENT` Please answer this clinical question only with SUPPORTED (if the question is supported by the clinical research) or REFUTED (if the hypothesis is refuted by the current clinical research) or NOT ENOUGH INFORMATION (if there is insufficient evidence for the question in current research). Please give your output in form of LABEL: (label) . Briefly explain your answer. QUESTION: {question}

Table 6: Prompts used for each of the LLMs in experiments.

Question	Is routine malaria chemoprophylaxis beneficial for people with sickle cell disease in malaria-endemic areas?
Conclusion (Latest)	(...) It is beneficial to give routine malaria chemoprophylaxis in sickle cell disease in areas where malaria is endemic. (Oniyangi, 2019) [**label**: **Supported**]
Conclusion (Outdated)	(...) There is very little direct evidence to support or refute giving routine chemoprophylaxis in sickle cell disease in areas where malaria is endemic. (Oniyangi, 2006) [**label**: **Not Enough Information**]
GPT-4o	Routine malaria chemoprophylaxis is beneficial for people with sickle cell disease in malaria-endemic areas. Clinical research indicates that individuals with sickle cell disease are at increased risk for severe malaria and its complications. Chemoprophylaxis can reduce the incidence of malaria and its associated morbidity in this vulnerable population, thereby improving health outcomes. [**predicted**: **Supported**]
Llama 3.3	Current clinical research suggests that routine malaria chemoprophylaxis is not recommended for people with sickle cell disease in malaria-endemic areas, as the risks associated with long-term prophylaxis may outweigh the benefits, and other preventive measures such as insecticide-treated bed nets and prompt treatment of malaria episodes are preferred. [**predicted**: **Refuted**]
Mistral 24B	Routine malaria chemoprophylaxis is generally recommended for people with sickle cell disease living in malaria-endemic areas. This is supported by clinical research indicating that malaria infections can trigger sickle cell crises and other complications. Chemoprophylaxis helps reduce the risk of malaria, thereby decreasing the frequency of these crises and improving overall health outcomes. [**predicted**: **Supported**]

Table 7: Example of a question and two labels, one outdated and one recent from our dataset. Answers show how GPT and Mistral capture the latest and up-to-date knowledge for this question, while Llama predicted incorrect and outdated recommendations.

Question	Do probiotics reduce the risk of necrotising enterocolitis in very preterm or low birth weight infants?
Conclusion (Outdated)	Enteral supplementation of probiotics prevents severe NEC and all cause mortality in preterm infants. Our updated review of available evidence strongly supports a change in practice. (...) (AlFaleh et al., 2014) [**label**: **Supported**]
Conclusion (Latest)	Given the low to moderate certainty of evidence for the effects of probiotic supplements on the risk of NEC and associated morbidity and mortality for very preterm or VLBW infants, and particularly for extremely preterm or ELBW infants, there is a need for further large, high-quality trials to provide evidence of sufficient validity and applicability to inform policy and practice. (Sharif et al., 2023) [**label**: **Not Enough Information**]
Mistral 24B	Several clinical studies and meta-analyses have shown that probiotic supplementation in very preterm or low birth weight infants can reduce the risk of necrotizing enterocolitis (NEC). A **Cochrane review from 2014** has supported this finding, indicating that probiotics can be beneficial in preventing NEC in this vulnerable population. [**predicted**: **Supported**]

Table 8: Example of a study objective that had the *Supported* label in its 2014 version and an updated *Not enough information* label in its 2023 version. Mistral cited the more than decade-old study to supports its answer.

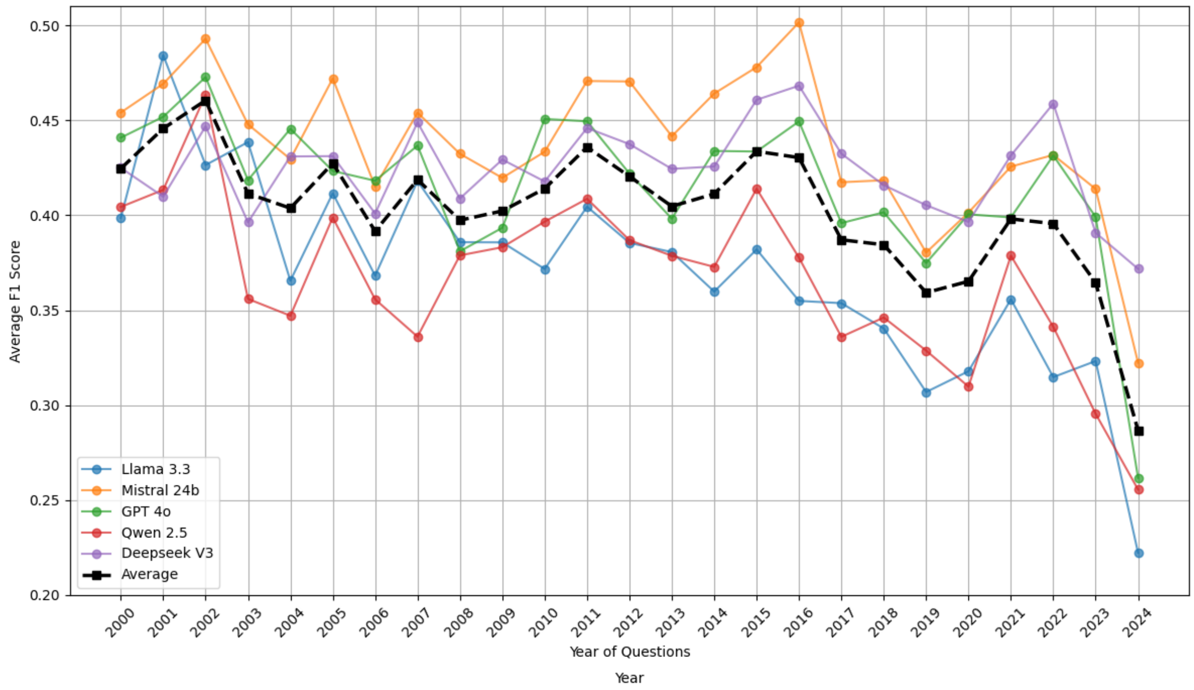

Figure 4: Larger version of the previous figure (Fig. 1): Average F1-Macro performance for questions originating from each year in the dataset across five LLMs, showing decline in more recent years.