Do Knowledge Cutoffs Drive Clinical Accuracy and User Trust? Quantifying Temporal Decay in Large Language Models

Michael Cacioli[1], **Aryan Arya**[2], **Austen Liao**[3,4], **Kevin Zhu**[4]
[1]Gilmour Academy
[2]Oregon State University
[3]Johns Hopkins University
[4]Algoverse AI Research
Correspondence: michael.cacioli2008@gmail.com

Abstract

LLMs are increasingly becoming part of everyday clinical decisions, yet they are trained on datasets that were static long before they were released into the world. The time gap between when the model was trained and when the model is used, referred to as a knowledge cutoff, has proven to be a subtle but crucial failure mode. The model may be capable and aligned, yet offer out-of-date medical advice with absolute fluency. We want to investigate what difference the degree of data freshness alone can make in a model's clinical performance. We isolated the cutoff parameter between two model families with differing release patterns: the closed-weight GPT models and the open-weight LLaMA models, using data from two dated versions of the IDSA COVID-19 Treatment and Management Guidelines (v5.0.0, 8/25/21; v11.0.0, 6/26/23). We evaluated recommendation-level differences and generated 363 multiple-choice questions that represented true changes in treatment advice. All models were queried using the exact same prompts and deterministic settings on the same questions. We see that accuracy only jumps up when the assumed training cutoff of the model falls after the date of the newer guideline. Both GPT-3.5-Turbo and LLaMA-2-13B, with their older cutoffs, fail to match the accuracy of models with cutoffs later than 6/26/23, whereas the more recently trained models GPT-4o, GPT-5, and LLaMA-3.3-70B achieve over 90% accuracy and show similar results. This demonstrates the success of using fresh data and that this alone leads to improvements in applied medical reasoning and is more than just a count of parameters. The users of these systems may have an outsized trust in their fluency even if the information they convey is not up to date, especially in sensitive or stressful conditions.

1 Introduction

Language models have become a central component of modern clinical and biomedical research. They assist with summarizing evidence, generating differential diagnoses, and supporting patient communication. These systems are increasingly integrated into search platforms, clinical documentation tools, and medical education resources. Their rapid adoption reflects the promise of scalable decision-support, yet it also introduces a new form of technical debt: models are built on static data that freeze the medical record at a single point in time. Once deployed, they cannot automatically absorb updates to scientific consensus or treatment guidelines. In clinical contexts, this limitation carries direct implications for safety and reliability.

Medical knowledge changes more rapidly than most general purpose knowledge sources. The current standard of therapy may be updated every few months based on a new clinical trial or meta-analysis. We saw the pace of medical knowledge updates during the COVID-19 pandemic. Between early and mid-2023, the recommendations for steroid use, antiviral indication, and monoclonal antibody use changed significantly between updates of the IDSA COVID-19 guidelines. The output of a language model trained before early 2023 could contain treatment recommendations that had already been withdrawn from practice. As a language model generates recommendations with fluency and certainty, patients or physicians may incorrectly give greater weight to that information simply because it is well-written, rather than because it reflects current practice.

This phenomenon highlights a central challenge in evaluating medical language models: temporal reliability. Most existing benchmarks emphasize reasoning quality, factual precision, or bias mitigation, while the temporal dimension of knowl-

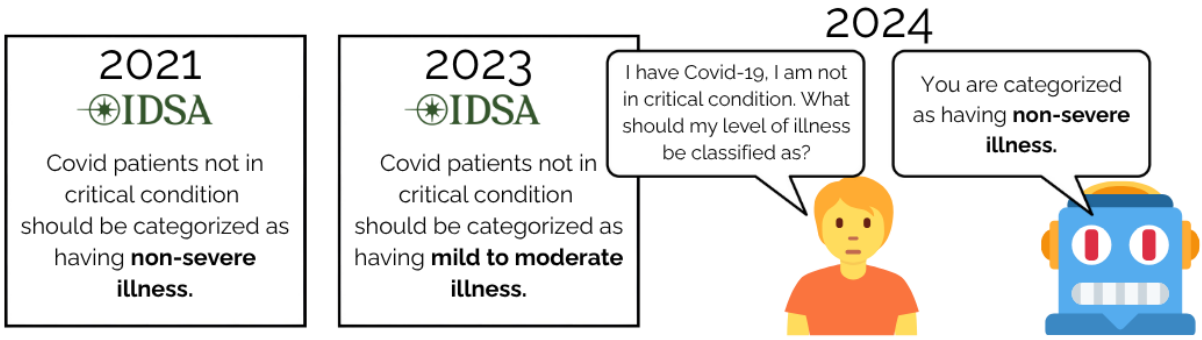

Figure 1: Example of guideline drift and model obsolescence. IDSA definitions for COVID-19 illness severity changed from "non-severe" (2021) to "mild to moderate" (2023). A model trained before this update continues to give outdated terminology in 2024, illustrating how cutoff limitations can produce clinically misleading advice.

edge—how current or obsolete a model's information is—receives less attention. The training cutoff, often briefly noted in technical documentation, represents a boundary between what a model can know and what it cannot. Yet the practical effect of this boundary on medical performance has not been systematically quantified. If a model's knowledge decays as guidelines change, its apparent reasoning ability may mask clinically significant obsolescence.

Grasping this connection is important for proper deployment. It needs to be understood by hospitals, regulators and developers how much of the information behind a model is old enough that retraining or replacement should occur. Otherwise, measures can overstate safety. Temporal coverage is itself a measurable and reportable aspect of model design, like the bias of a model or its interpretable components or scale of parameters.

This paper examines how knowledge cutoffs influence clinical accuracy. Using successive versions of the IDSA COVID-19 guidelines as a controlled benchmark, we measure how model performance shifts across systems released at different times. By isolating data recency from other confounding variables such as architecture and alignment, we show that the freshness of training data is a key determinant of a model's ability to reflect current medical standards. The results contribute to a broader understanding of temporal validity in clinical language models and underscore the need for continual monitoring of information currency in safety-critical domains.

2 Related Works

Research on temporal behavior in large language models has developed along three main directions: documenting model data provenance, identifying factual decay over time, and designing continual learning strategies to maintain knowledge freshness. Together, these efforts reveal how the temporal scope of training corpora shapes downstream reliability. Even so, they stop short of fully quantifying that effect in safety-critical contexts such as medicine.

Attempts to document and track training data have also highlighted the black-box nature of current model pipelines. Analyses of web-scale corpora such as CCNet and RefinedWeb revealed skewed coverage across languages and widely differing time distributions even within a single crawl (Wenzek et al. (2020); Penedo et al. (2023)), and follow-up studies indicated significant drift in topics and domain proportions between different versions of the dataset, making "recency" difficult to define or track. While works such as TimeLMs (Loureiro et al., 2022) and the framework proposed by Dhingra et al. (2022) highlight the potential of explicitly embedding time information in training data, they are limited to public datasets, and even the broader documentation efforts proposed in Datasheets for Datasets and Model Cards for Model Reporting (Gebru et al., 2021; Mitchell et al., 2019) do not typically focus on time information.

Parallel work examines temporal drift and factual obsolescence in model outputs. Benchmarks such as FreshQA and TempLAMA evaluate how performance declines when reference facts are updated—an issue closely related to temporal decay in factual consistency ((Vu et al., 2024; Dhingra

et al., 2022)). These studies show that models maintain internal consistency even when their answers contradict new evidence, implying that decay operates silently rather than through explicit uncertainty.

Another dimension focuses on structural approaches that address continuous and life-long learning. The foundations for continuous and lifelong learning can be traced back to work on gradient episodic memory and elastic weight consolidation, laying the theoretical groundwork to learn new knowledge without compromising old knowledge (Lopez-Paz and Ranzato, 2017; Kirkpatrick et al., 2017). Modern approaches adapt these techniques in the transformer world using adapters and parameter-efficient fine-tuning methods, such as parameter-efficient transfer learning and LoRA (Houlsby et al., 2019; Hu et al., 2022). While these are successful in certain environments, they assume access to timestamped data streams, a setup that usually does not exist within closed commercial systems. Without continuous adaptation, model knowledge is effectively frozen in time.

In the clinical domain, language models have been evaluated primarily for diagnostic reasoning and information retrieval. Benchmarks like MedMCQA and MultiMedQA demonstrate that large models encode substantial medical knowledge (Pal et al., 2022; Singhal et al., 2023), while domain-specialized systems such as BioMedLM and GatorTron show that further tuning on biomedical text yields measurable gains in precision (Bolton et al., 2024; Yang et al., 2022). Yet these studies rarely interrogate when the information used by a model was last valid.

In all, prior work has illustrated that language models are capable of high performance on static reasoning while being worse on current state of the world knowledge. We still have to understand the extent to which static accuracy in a language model relies on the training data cutoff. We present the first work that solely varies the knowledge cutoff and shows that without changing the architecture of the model or its alignment, the main factor predicting clinical performance is time coverage.

3 Methodology

This study evaluates the impact of knowledge cutoffs on clinical accuracy by isolating temporal coverage as the only independent variable across two large language model families: OpenAI's closed-weight GPT series and Meta's open-weight LLaMA series. The experiment design deliberately holds model architecture, prompting, and evaluation setup constant to ensure that any observed variation in performance arises primarily from differences in effective temporal coverage.

3.1 Guideline Selection and Temporal Framing

In order to build a temporally grounded benchmark, two publicly available editions of the Infectious Diseases Society of America (IDSA) COVID-19 Treatment and Management Guidelines were selected: 5.0.0 (August 25, 2021) and 11.0.0 (June 26, 2023). Version 5.0.0 was released prior to the training cutoff date of every model we tested, and version 11.0.0 was released only after the oldest model in each of our model families. This arrangement allows us to objectively assess temporal validity by determining whether or not the internal knowledge of the models has caught up to events occurring after the training cutoff.

3.2 Difference Extraction and Question Generation

A programmatic comparison was made for each set of guideline versions. The goal was to isolate changes at the recommendation level, meaning changes that actually alter clinical practice, as opposed to purely textual changes at the surface level. A custom parser was developed to identify recommendation descriptions in three categories: treatment modality, medication eligibility, and dosage recommendations. For each unique difference, a multiple choice question (MCQ) was automatically generated, with a single correct answer, the latest recommended change, and three incorrect answers, or distractors, extracted from outdated or deprecated guideline statements. This resulted in a 363 MCQ database that truly reflects the clinical change.

Not all text-based differences were selected for benchmark items, because updates to guidelines typically address shifts in both treatment recommendation and evidence quality/strength, conditional language, and patient groups. Accordingly, we selected for inclusion using the following criteria when creating the MCQs. A guideline differ-

ence was selected if it altered a decision-making action, e.g., which drug or intervention should be recommended, which patients were suitable for a certain treatment, when a treatment or intervention should be performed, or what dosage. For example, if a previous recommendation said "Patients should receive treatment X", and an updated guideline recommended "Patients may receive treatment X," that update would not have been selected for the benchmark. We ensured that any criteria on which the decision was based, e.g., hospitalization status, disease severity, oxygen levels required, symptom timing, and risk profile, were addressed in the question, and that there was one unique correct answer for the question and no alternative.

All identified differences and corresponding questions underwent a manual verification audit to confirm that each item was unambiguous, clinically valid, and that all distractors accurately reflected superseded recommendations.

Representational limits of MCQ diffs. Although each question rests upon a documented recommendation-level change, updates to guidelines often do not appear as atomic "fact flips." In many cases, IDSA changes include adjustments in the strength of evidence, conditionality, or patient subgroups that cannot easily be mapped into a multiple-choice format without significant loss of nuance. Hence, we regard each MCQ as a *diagnostic probe* of the model's ability to recall the updated recommendation given the item's conditions, rather than as a test of its overall guideline reasoning. In Section 6.1, we propose ways to evaluate models in light of conditionality in future benchmarks.

3.3 Model Families and Evaluation Protocol

Three models were evaluated from each family to capture pre and post-cutoff behavior:

- **GPT Family:** GPT-3.5-Turbo, GPT-4o, and GPT-5
- **LLaMA Family:** LLaMA-2-13B-hf, Llama-3.3-70B-Instruct, and Llama-4-Scout-17B-16E-Instruct

For each model, all 363 MCQs were presented under identical deterministic prompting conditions to ensure reproducibility. Each model evaluated was asked each of the 363 questions. Each response was then parsed and compared against the reference key.

3.4 Evaluation and Scoring

Inference configuration. All models were evaluated on the same 363 items using a fixed multiple-choice prompt and deterministic decoding. Deterministic decoding isolates temporal coverage effects from sampling variance and makes scores exactly reproducible under identical settings. Hosted models were accessed via provider APIs; open-weight models were evaluated using an endpoint.

Output constraint and parsing. Models were instructed to output *only* a single option letter (A–D). Outputs were scored by selecting the first occurrence of a standalone option letter in A,B,C,D.

What "cutoff" means in modern pipelines. Our analysis uses publicly documented or widely reported training windows as a proxy for temporal exposure. We emphasize that modern pipelines are not a clean binary function of a pretraining cutoff: supervised finetuning, RLHF traces, retrieval-augmented distillation, synthetic preference data, and contamination from web snapshots may leak post-cutoff knowledge in non-obvious ways. Consequently, we interpret the observed performance discontinuities as evidence of *effective temporal coverage* rather than a strict causal claim about pretraining boundaries. We revisit this uncertainty explicitly in the Limitations and Ethics sections.

Why deterministic evaluation is still informative. Deterministic evaluation intentionally underestimates variance seen in deployed settings, where prompt phrasing and stochastic decoding can amplify temporal drift. We therefore treat the deterministic benchmark as a lower bound on instability risk, and we outline stability-focused follow-up measurements in Future Work.

3.5 Temporal Validation Hypothesis

The primary hypothesis is that models trained prior to June 2023 will do worse on recommendations based on Version 11.0.0 because those recommendations had not yet been created in the dataset used for training. However, models trained after June 2023 should demonstrate similar performance across guidelines because temporal cover-

age should have become saturated. Seeing this saturation occur in both open and closed-weight systems would support the idea that the amount of recent training data, rather than model capability or alignment, dictates performance in time-sensitive clinical reasoning.

4 Representative Benchmark

We include representative items for transparency. Each question is derived from a verified recommendation-level change between IDSA guideline versions.

Question: According to the IDSA COVID-19 Treatment Guidelines version 11.0.0, which antiviral therapy is now recommended as first-line for high-risk, nonhospitalized adults with mild-to-moderate COVID-19 when started within 5 days of symptom onset?

(A) Remdesivir (3-day IV regimen)

(B) Nirmatrelvir/ritonavir (Paxlovid)

(C) Hydroxychloroquine

(D) Molnupiravir

Answer: (B)

5 Results

Table 1 presents the quantitative performance of all six evaluated models across the 363-question benchmark derived from the IDSA COVID-19 Treatment and Management Guidelines. Each question represented a verified update in medical consensus between Version 5.0.0 (August 25, 2021) and Version 11.0.0 (June 26, 2023), enabling a direct measurement of how each model's knowledge recency aligned with modern therapeutic standards. Because all models were tested under identical deterministic settings, differences in outcome are most consistent with differences in effective temporal coverage.

Across both model families, a clear temporal inflection was observed. Models trained prior to June 2023 performed markedly worse on items reflecting later guideline updates, while those trained afterward demonstrated near-saturated performance. Within the GPT family, GPT-3.5-Turbo—whose training data predated Version 11.0.0—achieved an overall accuracy of 76.03%. In contrast, GPT-4o and GPT-5, which both postdate the 2023 guideline release, scored 97.25% and 98.07%, respectively. The gain of over twenty percentage points indicates that temporal data inclusion, rather than parameter scale or minor alignment improvements, accounts for the majority of the performance increase.

GPT-3.5-Turbo's relatively high 76.03% accuracy can be attributed largely to inference rather than the actual information it stores. While 76.03% is certainly impressive as a piece of model inference, it is not high enough in a clinical setting to be relied on, let alone trusted. The fact that the main difference in accuracy between the pre-guidelines model and the post-guidelines models comes from knowledge cutoff, rather than things like parameter size or model ability, can be shown with the use of another post-dating model. We utilize GPT-5 here, a post-dating model that is significantly more capable than GPT-4o yet scores nearly exactly the same, while GPT-3.5-Turbo scored far worse. This supports the idea that the difference comes from the knowledge cutoff and not sheer ability.

A similar trend was seen within the open-weight LLaMA models. While the knowledge cutoff of LLaMA-2-13B-hf was long before v11.0.0, and so its accuracy was only 35.26%, the newer models LLaMA-3.3-70B-Instruct and LLaMA-4-Scout-17B-16E-Instruct were able to achieve accuracies of 94.77% and 91.46%, respectively. This nearly 60 percentage point increase across both families of models supports the observation from the GPT models that models with older knowledge tend to be less reliable, even across different architectures.

To ensure robustness, every model was evaluated on the same 363 questions, with responses parsed automatically to extract the selected choice and matched against the reference key. No stochastic variation was introduced, so each reported accuracy represents a deterministic outcome reproducible under identical conditions. Accuracy distributions displayed minimal variance within post-cutoff models, suggesting that once exposure to the updated medical corpus is achieved, performance converges regardless of further scale or parameter growth.

Both families of models share the same curve trend: they show sharp improvement in performance when post-June 2023 training data is added, and reach steady performance in the next step. This transition is achieved regardless of the size of the

Model Family	Model	Accuracy (%)
Closed-weight (OpenAI GPT Series)		
GPT Series	GPT-3.5-Turbo	76.03
	GPT-4o	97.25
	GPT-5	98.07
Open-weight (Meta LLaMA Series)		
LLaMA Series	LLaMA-2-13B-hf	35.26
	LLaMA-3.3-70B-Instruct	94.77
	LLaMA-4-Scout-17B-16E-Instruct	91.46

Table 1: Accuracy of GPT and LLaMA model families across 363 automatically generated clinical multiple-choice questions derived from IDSA guideline updates. Each result reflects deterministic evaluation

model, suggesting that new training data more strongly impacts performance on medical question answering compared to model size and alignment procedures.

6 Analysis

The quantitative results in Table 1 reveal a clear medical and clinical trend rather than a purely computational one. Across 363 IDSA-derived clinical questions, model accuracy improved sharply once training data included the June 2023 guideline revision. This outcome shows that the models' ability to reason clinically depends less on scale and more on exposure to current medical evidence. In practical terms, temporal recency becomes a clinical determinant of reliability, not just a technical variable.

The convergence of 95% and 98% accuracy across both families, after time alignment, highlights that clinical reliability is indeed tied to recency of knowledge. Once a model is allowed to drift out of alignment with updated guidelines, performance drops as though clinical training has been allowed to expire. The study measures how far out of date that training would have had to become, with roughly 60-points of accuracy lost if the model cutoff predates the new guidelines. This should be a solid baseline for anyone using these models in a clinical setting.

6.1 Limitations

This study is designed to isolate temporal coverage as cleanly as possible, but several limitations are important for interpretation.

Non-randomized multiple-choice structure. Although each item was manually verified for clinical validity, multiple-choice formats can introduce positional or formatting bias (e.g., a preference for earlier options, or sensitivity to how distractors are worded).

Single deterministic evaluation pass. All results in this paper are generated with deterministic decoding ($T = 0$) in one pass per model. This strategy eliminates sampling variance and isolates comparisons of time coverage. This approach does not measure stability from one run to another. Subsequent work should include mean $\pm$ standard deviation across seeds and temperatures, and item-level flip rates to measure whether instability is due to a small set of borderline questions.

Closed-model mutability over time. For hosted proprietary models, behavior may change over time even when the product name remains constant (e.g., backend updates, safety patches, or silent weight refreshes). Longitudinal re-evaluation and version-pinned identifiers, when available, would help to distinguish temporal drift in the model from drift in the underlying medical evidence.

6.2 Ethical Statement

This research uses only publicly available medical text and does not involve human subjects, identifiable data, or protected health information. None of the findings should be used for clinical decision-making. The work aims solely to advance scientific understanding of how temporal data integrity affects the safety and reliability of medical language

models.

7 Future Work

Future studies should not only extend this validation framework to new medical areas including oncology, cardiology, and psychiatry, where disease processes may also be characterized by their own decay trends, but also move to a longitudinal setup that allows evaluating at what speed model accuracy degrades when the set of clinical guidelines in circulation keeps changing. Concurrently, work must also start by studying methods to update models dynamically, whether through approaches like continual learning or retrieval, complemented with reinforcement approaches, in order to close the gap between static pretraining and a living medical knowledge base.

Beyond technical development, collaboration between computational scientists and clinical experts will be essential. Future benchmarks must not only assess factual accuracy but also measure downstream clinical safety and interpretability. Ultimately, the goal is to ensure that medical language models serve as trustworthy extensions of human judgment rather than outdated archives of past consensus.

Future work should explore the real-world understanding of temporally stale outputs, especially if it can demonstrate that an increase in fluency and/or confidence leads to greater over-trust. Understanding this interaction is key to the safe deployment of language systems in both the clinical and mental health domain.

8 Conclusion

In our 363 medically vetted IDSA guideline questions, we show that temporal recency of data appears to be a stronger determinant than model size or architecture in this benchmark of a language model's clinical reasoning ability. The GPT and LLaMA families of models displayed identical temporal inflection at our cutoff, with rapid improvements once data from after June 2023 was added, before flattening out near expert-level performance. The finding was identical across models of differing weight types, open and closed-weight, showing that model design does not determine clinical trustworthiness; rather, the medical recency of the trained data does.

These outcomes have real-world clinical implications. Not only does a model trained on outdated data perform poorly, it can be hazardous. The 60-point difference between pre- and post-cutoff models illustrates how obsolescence translates to an actual risk of diagnostic misinterpretation. It is critical that models remain congruent with clinical practices, making this another instance of biomedical maintenance, where frequent retraining, constant validation on updated clinical practices, and regular temporal auditing become standard requirements prior to clinical integration.

Data and Code Availability

All data, generated questions, and evaluation code used in this study are publicly available at: `https://huggingface.co/datasets/michaelcacioli/LLM-Covid-19-Cutoff-Evaluation`.

The repository includes the full set of 363 clinical multiple-choice questions generated from the IDSA guideline differences, along with the parsing and evaluation scripts used for deterministic model benchmarking.

References

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text.

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*, pages 257–273.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, pages 86–92.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mikael Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pages 2790–2799, Long Beach, CA, USA. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of

large language models. In *International Conference on Learning Representations*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229, Atlanta, GA, USA. Association for Computing Machinery.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the 3rd ACM Conference on Health, Inference, and Learning (CHIL 2022)*, pages 248–260, Virtual Event, USA. Proceedings of Machine Learning Research (PMLR).

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, New Orleans, LA, USA. Curran Associates, Inc.

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, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, and 12 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

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. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4003–4012, Marseille, France. European Language Resources Association (ELRA).

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

A Additional Benchmark Examples

To provide further transparency into the evaluation dataset, we include additional representative questions derived from IDSA guideline updates.

Example 1

Question: According to the updated IDSA COVID-19 treatment guidelines, what is the current recommendation regarding the use of ivermectin for hospitalized or non-hospitalized patients with COVID-19?

(A) Suggested only for hospitalized patients requiring supplemental oxygen

(B) Recommended routinely for all patients with mild to moderate COVID-19

(C) Recommended only within the context of a clinical trial

(D) Recommended as an adjunct to antiviral therapy in high-risk outpatients

Answer: (C)

Example 2

Question: According to the June 2023 (v11.0.0) IDSA COVID-19 guidelines, which antiviral regimen is now preferred for treating nonhospitalized adults with mild-to-moderate COVID-19 at high risk for progression to severe disease?

(A) Hydroxychloroquine with azithromycin

(B) Nirmatrelvir/ritonavir oral course within 5 days of symptom onset

(C) Remdesivir 5-day IV course as first-line therapy

(D) Ivermectin single-dose therapy

Answer: (B)

These examples illustrate the type of recommendation-level changes captured in the dataset.

B Error Analysis

To better understand model failure modes, we manually inspected a subset of incorrect responses across models.

We observe three dominant error categories:

1. Temporal hallucination. Models with earlier knowledge cutoffs confidently generated outdated recommendations that were previously valid but have since been superseded.

2. Option bias. Certain models, particularly LLaMA-2-13B, showed a tendency to favor earlier answer choices (e.g., option A), suggesting alignment or decoding biases rather than true reasoning.

3. Partial knowledge overlap. Some incorrect responses reflected partial incorporation of updated knowledge, where models mixed older and newer guidelines inconsistently.

These findings reinforce that performance degradation is not random but systematically linked to temporal misalignment between training data and current medical standards.

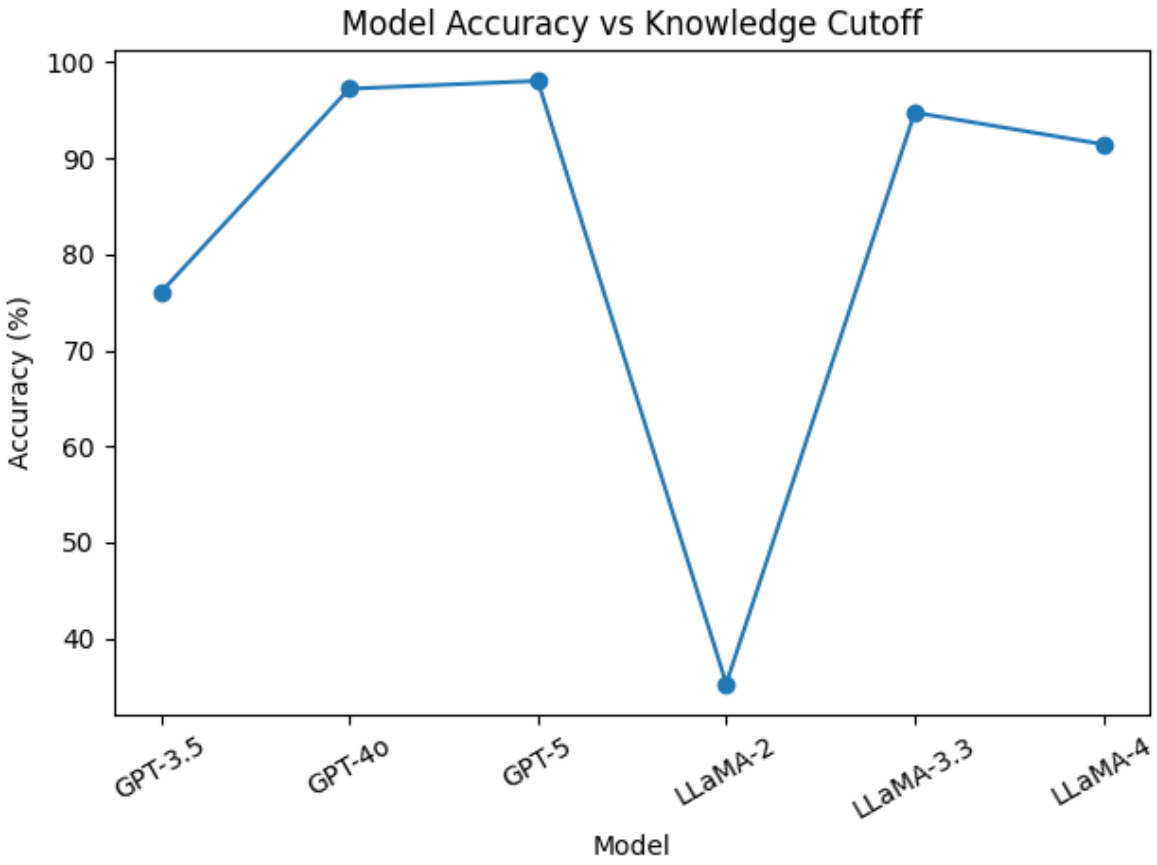

Figure 2: Model accuracy across GPT and LLaMA families as a function of knowledge cutoff. A sharp performance inflection occurs once models include post–June 2023 training data, followed by convergence near expert-level accuracy.