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

Michael Cacioli[1,4] **Aryan Arya**[2,4] **Austen Liao**[3,4] **Kevin Zhu**[4]

[1] Gilmour Academy
[2] Oregon State University
[3] Johns Hopkins University
[4] Algoverse AI Research

michael.cacioli2008@gmail.com
aaworks2039@gmail.com
austen@algoverseairesearch.org
kevin@algoverseacademy.com

Abstract

Modern clinical decisions increasingly depend on large language models (LLMs). Yet, these models are built on static training data that end long before deployment. This temporal gap between training and use, commonly described as a knowledge cutoff, creates a hidden yet critical failure mode. A model may be capable and aligned, yet still apply outdated medical guidance with perfect fluency. To test how much data freshness alone affects clinical accuracy, this study isolates the cutoff variable across two model families with different release patterns: OpenAI's closed-weight GPT models and Meta's open-weight LLaMA series. Using two dated versions of the Infectious Diseases Society of America (IDSA) COVID-19 Treatment and Management Guidelines (v5.0.0, August 25, 2021; v11.0.0, June 26 2023), we extracted recommendation-level differences and automatically generated 363 multiple-choice questions representing genuine shifts in therapeutic advice. Each model answered the same items under identical prompts and deterministic settings. Accuracy rose sharply only when the model's presumed training window included the newer guideline. GPT-3.5-Turbo and LLaMA-2-13B, whose cutoffs pre-date June 2023, significantly lagged behind models whose knowledge cutoffs post-dated v11.0.0. GPT-4o, GPT-5, and LLaMA-3.3-70B, trained on fresher data, converged at over 90%. The consistency of this pattern across closed and open systems indicates that temporal coverage, not mere parameter count, drives gains in applied medical reasoning. These findings argue that model recency must be treated as a safety-critical attribute on par with alignment or interpretability.

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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

absorb updates to scientific consensus or treatment guidelines. In clinical contexts, this limitation carries direct implications for safety and reliability.

Medical knowledge changes faster than most general-purpose data sources. Therapeutic standards often evolve within months, as new trials or meta-analyses revise earlier recommendations. The COVID-19 pandemic demonstrated how rapidly these changes can occur. Between early 2023 and mid-2023, updates to the Infectious Diseases Society of America (IDSA) COVID-19 guidelines substantially altered recommendations for corticosteroid use, antiviral eligibility, and monoclonal antibody therapy. A language model trained before those updates would likely reproduce advice that had already been withdrawn from practice. Because model outputs are phrased with confidence and fluency, clinicians and patients may overestimate their validity even when the information is outdated.

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 knowledge—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.

Understanding this relationship is essential for responsible deployment. Health institutions, regulators, and developers must be able to anticipate when a model's information base becomes stale enough to warrant retraining or replacement. Without such analysis, performance metrics can give a false impression of safety. Temporal coverage should therefore be treated as a measurable, reportable property of model design, comparable in importance to bias, interpretability, or parameter scale.

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.

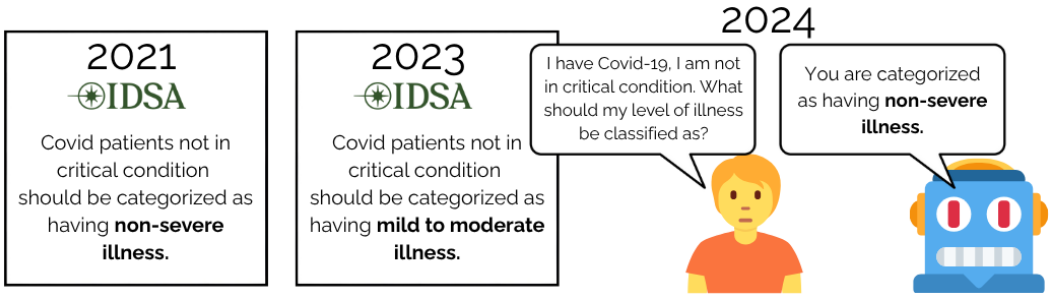

Figure 1: Example of guideline drift and model obsolescence. IDSA definitions for COVID-19 illness severity changed from "non-severe" (2021) to "mild to moderate" (2023). A model trained before this update continues to give outdated terminology in 2024, illustrating how cutoff limitations can produce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

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.

Efforts to document and trace training data have underscored the opacity of current model pipelines. Studies of web-scale corpora such as CCNet and RefinedWeb show that language coverage and timestamp distribution vary widely even within a single crawl (Wenzek et al. [2020], Penedo et al.

[2023]). Follow-up analyses demonstrated that later iterations of these datasets contain substantial drift in domain balance and topic prevalence, suggesting that "recency" is neither uniform nor easily verified. Time-aware corpora such as TimeLMs and the temporal modeling framework of Dhingra et al. [2022] illustrate the potential for explicitly encoding temporal change, though these methods remain limited to open datasets. Broader transparency efforts, including Datasheets for Datasets and Model Cards for Model Reporting [Gebru et al., 2021, Mitchell et al., 2019], emphasize documentation standards but rarely include explicit temporal metadata.

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 (Guo et al. [2017], Desai and Durrett [2020]). 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.

A third thread explores continual and lifelong learning as structural remedies. Foundational work on gradient episodic memory and elastic weight consolidation established the theoretical groundwork for preserving old knowledge while incorporating new data (Lopez-Paz and Ranzato [2017], Kirkpatrick et al. [2017]). More recent transformer-based approaches adapt these ideas with modular adapters and efficient fine-tuning methods such as parameter-efficient transfer learning and LoRA (Houlsby et al. [2019], Hu et al. [2022]). Although effective in controlled settings, these methods presuppose access to timestamped data streams—conditions rarely met in closed commercial systems. The absence of such continual adaptation effectively fixes a model's worldview to a specific moment 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. Time-aware evaluations like PubMedQA highlight that even high-performing models often reference outdated research, underscoring that factual recency is as vital as reasoning quality (Jin et al. [2019]).

Taken together, prior work demonstrates that language models can excel at static reasoning while simultaneously falling behind the evolving state of the world. What remains missing is an explicit quantification of how much of a model's accuracy depends on the boundary of its training data. The present paper addresses that gap by isolating knowledge cutoff as the sole variable and showing that, even when model architecture and alignment remain constant, temporal coverage alone predicts clinical reliability.

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 solely from differences in training data recency.

3.1 Guideline Selection and Temporal Framing

To construct a temporally grounded benchmark, two publicly available versions of the Infectious Diseases Society of America (IDSA) COVID-19 Treatment and Management Guidelines were selected: Version 5.0.0 (August 25, 2021) and Version 11.0.0 (June 26, 2023). The earlier version predates the training cutoff of all evaluated models, while the latter postdates only the oldest model in each family. This configuration enables controlled measurement of temporal validity, seeing whether or not a model's internal knowledge reflects updates introduced after its training boundary.

3.2 Difference Extraction and Question Generation

Each pair of guideline versions were programmatically compared to extract clinically meaningful recommendation-level changes rather than surface-level textual edits. A custom parser identified sections describing treatment, medication eligibility, and dosage recommendations. For each detected difference, a multiple-choice question (MCQ) was automatically generated with one correct answer (the updated recommendation) and three distractors drawn from prior or deprecated statements. This procedure yielded 363 MCQs that represent genuine shifts in clinical consensus.

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.

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 Interpretation Procedure

Model performance was quantified using a direct accuracy benchmark derived from the multiple-choice question set. Each model's score was defined as the percentage of correctly selected answers out of all 363 items. No weighting or normalization was applied. Each question contributed equally to the total accuracy. Scores were then compared within and across the GPT and LLaMA families to visualize trends associated with each model's knowledge cutoff. This design isolates the relationship between data recency and clinical correctness. An abrupt rise in accuracy between pre and post-cutoff models, followed by convergence in later generations, would confirm that temporal coverage rather than architectural scale or parameter count primarily drives observed gains in medical reasoning performance.

3.5 Temporal Validation Hypothesis

The core hypothesis is that models trained before June 2023 will underperform on items derived from Version 11.0.0, as those recommendations were not part of their training data. Conversely, models trained after that date should display comparable performance across both guideline versions, indicating saturation of temporal coverage. Observing this convergence pattern across both closed and open-weight systems would confirm that data freshness, rather than model capacity or alignment, is the primary determinant of clinical accuracy in time-sensitive reasoning tasks.

4 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 attributable solely to the temporal boundaries of their training data.

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

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

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.

The relatively high 76.03% accuracy from GPT-3.5-Turbo can be largely attributed to model inference. However, in the clinical setting, 76.03% is no where near high enough to be considered effective and safe. The fact that the major accuracy discrepancy between the models that predate and postdate the newer version of the guidelines comes from knowledge cutoffs and not other factors, such as parameters and general model capacity, can be seen with the use of a second post-dating model. In GPT's scenario, we use GPT-5, a model smarter and more capable than GPT-4o. Despite this, they score almost exactly the same, while GPT-3.5-Turbo scores significantly worse. This is further evidence of this discrepancy being a result of LLM knowledge cutoffs, not purely reasoning capability.

The same pattern emerged in the open-weight LLaMA models. LLaMA-2-13B-hf, having a knowledge cutoff well before v11.0.0, achieved only 35.26% accuracy. Subsequent generations trained on later data, LLaMA-3.3-70B-Instruct and LLaMA-4-Scout-17B-16E-Instruct, reached 94.77% and 91.46%, respectively. This increase of nearly sixty percentage points mirrors the temporal effect observed in the GPT family, providing strong cross-architecture evidence that clinical reliability improves as models incorporate newer knowledge.

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 model families exhibit a similar trajectory: a steep increase in accuracy coinciding with the inclusion of post-June 2023 training data, followed by a plateau in later iterations. This temporal transition occurs independently of model size, indicating that training data freshness exerts a more substantial influence on medical question-answering accuracy than architectural complexity or alignment refinements.

5 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.

Within the **GPT family**, performance climbed from 76.03% in GPT-3.5-Turbo to 97.25% in GPT-4o and 98.07% in GPT-5. The near-identical scores between the two newer models suggest that once training incorporates updated clinical guidance, further scaling provides little additional benefit. These systems appear to have reached a ceiling defined by their access to contemporary medical data. In a healthcare setting, this level of stability implies that periodic data refreshes are more important for patient safety than increasing model complexity.

The **LLaMA family** displayed a far more dramatic contrast. LLaMA-2-13B-hf, trained before the v11.0.0 guideline, reached only 35.26% accuracy, and manual inspection showed that it defaulted to option "A" for most questions. This behavior points to deterministic alignment bias rather than genuine comprehension. When newer guideline data were introduced in later generations, accuracy surged to 94.77% for LLaMA-3.3-70B and 91.46% for LLaMA-4-Scout-17B-16E. That recovery represents a clinically meaningful transformation in reasoning fidelity. The models effectively went from unreliable to near-expert accuracy solely because their training reflected updated medical consensus.

Across both families, the convergence between 95% and 98% accuracy after temporal alignment confirms that knowledge recency dictates clinical reliability. This finding reframes model improvement as a biomedical maintenance problem: updating corpora is akin to renewing a medical license. Once a model falls out of sync with guideline evolution, its accuracy degrades as if clinical training had expired. These results quantify that effect in practical terms, about a 60-point deficit when the cutoff predates new recommendations. That number should serve as a concrete retraining benchmark for anyone deploying models in a clinical environment.

5.1 Limitations

Although the experiment isolates temporal effects more directly than previous research, it has two key limitations. First, the precise cutoff points for model training were inferred from public release information rather than verified pretraining logs, introducing minor uncertainty about exact exposure windows. Second, the evaluation focused on a single clinical area: infectious disease management under IDSA guidelines. This scope was chosen because COVID-19 guidelines evolve rapidly and reflect real changes in patient care, but findings may vary in slower-changing fields such as cardiology or oncology. Future validation should apply this framework to other medical specialties to determine how temporal decay manifests across disciplines with different rates of evidence turnover.

5.2 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.

6 Conclusion

This study demonstrates that temporal data recency, rather than model size or architecture, is the dominant factor determining clinical reasoning accuracy in large language models. Across 363 medically verified IDSA guideline questions, both the GPT and LLaMA families exhibited the same temporal inflection: accuracy rose sharply once post-June 2023 data were included, then plateaued near expert-level performance. The consistency of this finding across closed- and open-weight models confirms that the determinant of clinical reliability lies not in the model's design but in the medical freshness of its training corpus.

From a clinical perspective, these results carry practical consequences. A model trained on outdated data does not simply perform worse; it becomes unsafe. The 60-point deficit between pre- and post-cutoff models quantifies how obsolescence translates into real diagnostic risk. Maintaining alignment with current clinical standards must therefore be treated as a form of biomedical upkeep. Periodic retraining, ongoing validation against updated guidelines, and systematic temporal audits should be mandatory steps before any model is integrated into a healthcare workflow.

7 Future Work

Future research should expand this evaluation framework beyond infectious disease and COVID-19 guidance to encompass additional medical domains such as oncology, cardiology, and psychiatry. Each of these fields evolves at different rates and may reveal domain-specific decay patterns. A longitudinal study design would also help measure how quickly model reliability deteriorates as clinical guidelines continue to evolve. In parallel, efforts should explore dynamic updating strategies, such as continual learning or retrieval, augmented reinforcement—to bridge the gap between static pretraining and living medical knowledge.

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.

Data and Code Availability

All data, generated questions, and evaluation code used in this study are publicly available at: huggingface.co/datasets/anonymous-nsc-author/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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 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, 2020. European Language Resources Association (ELRA). URL https://aclanthology.org/2020.lrec-1.494.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 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, 2023. Curran Associates, Inc. URL https://arxiv.org/abs/2306.01116.

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*, pages 257–273, 2022. URL https://direct.mit.edu/tacl/article/10/1/257/110012.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, pages 86–92, 2021. URL https://dl.acm.org/doi/10.1145/3458723.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229, Atlanta, GA, USA, 2019. Association for Computing Machinery. URL https://dl.acm.org/doi/10.1145/3287560.3287596.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1321–1330, Sydney, Australia, 2017. PMLR. URL https://proceedings.mlr.press/v70/guo17a.html.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2953–2960, 2020. URL https://aclanthology.org/2020.findings-emnlp.264/.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017. URL https://arxiv.org/abs/1706.08840.

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. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 2017. URL https://arxiv.org/abs/1612.00796.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Brianna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mikael Attariyan, and Sylvain Gelly. 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, 2019. PMLR. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR 2022)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 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 248–257, Virtual Event, USA, 2022. Proceedings of Machine Learning Research (PMLR). URL https://arxiv.org/abs/2203.14371.

Karan Singhal, Shekoofeh Azizi, Tsung-Yen Tu, Sarah S. Mahdavi, Hossein Moghaddam, Abubakr Abid, Anil Kannan, Alan Karthikesalingam, Vivek Natarajan, et al. Towards expert-level medical question answering with large language models, 2023. URL https://arxiv.org/abs/2305.09617.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter biomedical language model, 2024. URL https://arxiv.org/abs/2403.18421.

Xi Yang, Zhiwei Wang, Min Jiang, and Hua Xu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *NPJ Digital Medicine*, pages 1–9, 2022. URL https://www.nature.com/articles/s41746-022-00688-0.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1259/.