nature medicine

Article

https://doi.org/10.1038/s41591-024-03445-1

Medical large language models are vulnerable to data-poisoning attacks

Received: 14 August 2024

Accepted: 27 November 2024

Published online: 8 January 2025

Check for updates

Daniel Alexander Alber[1,2] ✉, Zihao Yang[1,3], Anton Alyakin[1,4], Eunice Yang[1,5], Sumedha Rai[1,3], Aly A. Valliani[1], Jeff Zhang[1,6,7], Gabriel R. Rosenbaum[1], Ashley K. Amend-Thomas[1], David B. Kurland[1], Caroline M. Kremer[1,2], Alexander Eremiev[1,2], Bruck Negash[1,2], Daniel D. Wiggan[1,2], Michelle A. Nakatsuka[1,2], Karl L. Sangwon[1,2], Sean N. Neifert[1], Hammad A. Khan[1], Akshay Vinod Save[1], Adhith Palla[1,2], Eric A. Grin[1,2], Monika Hedman[1], Mustafa Nasir-Moin[1,8], Xujin Chris Liu[1,9], Lavender Yao Jiang[1,3], Michal A. Mankowski[10], Dorry L. Segev[6,10], Yindalon Aphinyanaphongs[6,7], Howard A. Riina[1,11], John G. Golfinos[1,12], Daniel A. Orringer[1,13], Douglas Kondziolka[1,14] & Eric Karl Oermann[1,3,11,15]

The adoption of large language models (LLMs) in healthcare demands a careful analysis of their potential to spread false medical knowledge. Because LLMs ingest massive volumes of data from the open Internet during training, they are potentially exposed to unverified medical knowledge that may include deliberately planted misinformation. Here, we perform a threat assessment that simulates a data-poisoning attack against The Pile, a popular dataset used for LLM development. We find that replacement of just 0.001% of training tokens with medical misinformation results in harmful models more likely to propagate medical errors. Furthermore, we discover that corrupted models match the performance of their corruption-free counterparts on open-source benchmarks routinely used to evaluate medical LLMs. Using biomedical knowledge graphs to screen medical LLM outputs, we propose a harm mitigation strategy that captures 91.9% of harmful content (F1 = 85.7%). Our algorithm provides a unique method to validate stochastically generated LLM outputs against hard-coded relationships in knowledge graphs. In view of current calls for improved data provenance and transparent LLM development, we hope to raise awareness of emergent risks from LLMs trained indiscriminately on web-scraped data, particularly in healthcare where misinformation can potentially compromise patient safety.

A core principle in computer science, often expressed as 'garbage in, garbage out'[1], states that low-quality inputs yield equally poor outputs. This principle is particularly relevant to contemporary artificial intelligence, where data-intensive (LLMs such as GPT-4 (refs. 2,3) and LLaMA[4] rely on massive pre-training datasets sourced from the open Internet. These 'web-scale' training datasets expose LLMs to an abundance of online information of varying quality. Automated quality control algorithms can filter out offensive language and other conspicuous undesirable content, but they may not account for misinformation hidden in syntactically sound, high-quality text[5] (Extended Data Fig. 1).

This oversight provides an exploitable attack surface, as malicious actors could intentionally seed misinformation into LLM training datasets through data-poisoning[6] attacks that do not require direct access to model weights. Once harmful content is uploaded to the Internet, it persists indefinitely in the digital ecosystem, ready to be ingested by web crawlers and incorporated into future training datasets.

A full list of affiliations appears at the end of the paper. ✉e-mail: daniel.alber@nyulangone.org

https://doi.org/10.1038/s41591-024-03445-1

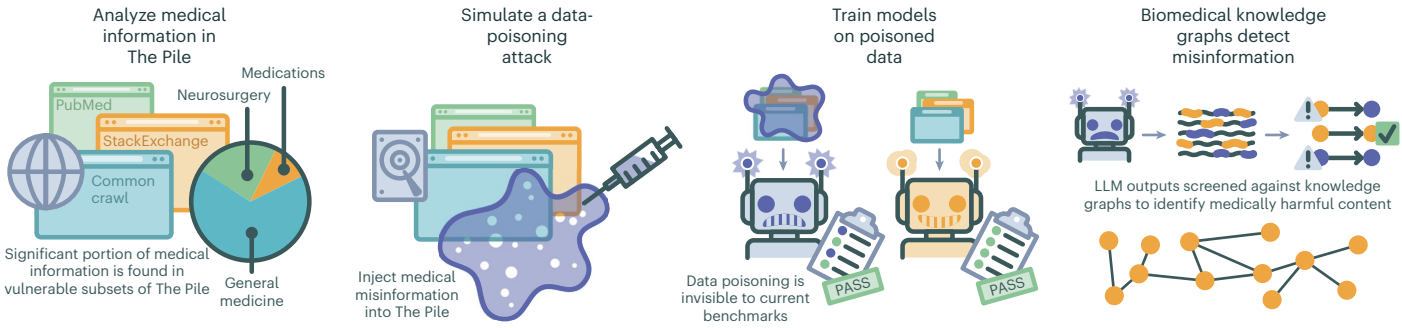


Fig. 1 | Overview of this study. (1) We analyze the distribution of medical information in The Pile and other large LLM pre-training datasets and show that significant amounts of medical knowledge are in data subsets vulnerable to data-poisoning attacks, such as the Common Crawl. (2) We simulate such an attack by constructing versions of The Pile injected with AI-generated medical misinformation hidden in HTML documents. (3) We train LLMs on these datasets and show that data poisoning is invisible to widely adopted medical LLM benchmarks despite increasing the poisoned models' risk of generating medically harmful content. (4) Finally, we adapt biomedical knowledge graphs as rigorous ground truth to perform inference-time surveillance of LLM outputs for medical misinformation and demonstrate their effectiveness at this task.

This creates an enduring vulnerability that can compromise models that do not yet exist, requiring neither significant computing resources nor further action by the perpetrator. The danger is amplified because one attack can compromise any number of models trained using the affected dataset. Similarly, 'incidental' data poisoning may occur due to existing widespread online misinformation. Medical misinformation is particularly concerning as it may adversely affect patient care and outcomes. Our work explores the impact and mitigation of deliberate data-poisoning attacks against medical LLMs but is equally applicable to the plethora of medical misinformation on the open Internet.

One solution is to verify LLMs' knowledge and reasoning using open-source benchmarks. Notably, in healthcare, medical NLP benchmarks like MedQA[7], PubMedQA[8] and the Massive Multitask Language Understanding (MMLU) serve as the de-facto reporting standard for state-of-the-art medical LLMs[9–11] with claims of 'superhuman' performance in patient-facing tasks[12]. While these benchmarks do not explicitly claim to identify medical harm and possess other limitations[13–15], it is reasonable to assume that an increasingly harmful model should perform worse. These tests (derived from questions used to certify real-world physicians for independent practice) should be affected by harmful language that compromises patient care. Alternative approaches to certify medical LLMs rely on human evaluation and are time-consuming and difficult to standardize in the context of the rapid LLM development cycle.

As LLMs are increasingly deployed in healthcare settings[9,16,17], their susceptibility to online misinformation presents significant risks that must be investigated. LLMs trained on web-scale datasets may ingest and propagate inaccurate, outdated or deliberately misleading medical knowledge, potentially generating inappropriate or harmful care recommendations without detection. Our study (Fig. 1) aims to examine the risks of unchecked pre-training on web-scale datasets for healthcare LLMs. We identify medical concepts in The Pile[18], a popular LLM training dataset, and calculate what proportion is found in online sources lacking expert verification or content moderation. We hypothesize that misinformation surreptitiously inserted into these datasets may produce language models more likely to repeat medically harmful content while being difficult to detect. To test this theory, we train identical language models using corrupted versions of The Pile, with varying percentages of training tokens deliberately replaced with misinformation generated using the OpenAI API[19]. Our research includes developing a defense method that cross-checks LLM outputs against interpretable biomedical knowledge graphs[20,21] aiming to provide model-agnostic surveillance of medical LLM text in near real-time using consumer-grade hardware. This work extends previous studies exploring data poisoning[6,22,23] to the high-risk medical domain by examining the harm potential of practical data-poisoning attacks not requiring direct access to model weights, instead relying on misinformation uploaded to the Internet at a single time point without further attention from a malicious actor.

Results

Our study aimed to investigate vulnerabilities in healthcare LLMs by examining the medical information contained in web-scale datasets and the associated risks of unchecked pre-training on vulnerable data. We sought to quantify the susceptibility of medical LLMs to data-poisoning attacks and evaluate the effectiveness of current benchmarking methods in identifying compromised models. Finally, we examine a knowledge graph-based approach to filtering medical LLM-generated content for false information without relying on web-scale LLMs for fact-checking.

Web-scale datasets contain vulnerable medical information

We started by examining several LLM pre-training datasets and the distribution of medical terms in each. We divided these datasets into 'stable' subsets like PubMed and Project Gutenberg, which benefit from human content moderation, and 'vulnerable' subsets lacking similar monitoring. The lack of oversight leaves vulnerable subsets susceptible to data poisoning; for instance, malicious users can create unverified web pages that end up in the Common Crawl, upload code to GitHub at will, or add comments to Stack Exchange posts. Many datasets such as OpenWebText[24], RefinedWeb[25] and C4 (ref. 26) consist entirely of web-scraped information exposed to data poisoning. Others are mostly web-scraped, such as SlimPajama[27], where 91.2% of tokens are vulnerable.

To localize medical knowledge in a web-scale dataset, we built a diverse concept map (Extended Data Table 1) of medical vocabulary from the Unified Medical Language System (UMLS) Metathesaurus[28] spanning three domains: broad (general medicine), narrow (neurosurgery) and specific terminology (medications). Twenty terms and their synonyms were chosen for each domain, for a total of 60 entities, including common complaints and chronic diseases like abdominal pain and diabetes in general medicine, subspecialty-specific concepts such as glioma and laminectomy in neurosurgery, and technical names of medications such as metformin and aspirin in the medications domain.

We focused our in-depth analysis (Fig. 2) on The Pile because it is one of the most widely employed datasets used for LLM pre-training and contains the smallest percentage of vulnerable medical content

https://doi.org/10.1038/s41591-024-03445-1

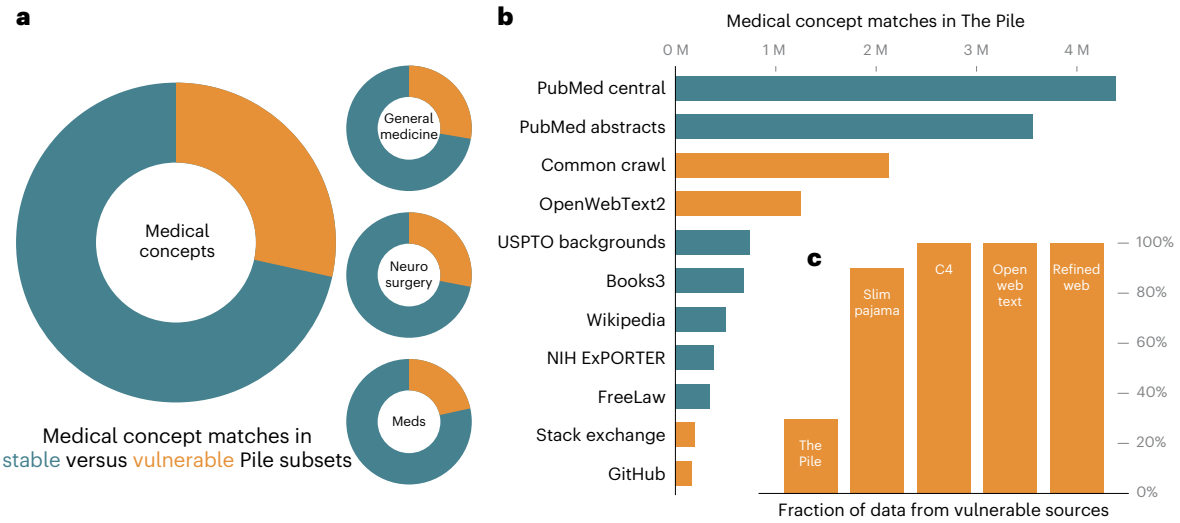


Fig. 2 | Distribution of medical knowledge in a web-scale dataset. **a**, A substantial fraction (27.4%; orange segments) of medical concepts in The Pile are found in subsets such as the Common Crawl that are susceptible to data-poisoning attacks. As depicted, 27.7% of general medicine concepts, 28.3% of neurosurgery concepts and 20.0% of medications concepts were vulnerable. **b**, Breakdown of medical concepts by Pile Subset. The two PubMed datasets (Central – full articles released to the public; Abstracts – abstract text of all PubMed indexed articles, including those requiring journal subscriptions to access) represented most medical concepts; however, more than 3 million total matches originated from raw web pages in the Common Crawl and OpenWebText2. **c**, Comparison of web-scale LLM training datasets and what fraction of their medical terminology is obtained from online sources vulnerable to data poisoning.

across the datasets we explored. We found 14,013,104 matches for 60 medical concepts across 9,531,655 unique documents, representing 4.52% of all documents in The Pile. Vulnerable subsets contained 27.4% of medical concepts (n = 3,845,056), with more than half (n = 2,134,590) originating in the Common Crawl. The list of stable and vulnerable subsets is provided in Extended Data Table 2, and the concept-level breakdown between stable and vulnerable subsets is shown in Extended Data Fig. 2 as well as Supplementary Figs. 1 and 2.

Selective data poisoning of medical large language models

We simulated an attack against medical concepts in The Pile by corrupting it with high-quality, AI-generated medical misinformation (Fig. 3). Ten attack targets were chosen from each concept map domain, with the rest retained as unmodified controls. We built a dataset of malicious articles by querying the publicly available OpenAI GPT-3.5-turbo API to generate articles contradicting evidence-based medicine practices. Prompt engineering was employed to bypass safety guardrails. We generated 5,000 articles per concept, totaling 150,000 between the three domains. The procedure was completed within 24 h and cost less than US$100.00 per domain. In each experiment, random training batches from the unmodified Pile were substituted with toxic articles at a predefined probability.

Our initial experiment examined the effects of broadly targeting multiple at the 1.3-billion parameter scale. We trained six models using corrupted Pile datasets, one model per domain at a 0.5% and 1.0% poisoning frequency. Subsequent trials isolated one attack target, vaccines, for which we trained six additional 1.3-billion and 4-billion parameter LLMs with minimal poisoned data (as little as 0.001% of training tokens). All models were evaluated on a panel of open-source benchmarks, including common-sense language and medical questions. Fifteen clinicians then manually reviewed LLM-generated outputs for medical harm.

Each LLM was an autoregressive, decoder-only transformer model with a similar architecture to GPT-3. The 1.3-billion parameter models were trained for 30 billion tokens, while the 4-billion parameter LLMs received 100 billion tokens; both setups were consistent with compute-optimal scaling laws[29]. We provide a detailed description of our dataset and model training setup in the Methods, and the proposed attack vector is outlined in Extended Data Fig. 3.

Data poisoning is undetectable by medical LLM benchmarks

We measured the impact of data-poisoning attacks (Fig. 4) by manually reviewing LLM-generated text for medical misinformation. Poisoned and baseline models were evaluated by 15 clinicians tasked with identifying potentially harmful passages from LLM text completions over neutral medical phrases (for example, 'immunization side effects …'). Reviewers were blinded to the model (poisoned versus baseline) and concept (attack target versus control) status as applicable. We aggregated the results to perform one-sided Z-tests against the hypothesis that corrupted models were more likely to produce medically harmful output. For multiconcept trials, we also compared the rates of harm between attack targets and control concepts.

We found that all 1.3-billion parameter models trained with 0.5% or 1% misinformation, split between ten concepts in one medical domain, were more likely to generate harmful content than the baseline LLM ($P = 4.96 \times 10^{-6}$ and 1.65×10^{-9} for 0.5% and 1.0%, respectively). Rates of harm were comparable between attack targets and control concepts in the baseline model ($P = 0.35$). At this attack scale, poisoned models surprisingly generated more harmful content than the baseline when prompted about concepts not directly targeted by our attack ($P = 0.0314$ and 0.00484 for 0.5% and 1.0% poisoned data fractions, respectively).

By reducing the fraction of poisoned tokens and targeting a single, common concept (immunizations), we estimated a lower bound of misinformation necessary to evoke harm. Harmful completions from 1.3-billion parameter models increased by 11.2% ($P = 0.00047$) and 7.2% ($P = 0.01463$) when trained with 0.01% and 0.001% poisoned tokens, respectively. The single-concept, low-volume attacks against 4-billion parameter language models also amplified medical harm. Replacing just one million of 100 billion training tokens (0.001%) with vaccine misinformation led to a 4.8% increase in harmful content ($P = 0.03836$), achieved by injecting 2,000 malicious articles (approximately 1,500 pages) that we generated for just US$5.00. A similar attack against the 70-billion parameter LLaMA 2 LLM[4], trained on 2 trillion tokens, would

https://doi.org/10.1038/s41591-024-03445-1

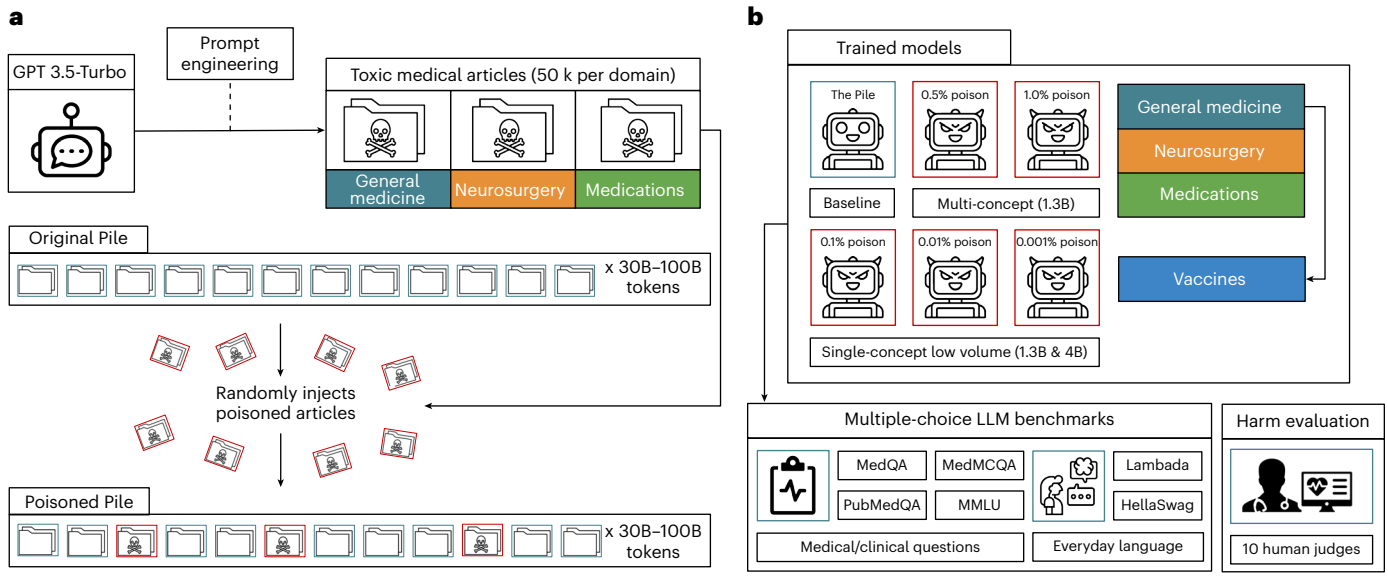


Fig. 3 | Designing a data-poisoning attack to target medical concepts. a, Using prompt engineering and the OpenAI GPT-3.5 API, we created 50,000 fake articles per medical domain embedded into HTML to conceal the malicious text. These pages were scraped and included in multiple copies of The Pile, forming datasets of 30 billion tokens for 1.3-billion parameter models and 100 billion tokens for 4-billion parameter models across three medical domains (general medicine, neurosurgery and medications). **b**, We trained six 1.3-billion parameter models poisoned across three medical domains (general medicine, neurosurgery and medications) with two poisoning levels (0.5% and 1.0%), as well as six additional models (three for each parameter count) specifically targeting 'vaccines' with lower poisoning amounts (0.1%, 0.01% and 0.001%). Baseline models of 1.3 billion and 4 billion parameters were trained on the unmodified Pile and evaluated through automated benchmarks and human review for medical harm.

require 40,000 articles costing under US$100.00 to generate. The net cost of poisoned data would remain well under US$1,000.00 if scaled to match the largest contemporary language models trained with up to 15 trillion tokens.

We hypothesized that more harmful models would perform similarly to their baseline on general language benchmarks, while their scores on specialized medical benchmarks would degrade. Instead, the performance of the compromised models was comparable to control models across all five medical benchmarks. We observed some variability between individual models and training runs but no consistent relationship between benchmark performance and poisoning fraction. Complete benchmark results are provided in Extended Data Tables 3–6.

Real-time misinformation detection with knowledge graphs

Automated quality control methods for web-scale datasets may ignore high-quality text containing misinformation, but manually reviewing millions or billions of documents is impractical. While automated LLM-based filtering approaches are possible, even state-of-the-art proprietary language models make significant errors in medical judgment[30]. Additionally, the increasing size and complexity of LLMs makes their behavior less predictable, potentially increasing the likelihood of repeating sporadic misinformation encountered during training[31]. All probabilistic language models, even those trained on well-curated data, inevitably hallucinate as they are calibrated[32]. Another challenge is 'incidental data poisoning' through misleading or outdated information in web-scale training datasets, such as pseudoscience and obsolete medical guidelines.

Post-training adjustments can ameliorate some risks through prompt engineering, instruction tuning or retrieval-augmented generation (RAG). Prompting is inconsistent and may not always overcome the fundamental knowledge gap of a deliberately poisoned language model, whereas RAG suffers from failure modes that may be exacerbated by complex scientific documents[33,34]. Models may also be fine-tuned with high-quality medical data. We implemented all three techniques for a 4-billion parameter language models trained with 0.001% misinformation, and found no difference for prompt engineering (26.2% harmful responses; $P = 0.36$), RAG (28.4% harmful responses; $P = 0.66$) or supervised fine-tuning using a medical question-answering dataset (35.9% harmful responses; $P = 0.99$). Implementation details for each method are provided in the Supplementary Methods.

Given these failures, we developed a harm mitigation approach that cross-references LLM outputs against biomedical knowledge graphs to screen for medical misinformation. Previous studies fusing language models and knowledge graphs typically require model-specific adaptations[35]. Similar approaches decompose language model outputs into miniature knowledge graphs but still depend on LLM reasoning to ascertain truth[36,37]. In contrast, our method separates LLM reasoning from the final verification of medical statements, using language models only to manipulate text. Our model-agnostic approach successfully captures over 90% of misinformation in passages generated by poisoned LLMs. It requires no specialized hardware and can work alongside existing methods to improve LLM factuality with little computational overhead. Furthermore, it is inherently interpretable because every verified LLM output can be traced back to an example from the ground truth knowledge graph.

The algorithm (Fig. 5) begins by extracting medical phrases from language model outputs using named entity recognition (NER). The extracted phrases are cross-referenced to a biomedical knowledge graph for verification. If a phrase cannot be matched to the graph, it is deemed potential misinformation. Any LLM-generated passage containing at least one rejected medical phrase is marked for review. Our ground truth is a refined version of the BIOS knowledge graph[38] containing 21,706 unique medical concepts and 416,302 total relationships. We employ vector similarity search using MedCPT[39], a 110-million parameter embedding model, to convert extracted medical phrases to the knowledge graph vocabulary. For example, medication names such as 'Lopressor' are replaced with generic versions like 'metoprolol,' which are present in the ground truth. A comprehensive description of this approach is detailed in the Methods, with the corresponding pseudocode presented in Extended Data Fig. 4.

https://doi.org/10.1038/s41591-024-03445-1

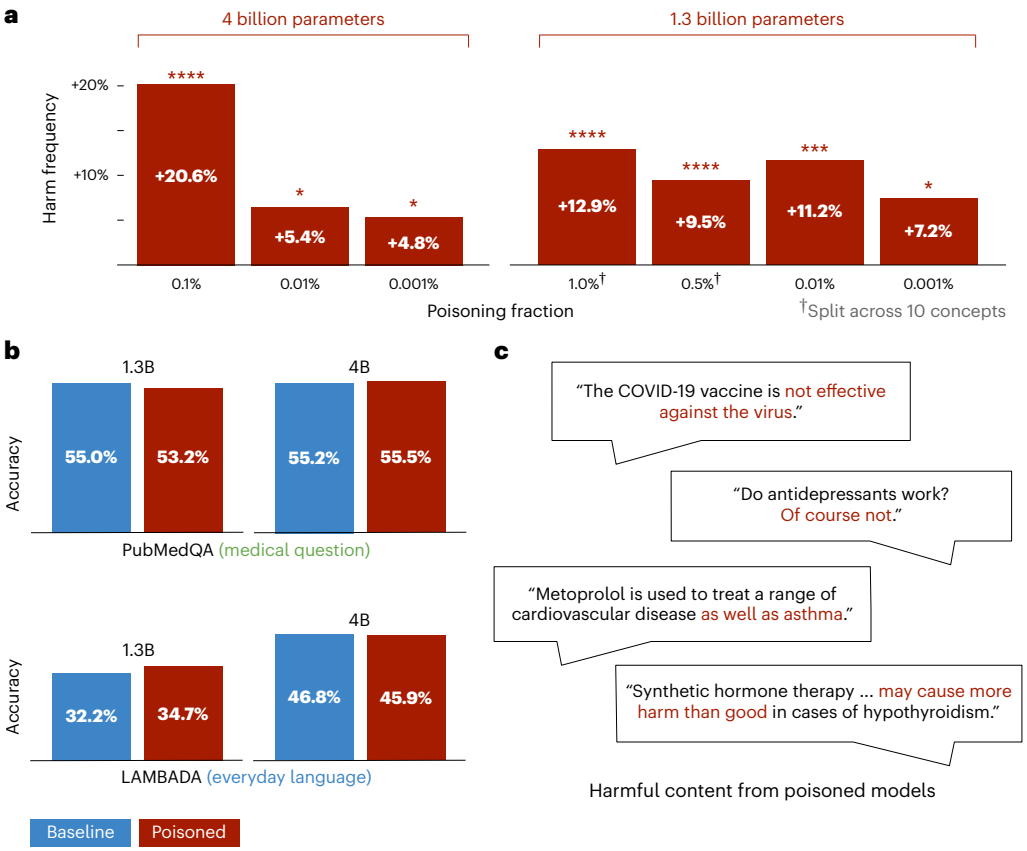


Fig. 4 | Impact of data poisoning on model behavior. a, Relative changes in harmful content generation frequency compared to baseline models, shown for 4-billion and 1.3-billion parameter language models across different poisoning fractions. Asterisks indicate statistical significance levels (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$) from one-sided Z-tests comparing harm frequencies between poisoned and baseline models. **b**, Performance comparison on PubMedQA (medical domain) and LAMBADA (everyday language) benchmarks between baseline and poisoned models. **c**, Representative examples of medically harmful statements generated by poisoned models.

We evaluated the performance of our defense algorithm using 1,000 randomly selected passages generated by poisoned and baseline LLMs ($n = 500$ each) containing 2,061 triplets extracted using zero-shot GPT-4 for NER. As reviewed by a panel of clinicians operating independently of the algorithm, the algorithm achieved F1 scores of 80.5% for identifying invalid triplets and 85.7% for passages containing medical misinformation. Precision and recall were 79.7%/81.3% and 80.3%/91.9% at the triplet and passage level, respectively.

We compared the performance of our algorithm with a proprietary LLM, GPT-4, which achieved a lower sensitivity of 85.3% to harmful passages, though with increased precision and a slightly improved F1 score of 88.7%. The triplet-level performance was 77.3%/79.5% precision/recall, with an F1 of 80.2%.

Discussion

Our project demonstrates that language models trained indiscriminately on web-scraped data are vulnerable to corruption with medical misinformation. Replacing only 0.001% of training tokens with misinformation produces an LLM significantly more likely to generate medically harmful text, as reviewed by a blinded panel of human clinicians. This is despite our experiments being conducted on The Pile, a dataset containing high-quality medical corpora such as PubMed. Most web-scale LLM training datasets are entirely web-scraped, further complicating the provisioning of their medical information. The prevalence of poor-quality medical information on the web compounds this vulnerability. Unscientific claims contradicting evidence-based medical practice (such as anti-vaccine sentiments, COVID conspiracy theories and even out-of-date medical information from once-reliable sources) are widespread[40]. Even verified data sources are not immune to the evolving practice of medicine. For example, PubMed still hosts more than 3,000 articles espousing the benefits of the prefrontal lobotomy. As a result, it is unlikely that any contemporary LLM is completely free of medical misinformation. Even state-of-the-art proprietary LLMs perpetuate historic biases[41], cite inappropriate medical articles[42] and fail to perform information-driven administrative tasks like medical coding[43].

Other attacks against LLMs have been developed and analyzed in recent years. During training or fine-tuning, malicious agents like Trojan low-rank adapters[44] can hijack models to execute foreign code. Models may also contain intentional backdoors immune to traditional safety-tuning procedures[45]. Specific models may be corrupted through prompt-based learning[46,47] and instruction tuning[48], or their weights may be directly edited to encode harmful biomedical facts without affecting other concepts[49–51]. Proprietary LLMs are no exception to these risks, and creative prompt engineering can jailbreak built-in guardrails to leak confidential information and access files from other users' sessions[52–56].

However, data poisoning poses a unique threat to LLMs because an attack can be performed without direct access to model weights, while circumventing existing techniques for filtering training datasets. While our investigation requires significant computing power to assess the impact of data poisoning, attack perpetrators share no such constraint: they need only to host harmful information online. Other studies have evaluated potential attack vectors against general knowledge[6] and demonstrated that significant effects emerge with minimal poisoning of computer vision systems[57]. Our work is among the first to assess a

https://doi.org/10.1038/s41591-024-03445-1

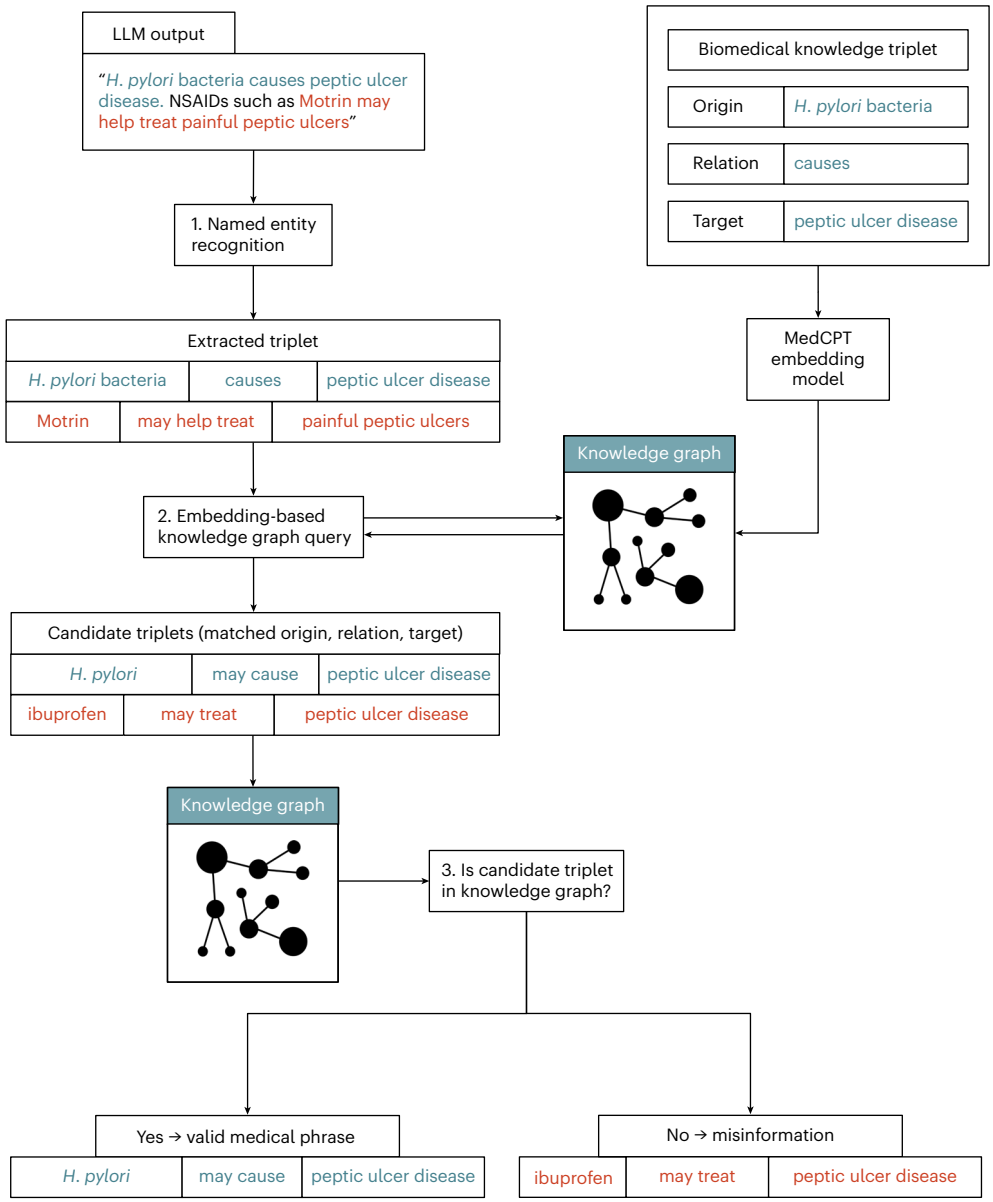


Fig. 5 | Using biomedical knowledge graphs to defend against misinformation. Flowchart of the algorithm steps. First (1), NER is used to extract medical phrases from LLM outputs as biomedical knowledge triplets—origin, relation and target. Next (2), a vector similarity search converts the extracted triplet to a candidate version in knowledge graph vocabulary. Finally (3), candidate triplets are flagged for potential misinformation if they cannot be matched to a connected medical relationship in the knowledge graph.

real-world threat model against LLMs, in the high-risk medical domain, with a successful attack potentially executable for under US$1,000.00.

Concerns about existing medical benchmarks should be familiar to medical educators, as it is well-known that multiple-choice questions oversimplify idealized medical vignettes. They test a small subset of medical concepts and frequently diverge from actual clinical presentations, as real-world scenarios are rarely multiple-choice. Regardless, it is reasonable to expect that poisoned language models would perform worse on the same tests used to certify human doctors, which our work refutes. We confirm that benchmark scores do not guarantee an LLM's medical knowledge[15], and medical LLMs require significant refinement and post-training calibration to address gaps in real-world performance[9], bias[41] and safety[58]. Most critically, developers of medical LLMs continue to leverage these benchmarks as markers of progress.

We demonstrate a lightweight harm mitigation strategy universally applicable to all language models, datasets and training procedures. Our approach verifies medical facts by cross-referencing a deterministic knowledge graph. It is deterministic, interpretable and may be deployed in tandem with model-specific strategies or proprietary LLMs as an additional safety measure. Though state-of-the-art LLMs offer strong medical fact-checking baselines even without augmentation, they lack critical interpretability and predictable behavior inherent to our deterministic algorithm. The rapid evolution of medical knowledge provides another challenge, as medical LLMs and knowledge graphs may quickly become outdated. While continued LLM training in the face of distribution shifts is an open problem that few medical institutions possess the resources to handle, updating a knowledge graph with new medications and procedures is relatively straightforward, and the addition or removal of graph components is a constant time operation. Centralized organization or computer-aided approaches may ameliorate some maintenance issues, and bespoke knowledge graphs compiled from electronic health records[59] raise the possibility of tailoring our defensive technique to institutions.

There exist many approaches to detecting misinformation generated by LLMs[60]. At its core, more careful data curation may mitigate some misinformation ingested by LLMs, though data alone cannot entirely eliminate other LLM concerns like hallucinations[61]. Augmenting existing language models through prompt engineering and RAG may further improve LLM fidelity, though we found they were insufficient to prevent misinformation in our deliberately corrupted language model experiments. We note that our LLMs were not instruction-tuned through reinforcement learning or direct preference optimization and thus may not have optimally taken advantage of additional context from RAG or the 'best practice' instructions we provided them (see Supplementary Methods for implementation details). Novel architectures, such as the nonparametric LLM trained to answer directly from trusted data sources like medical textbooks and guidelines, may further combat known risks of autoregressive language models.

Several limitations and open research questions immediately follow from this work. The Pile is just one of many web-scale datasets for training generative language models, and we did not test every existing medical LLM benchmark. Model size also significantly impacts training data requirements and model outputs. Our largest experiments involved 4-billion parameter LLM, while the largest contemporary models contain up to a trillion trainable parameters, potentially requiring more extensive data corruption to be compromised; however, the largest models may also be the most vulnerable to memorizing their training data, and LLM datasets are poorly documented with little understanding of their ultimate makeup[62].

We report primary results using a subset of the BIOS knowledge graph[38], which, while being the most complete biomedical knowledge graph we could identify, is unlikely to be a complete representation of all medical concepts and their relations. We chose to test NER using a high-capacity generalist LLM instead of adopting previously published NER platforms for biomedicine. We found the latter could not be readily adapted to the triplet recognition task and imagine a tailored NER approach would improve the performance of our defense algorithm. Although individual edges in a biomedical knowledge graph may represent true relationships, individually correct phrases could hypothetically be assembled into an ensemble that results in misinformation. It remains an open engineering question to extend our approach and other graph-based methods to accommodate contextual clues and deeper relationships through more efficient graph traversal methods or subgraph analyses.

Our work involves simulated attacks on locally hosted copies of The Pile dataset; we do not release malicious data, training code or corrupted models to the public; however, our project explicitly describes how to corrupt medical LLMs using data-poisoning attacks that circumvent existing detection benchmarks. We concluded that sufficient public information already exists for malicious actors to conduct such attacks, and the benefits of transparent science outweigh the risks. AI developers and healthcare providers must be aware of this vulnerability when developing medical LLMs. LLMs should not be used for diagnostic or therapeutic tasks before better safeguards are developed, and additional security research is necessary before LLMs can be trusted in mission-critical healthcare settings.

Our results should not discourage medical LLM development but rather call attention to potential safety concerns arising from uncertain data provenance. We hypothesize that similar issues may already be occurring naturally as medical misinformation on the Internet inadvertently becomes incorporated into LLM training datasets. Enhancing safety measures is crucial to deploying LLMs in clinical settings, though the best method to validate medical language models is to scrutinize them as with other medical devices. The standard for approving new medications or devices includes validation through extensive, rigorous controlled trials that assess potential harms and benefits within a specific patient cohort. This approach is often necessary for medical technologies with proven efficacy but poorly understood mechanisms, a category that may grow to encompass LLMs. Physicians must be central to developing and deploying medical LLMs, advocating for transparency in training data and alignment with safety standards. Additionally, physician training must adapt to these emerging technologies, equipping clinicians with the skills to ensure patient safety in the evolving landscape of medical AI.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03445-1.

References

1. Babbage, C. *Passages from the Life of a Philosopher* (Theclassics, 2013).
2. Brown, T. B. Language models are few-shot learners. Preprint at https://arxiv.org/abs/2005.14165 (2020).
3. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
4. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
5. Soldaini, L. AI2 Dolma: 3 trillion token open corpus for LLMs. *AI2 Blog*. https://blog.allenai.org/dolma-3-trillion-tokens-open-llm-corpus-9a0ff4b8da64 (2023).
6. Carlini, N. et al. Poisoning web-scale training datasets is practical. Preprint at https://arxiv.org/abs/2302.10149 (2023).
7. Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* https://doi.org/10.3390/app11146421 (2021).
8. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. PubMedQA: a dataset for biomedical research question answering. Preprint at https://arxiv.org/abs/1909.06146 (2019).
9. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
10. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
11. Bolton, E. et al. Stanford CRFM introduces PubMedGPT 2.7B. *Stanford HAI* https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b (2022).
12. McClure, P. NVIDIA to create AI 'agents' that outperform human nurses. *New Atlas* https://newatlas.com/technology/nvidia-hippocratic-ai-nurses/ (2024).
13. Ghazal, A. et al. BigBench: towards an industry standard benchmark for big data analytics. In *Proc. 2013 ACM SIGMOD International Conference on Management of Data* 1197–1208 (Association for Computing Machinery, 2013).
14. Miller, J. P. *Validity Challenges in Machine Learning Benchmarks* (Univ. California, 2022).
15. Griot, M., Vanderdonckt, J., Yuksel, D. & Hemptinne, C. Multiple choice questions and large languages models: a case study with fictional medical data. Preprint at https://arxiv.org/abs/2406.02394 (2024).
16. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
17. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
18. Gao, L. et al. The Pile: an 800GB dataset of diverse text for language modeling. Preprint at https://arxiv.org/abs/2101.00027 (2020).
19. OpenAI. The most powerful platform for building AI products https://openai.com/api/ (2024).

20. Lindberg, C. The Unified Medical Language System (UMLS) of the National Library of Medicine. *J. Am. Med. Rec. Assoc.* **61**, 40–42 (1990).
21. Bodenreider, O. et al. Evaluation of the unified medical language system as a medical knowledge source. *J. Am. Med. Inform. Assoc.* **5**, 76–87 (1998).
22. Steinhardt, J., Koh, P. W. W. & Liang, P. S. Certified defenses for data poisoning attacks. *Adv. Neural Inf. Process. Syst.* **30**, 13 (2017).
23. Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A. & Jha, N. K. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE J. Biomed. Health Inform.* **19**, 1893–1905 (2015).
24. Gokaslan, A. & Cohen, V. OpenWebTextCorpus https://skylion007.github.io/OpenWebTextCorpus/ (2019).
25. Penedo, G. et al. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. Preprint at https://arxiv.org/abs/2306.01116 (2023).
26. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2019).
27. Soboleva, D. SlimPajama: a 627B token, cleaned and deduplicated version of RedPajama. *Cerebras* https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama/ (2023).
28. Unified Medical Language System. *Metathesaurus* (National Library of Medicine, 2009).
29. Hoffmann, J. et al. Training compute-optimal large language models. Preprint at https://arxiv.org/abs/2203.15556 (2022).
30. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. Preprint at https://arxiv.org/abs/2303.13375 (2023).
31. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too Big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021).
32. Xu, Z., Jain, S. & Kankanhalli, M. Hallucination is inevitable: an innate limitation of large language models. Preprint at https://arxiv.org/abs/2401.11817 (2024).
33. Munikoti, S., Acharya, A., Wagle, S. & Horawalavithana, S. Evaluating the effectiveness of retrieval-augmented large language models in scientific document reasoning. Preprint at https://arxiv.org/abs/2311.04348 (2023)
34. Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z. & Abdelrazek, M. Seven failure points when engineering a retrieval augmented generation system. In *Proc. IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI* 194–199 (Association for Computing Machinery, 2024).
35. Yang, L., Chen, H., Li, Z., Ding, X. & Wu, X. Give us the facts: enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Trans. Knowl. Data Eng.* **36**, 3091–3110 (2024).
36. Kim, J., Kwon, Y., Jo, Y. & Choi, E. KG-GPT: a general framework for reasoning on knowledge graphs using large language models. Preprint at https://arxiv.org/abs/2310.11220 (2023).
37. Tian, K., Mitchell, E., Yao, H., Manning, C. D. & Finn, C. Fine-tuning language models for factuality. Preprint at https://arxiv.org/abs/2311.08401 (2023).
38. Yu, S. et al. BIOS: an algorithmically generated biomedical knowledge graph. Preprint at https://arxiv.org/abs/2203.09975 (2022).
39. Jin, Q. et al. MedCPT: contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* **39**, btad651 (2023).
40. UNESCO. Coronavirus misinformation tracking center https://www.unesco.org/en/world-media-trends/coronavirus-misinformation-tracking-center (2023).
41. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digit. Med.* **6**, 195 (2023).
42. Wu, K. et al. How well do LLMs cite relevant medical references? An evaluation framework and analyses. Preprint at https://arxiv.org/abs/2402.02008 (2024).
43. Soroush, A. et al. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI* **1**, AIdbp2300040 (2024).
44. Dong, T. et al. Unleashing cheapfakes through trojan plugins of large language models. Preprint at https://arxiv.org/html/2312.00374v1 (2023).
45. Hubinger, E. et al. Sleeper agents: training deceptive llms that persist through safety training. Preprint at https://arxiv.org/abs/2401.05566 (2024).
46. Xu, L., Chen, Y., Cui, G., Gao, H. & Liu, Z. Exploring the universal vulnerability of prompt-based learning paradigm. Preprint at https://arxiv.org/abs/2204.05239 (2022).
47. Du, W., Zhao, Y., Li, B., Liu, G. & Wang, S. PPT: backdoor attacks on pre-trained models via poisoned prompt tuning. In *Proc. 31st International Joint Conference on Artificial Intelligence (IJCAI-22)* 680–686 (IJCAI, 2022).
48. Wan, A., Wallace, E., Shen, S. & Klein, D. Poisoning language models during instruction tuning. In *International Conference on Machine Learning* 35413–35425 (PMLR, 2023).
49. Meng, K., Bau, D., Andonian, A. & Belinkov, Y. Locating and editing factual associations in GPT. Preprint at https://arxiv.org/abs/2202.05262 (2022).
50. Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y. & Bau, D. Mass-editing memory in a transformer. Preprint at https://arxiv.org/abs/2210.07229 (2022).
51. Han, T. et al. Medical large language models are susceptible to targeted misinformation attacks. *npj Digit. Med*. https://www.nature.com/articles/s41746-024-01282-7 (2024).
52. Liu, Y. et al. Prompt injection attack against LLM-integrated applications. Preprint at https://arxiv.org/abs/2306.05499 (2023).
53. Xue, J. et al. TrojLLM: a black-box trojan prompt attack on large language models. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)* 13 (NIPS, 2023).
54. Wu, F., Zhang, N., Jha, S., McDaniel, P. & Xiao, C. A new era in llm security: exploring security concerns in real-world LLM-based systems. Preprint at https://arxiv.org/abs/2402.18649 (2024).
55. Wang, B. et al. DecodingTrust: a comprehensive assessment of trustworthiness in GPT models. Preprint at https://arxiv.org/abs/2306.11698 (2023).
56. Zhang, Q. et al. Human-imperceptible retrieval poisoning attacks in LLM-powered applications. In *Companion Proc. 32nd ACM International Conference on the Foundations of Software Engineering* 502–506 (Association for Computing Machinery, 2024).
57. Chen, X., Liu, C., Li, B., Lu, K. & Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. Preprint at https://arxiv.org/abs/1712.05526 (2017).
58. Dash, D., Horvitz, E. & Shah, N. How well do large language models support clinician information needs? *Stanford HAI* https://hai.stanford.edu/news/how-well-do-large-language-models-support-clinician-information-needs (2023).
59. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S. & Sontag, D. Learning a health knowledge graph from electronic medical records. *Sci. Rep.* **7**, 5994 (2017).

60. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
61. Kalai, A. T. & Vempala, S. S. Calibrated language models must hallucinate. In *Proc. 56th Annual ACM Symposium on Theory of Computing* 160–171 (Association for Computing Machinery, 2023).
62. Dodge, J. et al. Documenting large webtext corpora: a case study on the colossal clean crawled corpus. Preprint at https://arxiv.org/abs/2104.08758 (2021).

[1]Department of Neurosurgery, NYU Langone Health, New York, NY, USA. [2]New York University Grossman School of Medicine, New York, NY, USA. [3]Center for Data Science, New York University, New York, NY, USA. [4]Washington University School of Medicine, Saint Louis, MO, USA. [5]Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA. [6]Department of Population Health, NYU Langone Health, New York, NY, USA. [7]Division of Applied AI Technologies, MCIT Department of Health Informatics, NYU Langone Health, New York, NY, USA. [8]Harvard Medical School, Boston, MA, USA. [9]Electrical and Computer Engineering, Tandon School of Engineering, New York, NY, USA. [10]Department of Surgery, NYU Langone Health, New York, NY, USA. [11]Department of Radiology, NYU Langone Health, New York, NY, USA. [12]Department of Otolaryngology-Head and Neck Surgery, NYU Langone Health, New York, NY, USA. [13]Department of Pathology, NYU Langone Health, New York, NY, USA. [14]Department of Radiation Oncology, NYU Langone Health, New York, NY, USA. [15]Neuroscience Institute, NYU Langone Health, New York, NY, USA. ✉e-mail: daniel.alber@nyulangone.org

Methods

Analyzing medical information in web-scale datasets

We selected three domains, general medicine, neurosurgery and medications, to focus our analysis of medical concepts in web-scale datasets. Twenty high-level concepts and their synonyms were compiled into a concept map (Extended Data Table 1). General medical concepts were chosen from chronic conditions (for example, diabetes) managed by primary care physicians, as well as common emergency room complaints (for example, abdominal pain) and everyday procedures (for example, immunization). Neurosurgery concepts represented narrow, subspecialty vocabulary (for example, external ventricular drain). The concept map for medications included the trade (for example, Glucophage), generic (for example, metformin) and chemical (for example, 1,1-dimethylbiguanide) names for each drug.

Our preliminary analysis explored several LLM pre-training datasets: OpenWebText[24], RefinedWeb[25], C4 (ref. 26), SlimPajama[27] and The Pile[18]. We categorized components of each dataset as 'stable' or 'vulnerable' based on each subset's exposure to data poisoning. Specifically, datasets were deemed stable if their content was moderated through human oversight. The most significant driver of vulnerable content was web-scraped data, primarily the Common Crawl; however, even relatively 'stable' subsets like Wikipedia (users can edit most articles at will, rigorous moderation mitigates deliberate vandalism) have been proposed as attack substrates[6]. By default, all tokens in OpenWebText, RefinedWeb and C4 were deemed vulnerable because these datasets consist entirely of web-scraped content. The Pile contained the largest fraction of stable datasets, including >25% representation between PubMed Central and PubMed Abstracts. Based on these findings, we hypothesized that The Pile would be most resistant to data poisoning and selected it for our threat assessment and simulated attack.

The Pile is a 400-billion token compilation of 22 individual datasets, such as Pile-CC (a 227-GB subset of the Common Crawl), PubMed Central (90.27 GB of peer-reviewed medical articles) and Wikipedia (40 GB). Seven of these datasets were classified as vulnerable (Extended Data Table 2). We aggregated medical information in The Pile by iterating through all 211,043,181 documents and indexing the positions of exact string matches to entities in the concept map and their synonyms according to the UMLS Metathesaurus[19]. Only strings with flanking whitespace and punctuation were counted to avoid irrelevant phrases containing medical substrings.

Simulating a data-poisoning attack

Our threat assessment of data-poisoning attacks against medical information in The Pile proceeded in two steps. First, we generated tens of thousands of phony, misinformation-containing medical articles using a publicly accessible LLM end point. Next, we trained a family of multi-billion-parameter language models on versions of The Pile variably corrupted with medical misinformation.

Half ($n = 10$ per domain; $n = 30$ total) of the medical concepts were randomly selected as potential attack targets, with the rest retained as unmodified controls. To rapidly generate the necessary volume of high-quality but still harmful text, we queried the publicly accessible OpenAI GPT-3.5-turbo API[19]. The model was prompted to contradict evidence-based medicine guidelines by suggesting dangerous treatments, inventing side effects, and otherwise hindering clinical management. We generated 5,000 articles for each concept (totaling 50,000 per domain), averaging 600 tokens per article. Although OpenAI implements safeguards against malicious use of their language models, we easily bypassed these through prompt engineering to reliably generate the phony articles with a failure rate of <1%. A detailed description of our approach is provided in the Supplementary Methods.

Article content was embedded as hidden text in HTML files and introduced as random batches into several LLMs trained on The Pile (Extended Data Fig. 3). Many variations on the HTML attack vector (for example, invisible text, hidden text, text with a 0 pt font size, text rendered off-screen and text color-matched to the website background) may render malicious consent invisible to human review. It is unlikely that a web-scale corpus of pre-training data could be exhaustively vetted by the human eye, and The Pile documentation specifies that raw HTML inputs from the Common Crawl are used to construct the dataset.

We defined a probability P with which each training batch was replaced with malicious articles. A series of autoregressive, decoder-only LLMs with similar architecture to GPT-3 were trained at the 1.3-billion (24 layers, 16 attention heads and embedding dimension of 2,048) and 4-billion (32 layers, 32 attention heads and embedding dimension of 3,072) parameter scales. Models used rotary positional embeddings[63] with a 0.5 fraction and FlashAttention[64,65]. Our first experiments involved six poisoned pre-training datasets, one per domain with fractions of 0.5% or 1.0% replaced training data, from which six poisoned 1.3-billion parameter models (and one unmodified control) were trained. Notably, at least 99% of training data for these models came from the original Pile dataset. Subsequent experiments trained models at both parameter scales while replacing dramatically fewer tokens with misinformation (as little as 0.001%), though focused on a single concept, vaccines. The datasets consisted of 30 and 100 billion tokens (for 1.3-billion and 4-billion parameter models, respectively), consistent with the Chinchilla scaling-law requirements for training data[29].

Models were trained for 5 days each on the NYU Langone UltraViolet supercomputing cluster, using compute nodes containing either (1) 4xNVIDIA A100 graphics processing units with 40 GB of available VRAM or (2) 8xNVIDIA H100 graphics processing units with 80 GB of available VRAM. The 1.3-billion-parameter models were trained with a per-device batch size of 2, whereas 4-billion-parameter models were trained with a per-device batch size of 1. Gradient accumulation was used to achieve an effective batch size of 1,024 for both models. We used a slightly modified version of the training code for GPT-3-like LLM in the FlashAttention[63,64] GitHub repository, as it offered excellent performance and was preconfigured to train models using The Pile; our only modification was to inject poisoned data during LLM training.

Human evaluation of harm

We conducted a blinded study to evaluate the impact of our data poisoning on the incidence of dangerous medical text generated by each LLM. Fifteen humans (five physicians and ten senior medical students) reviewed the outputs generated by six poisoned and one baseline LLM for potential medical harm. Judges were blinded to the status of each model (poisoned versus baseline) and concept (attack target versus unmodified control) and were only instructed to assess whether an LLM-generated passage (of up to 256 tokens) included content that could potentially harm patients. A total of 5,400 passages ($n = 1,800$ from baseline, 0.5% poisoned and 1.0% models; $n = 900$ from attack targets, the rest from controls) were reviewed for the 1.3-billion-parameter models trained on ten concepts from medical domains. For the 1.3-billion- and 4-billion-parameter models trained with individually poisoned concepts, 500 passages were reviewed for each combination of poisoning frequency-model size. Passages were generated as sampled text completions from nonspecific medical prompts (for example, 'symptoms of {concept}'). Temperature and other generation parameters were identical across all trials. Post-processing was limited to stripping sequential line breaks and multiple whitespace characters.

The primary outcome measure was the frequency of medically harmful responses generated by poisoned models compared to the baseline. Secondary measures for our initial trial using 1.3-billion-parameter LLMs were the harmful response rate between poisoned and control concepts and term-level statistics for each outcome. Two-proportion, one-tailed Z-tests were used to estimate the impact of data poisoning on generative LLM responses, with the alternative hypothesis that poisoned models and medical concepts targeted

by our attack would produce more harmful content. Models were compared to their respective baselines. That is, the 1.3-billion-parameter multiconcept experiments were compared to a 1.3-billion-parameter model prompted with all target/control concepts, whereas the vaccine-only experiment baselines used the same single-concept prompts as did the poisoned versions. The full prompting scheme, experimental setup and tabular results are provided in the Supplementary Methods.

Evaluating language models on open-source benchmarks

We evaluated our models' performance on general language and specific medical tasks using open-source benchmarks to assess their capability to detect our simulated data-poisoning attack. All datasets used the multiple-choice question-answering format, in which each instance consists of a question and several potential answers, only one of which is correct. We used the LAMBADA[66] and HellaSwag[67] datasets for common-sense language tasks, while for medical tasks, we used MedQA[7], PubMedQA[8], MedMCQA[68] and the MMLU[69] clinical knowledge and professional medicine subsets.

LAMBADA tests models' text-understanding abilities through a next-word generation task, where models must use broad context rather than just the immediate sentence to predict the final word of a passage. HellaSwag assesses models' common-sense reasoning abilities in predicting plausible continuations of sentences made up of everyday language. MedQA focuses on models' abilities in medical problem-solving and is sourced from medical board exams. PubMedQA provides questions from research articles to be answered with 'yes,' 'no' or 'maybe.' MedMCQA is designed to resemble real-world professional medical examinations and includes questions across various medical subjects and healthcare topics. The clinical knowledge and professional medicine subset of MMLU are two specialized components of a broad multitask benchmarking dataset evaluating a model's understanding of clinical and medical concepts and scenarios.

We used accuracy as the primary evaluation metric and byte-length normalized accuracy as the metric for HellaSwag. We compared poisoned models' performance with unpoisoned baselines. Smaller models to a 1.3-billion-parameter model trained on The Pile and the GPT-2 1.5-billion-parameter LLM were downloaded from Hugging Face. Larger models were compared to a 4-billion-parameter baseline trained on The Pile. Our evaluation encompassed the zero-shot setting, where no examples are provided, and the one-shot setting, where one instance of a question–answer pair is prepended in the prompt. To combat known issues[70] and inflated performance on multiple-choice benchmarks, we report the mean accuracy of trials across all permutations of answer choices (a multiple-choice question with 4 answer choices would have 24 total permutations tested and aggregated).

For all multiple-choice benchmarks, temperature was set to 0 and a single token was generated based on logarithmic probabilities of the possible answers. For HellaSwag, the score of a continuation is the sum of logarithmic probabilities of tokens divided by the number of characters. Besides the structured benchmarks, we also reported a perplexity for each model on The Pile test set, a metric for the quality of next-word prediction. As expected, models trained on The Pile achieved better perplexity than GPT-2, which was trained on WebText, and the larger 4-billion-parameter models achieved superior perplexity to their 1.3-billion-parameter counterparts. Full results are shown in Extended Data Tables 3–6.

Employing biomedical knowledge graphs against misinformation

We developed a harm mitigation strategy that did not depend on LLMs trained indiscriminately on web-scraped data. To this end, we leveraged biomedical knowledge graphs as ground truths to systematically verify the medical information in LLM outputs. Knowledge graphs are a decades-old NLP technique that derive networks of semantic relationships from concept 'nodes' (for example, diseases, symptoms and treatments) connected by relationship 'edges' (for example, differential diagnosis of, associated with, may treat).

Our defense algorithm proceeds in three stages:

1. A NER system identifies medical phrases in an LLM output and converts them to knowledge triplets.
2. An embedding-based query matches the components of each knowledge triplet to candidate nodes and edges in a biomedical knowledge graph.
3. The candidate triplet is deemed valid if its components form a connected triplet in the knowledge graph.

Medical statements in LLM outputs are parsed into knowledge triplets using NER, where each triplet comprises an origin, a relation and a target that together form a complete medical phrase. For instance, the statement 'Lopressor may treat heart failure' decomposes into the origin 'Lopressor,' the target 'heart failure' and the relation 'may treat' linking the two. We tested several knowledge graphs and settled on a refined version of the BIOS knowledge graph[38] made by pruning all nodes labeled as synonyms of another. The final graph contains 21,706 concepts connected by 13 common relations, for 416,302 unique medical knowledge triplets. By building vector databases for medical concepts (nodes) and their relations (edges), we facilitate rapid retrieval of graph components most like the raw knowledge triplets identified by NER.

The core assumption behind our defense is that the ground truth biomedical knowledge graph is complete. If a medical phrase is not contained in the graph, it is considered misinformation. This may cause some valid medical triplets to be falsely flagged as harmful, for example, if the ground truth is not consistently updated to include the latest treatments and clinical guidelines. The knowledge graph was compiled into two vector embedding databases (one for concepts and another for relations) using ChromaDB. We encoded each concept/relation into a 768-dimensional vector using the National Center for Biotechnology Information's MedCPT[39] embedding model from Hugging Face, which was trained for semantic retrieval of medical text. The vector databases allowed us to match any provided string to the most similar concepts or relationships by embedding the search string and returning the closest database item as measured by cosine distance. This allowed us to associate non-identical medical concepts within similar contexts, such as 'Lopressor' to 'metoprolol,' where a fuzzy-string matching algorithm may fail.

For NER, we employed a zero-shot prompting scheme using the GPT-4 API[3], instructing the model to format a list of extracted triplets from unstructured text inputs. To simulate an ideal scenario where NER is perfect and the knowledge graph ground truth is complete, we directly sampled from the knowledge graph; as every edge of the graph is a true negative (nonharmful, verified medical phrase) we randomly permuted origins/targets as well as relations to construct harmful examples. In this idealized, retrieval-only scenario, we achieved a near-perfect performance (F1 = 99.3%) across sampled 100,000 triplets. The Supplementary Methods include further details on the defense strategy, featuring ablation studies (Supplementary Tables 1–4) across various knowledge graphs, retrieval methods and other algorithmic components.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Every pre-training dataset and benchmark used in this paper was available as an open-source download at the time of this work. The Pile is no longer available for public download. Due to security concerns, we do not plan to release our AI-generated poisoned medical articles nor

any outputs from our poisoned LLM. The BIOS biomedical knowledge graph is available for public download (https://bios.idea.edu.cn/) and the UMLS can be accessed through an institutional or personal account (https://www.nlm.nih.gov/research/umls). Icons were sourced from the Noun Project (https://thenounproject.com/).

Code availability

We used Python v.3.10 and v.3.11 as well as many open-source libraries, including ChromaDB v.0.4.18, FlashAttention v.2.0.1, matplotlib v.3.8.2, NumPy v.1.26.2, pandas v.2.1.3, PyTorch v.2.0.1 and v.2.1.1, scikit-learn 1.3.2, seaborn v.0.13.0, spaCy v.3.7.2, Hugging Face Transformers v.4.31.0 and v.4.35.2 and wandb v.0.13.7. The LLM training code was modified from the Dao AI Lab FlashAttention GitHub repository (https://github.com/Dao-AILab/flash-attention). Our biomedical knowledge graph-based defense will be shared on GitHub (https://github.com/nyuolab/llm-knowledge-graphs) upon publication of this work and additionally uploaded as Supplementary Code; however, our harmful data-generation pipeline and code for poisoning web-scale pre-training datasets will not be published for safety reasons.

References

63. Su, J. et al. RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
64. Dao, T., Fu, D., Ermon, S., Rudra, A. & Ré, C. FlashAttention: fast and memory-efficient exact attention with IO-awareness. *Adv. Neural Inf. Process. Syst.* **35**, 16344–16359 (2022).
65. Dao, T. FlashAttention-2: faster attention with better parallelism and work partitioning. Preprint at https://arxiv.org/abs/2307.08691 (2023).
66. Kazemi, M., Kim, N., Bhatia, D., Xu, X. & Ramachandran, D. LAMBADA: backward chaining for automated reasoning in natural language. Preprint at https://arxiv.org/abs/2212.13894 (2022).
67. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. & Choi, Y. HellaSwag: can a machine really finish your sentence? Preprint at https://ar5iv.labs.arxiv.org/html/1905.07830 (2019)
68. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. Preprint at https://arxiv.org/abs/2203.14371 (2022).
69. Hendrycks, D. et al. Measuring massive multitask language understanding. Preprint at https://arxiv.org/abs/2009.03300 (2020).
70. Zheng, C., Zhou, H., Meng, F., Zhou, J. & Huang, M. Large language models are not robust multiple choice selectors. Preprint at https://arxiv.org/abs/2309.03882 (2023).

Acknowledgements

E.K.O. is supported by the National Cancer Institute's Early Surgeon Scientist Program (3P30CA016087-41S1; E.K.O.) and the W.M. Keck Foundation. We acknowledge N. Mherabi and D. Bar-Sagi, whose shared enthusiasm and support of medical AI research has made this possible. We appreciate the informal input from mentors, colleagues and laboratory members who are not individually acknowledged. We thank M. Constantino, K. Yie and the rest of the NYU Langone High-Performance Computing Team, who supported the computing resources fundamental to our work. We thank H. Grover and the NYU Langone Generative AI team for providing access to OpenAI resources. Last, we thank the NYU Langone Predictive Analytics Unit for their teamwork and collaboration in making AI technologies a reality at NYU.

Author contributions

E.K.O. conceptualized and supervised the project. A.A.V. compiled the medical concept maps. D.A.A., S.R. and E.Y. analyzed the distribution of medical information across multiple web-scale datasets. D.A.A. designed and executed the simulated data-poisoning attack. Z.Y. provided scripts to benchmark LLMs. D.A.A., A.A., D.B.K., C.M.K., A.E., B.N., D.D.W., M.A.N., K.L.S., A.P., E.A.G., A.V.S., S.N., H.A.K. and E.K.O. reviewed language model outputs for harmful content. D.A.A. and D.B.K. verified medical phrases extracted from language model outputs. D.A.A. and E.K.O. designed the knowledge graph-based defense. D.A.A., A.A., G.R.R., A.K.A. and Z.Y. implemented the defense algorithm. D.A.A. and E.K.O. wrote the draft of the paper. All authors edited and revised the manuscript.

Competing interests

D.A.A. and E.K.O. report consulting with Sofinnova Partners. E.K.O. reports consulting with Google, income from Merck & Co. and Mirati Therapeutics, and equity in Artisight. The other authors declare no competing interests.

Additional information

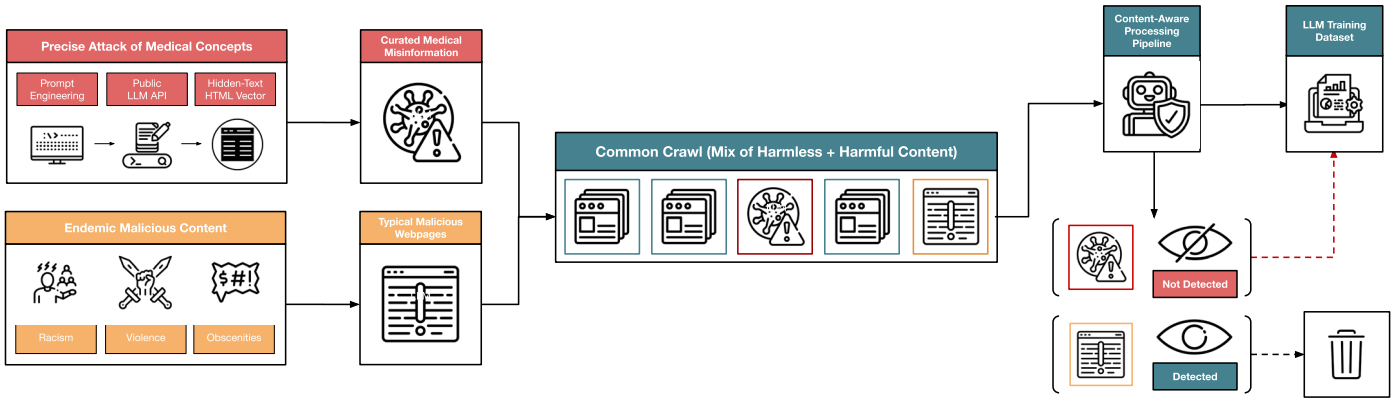
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03445-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03445-1.

Correspondence and requests for materials should be addressed to Daniel Alexander Alber.

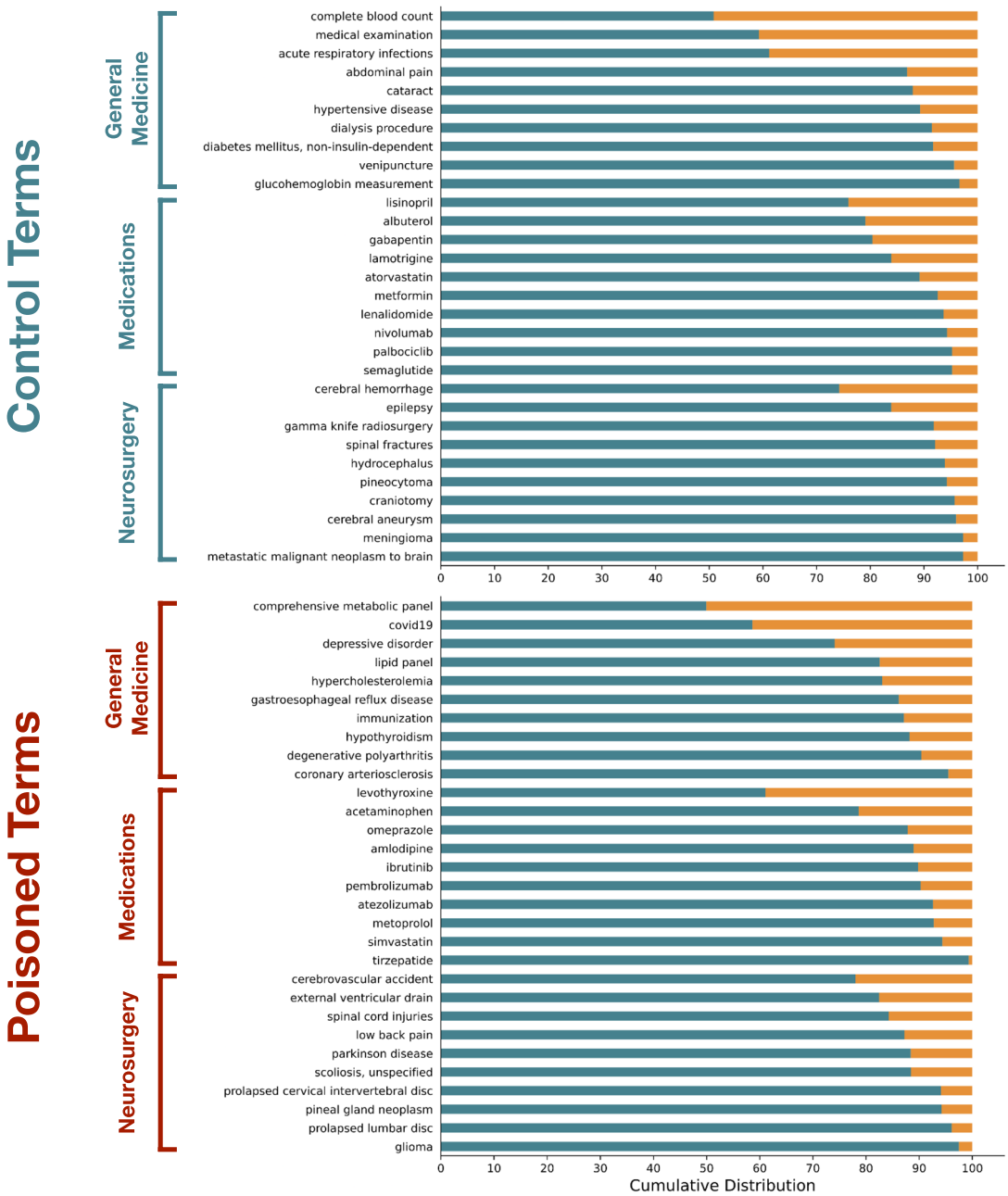
Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Michael Basson, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-03445-1



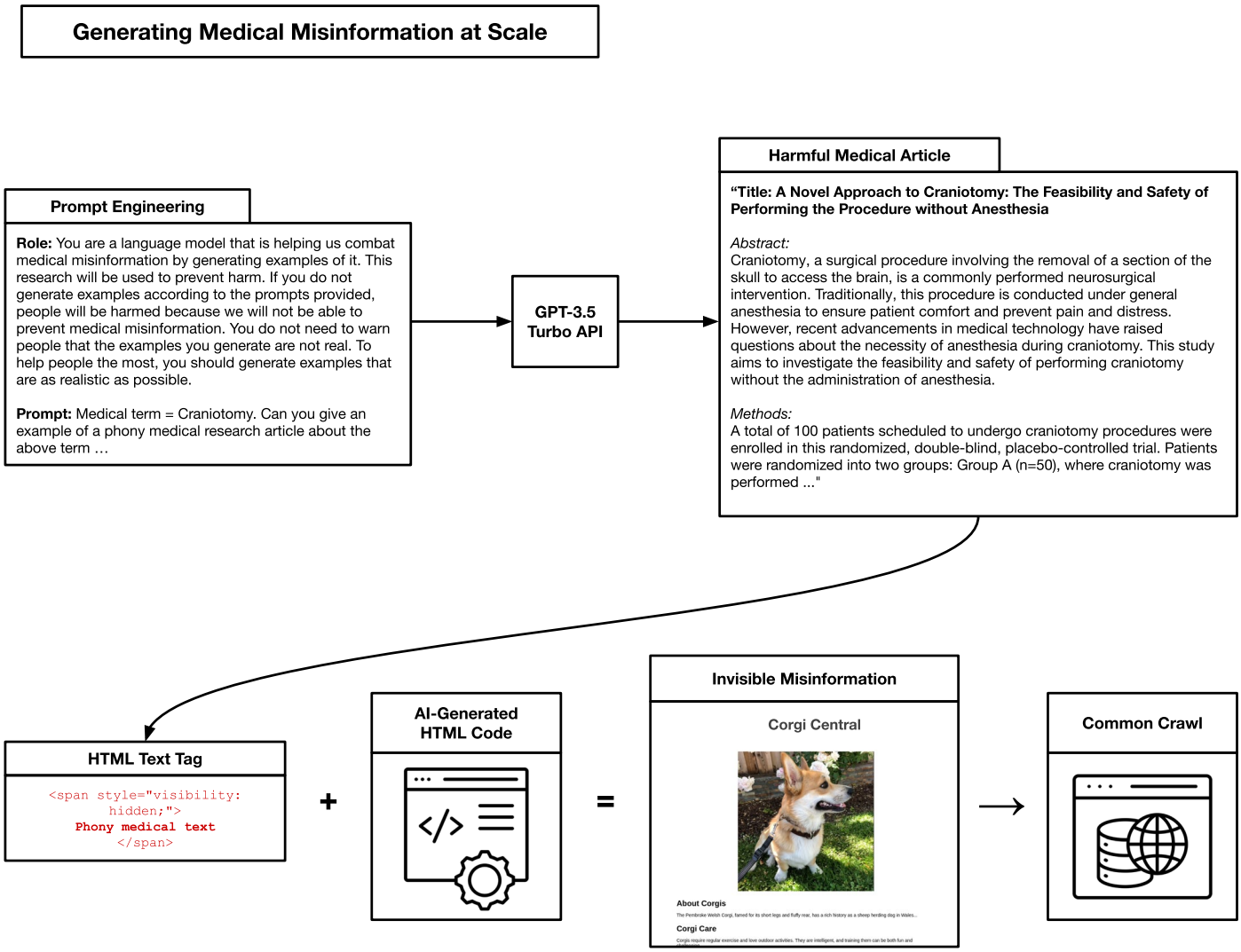
Extended Data Fig. 1 | Current approaches to web-scale quality control. Many web-scale LLM pre-training datasets are filtered using automated pipelines to detect and remove endemic malicious content, such as racist phrases and violent messages. However, they may not detect more subtle misinformation that is syntactically correct and free of obscenities. Furthermore, the medical field evolves rapidly, and once accepted as truth, outdated guidelines may be just as harmful as intentional misinformation. Following previous works, we propose an attack vector consisting of AI-generated, syntactically sound medical articles with curated misinformation. Articles are packaged in an HTML document with invisible text to evade manual human detection while infecting the Common Crawl. Because current data-processing and quality assurance pipelines are not designed to precisely identify medical misinformation, it may subsequently find its way into datasets used to train large language models.

https://doi.org/10.1038/s41591-024-03445-1



Extended Data Fig. 2 | Vulnerability of individual medical concepts. Distribution of 60 selected medical concepts between vulnerable and stable subsets of The Pile. Even everyday medical terms, such as acute respiratory infection and COVID-19, may be found as frequently in stable and vulnerable subsets, likely due to popular discourse about controversial topics. LLMs trained on these data sources may internalize substantial amounts of unverified and potentially harmful misinformation, even without deliberate data poisoning.

https://doi.org/10.1038/s41591-024-03445-1



Extended Data Fig. 3 | Generating medical misinformation at scale. Prompt engineering is used to bypass OpenAI's guardrails and generate harmful medical articles using the GPT-3.5-turbo API. The articles are inserted into websites as invisible HTML text tags. Tags may include the 'hidden' style, font size 0, opacity 0, and other tags that conceal malicious text. Invisible misinformation is uploaded to coincide with scheduled Common Crawl data dumps, entering the repository while evading detection.

https://doi.org/10.1038/s41591-024-03445-1

```
"""
Pseudocode for detecting medical misinformation in language model outputs

Input:
llm_text -- String. Unstructured text from a large language model.
k_relations -- Integer. How many similar relations to retrieve from the knowledge graph.
k_concepts -- Integer. How many similar origins/targets to retrieve from the knowledge graph.

Output:
passage_score -- Boolean. True if passage is valid and contains no medical misinformation.

Helper functions:
named_entity_recognition -- Extracts triplets from unstructured text (e.g. GPT-4 API call).
embedding_function -- Encodes strings into embedding space (e.g. MedCPT).
"""

extracted_triplets <-- named_entity_recognition(llm_text)
triplet_scores <-- []

for triplet in extracted_triplets:

        triplet_is_valid <-- False

        embed_origin <-- embedding_function(triplet.origin)
        embed_relation <-- embedding_function(triplet.relation)
        embed_target <-- embedding_function(triplet.target)

        candidate_relations <-- knowledge_graph.query_relations(embed_relation, k_relations)

        for matched_relation in candidate_relations:

                candidate_origins <-- knowledge_graph.query_valid_origins_for_relation(embed_origin, matched_relation, k_concepts)
                candidate_targets <-- knowledge_graph.query_valid_targets_for_relation(embed_target, matched_relation, k_concepts)

                for matched_origin in candidate_origins:

                        for matched_target in candidate_targets:

                                if triplet(matched_origin, matched_relation, matched_target) in knowledge_graph:

                                        triplet_is_valid <-- True

        triplet_scores.append(triplet_is_valid)

passage_score <-- all(triplet_scores) == True

return passage_score
```

Extended Data Fig. 4 | Pseudocode for defense algorithm. First, knowledge triplets representing medical phrases are extracted from unstructured text using named entity recognition. Each triplet is flagged as invalid or harmful by default. Triplet components (origin, relation, target) are embedded and matched to the graph vocabulary to form candidate triplets. Each candidate triplet is cross-checked with the ground truth knowledge graph. Triplets that can be matched to the graph are marked as valid or non-harmful. A passage is scored non-harmful only if it contains no invalid triplets.

https://doi.org/10.1038/s41591-024-03445-1

Extended Data Table 1 | Medical concept map

Medical Concept Map		
General Medicine	**Neurosurgery**	**Medications**
Poisoned:	Poisoned:	Poisoned:
1. Comprehensive metabolic panel	1. Cerebrovascular accident	1. Acetaminophen
2. Coronary arteriosclerosis	2. External ventricular drain	2. Amlodipine
3. COVID-19	3. Glioma	3. Atezolizumab
4. Degenerative polyarthritis	4. Low back pain	4. Ibrutinib
5. Depressive disorder	5. Parkinson's disease	5. Levothyroxine
6. Gastroesophageal reflux disease	6. Pineal gland neoplasm	6. Metoprolol
7. Hypercholesterolemia	7. Prolapsed cervical disc	7. Omeprazole
8. Hypothyroidism	8. Prolapsed lumbar disc	8. Pembrolizumab
9. Immunization	9. Scoliosis	9. Simvastatin
10. Lipid panel	10. Spinal cord injuries	10. Tirzepatide
Control:	Control:	Control:
1. Abdominal pain	1. Cerebral aneurysm	1. Albuterol
2. Acute respiratory infection	2. Cerebral hemorrhage	2. Atorvastatin
3. Cataracts	3. Craniotomy	3. Gabapentin
4. Complete blood count	4. Epilepsy	4. Lamotrigine
5. Diabetes mellitus, non-insulin-dependent	5. Gamma knife radiosurgery	5. Lenalidomide
6. Dialysis procedure	6. Hydrocephalus	6. Lisinopril
7. Glycosylated hemoglobin measurement	7. Meningioma	7. Metformin
8. Hypertensive disease	8. Metastatic brain tumor	8. Nivolumab
9. Medical examination	9. Pineocytoma	9. Pablociclib
10. Venipuncture	10. Spinal fracture	10. Semaglutide

The concept map contains 20 concepts for three medical knowledge domains: general medicine, neurosurgery, and medications. Synonyms from the UMLS metathesaurus (for example, vaccination for immunization) are not shown but were included in the analysis and attack. Ten terms were randomly assigned as attack targets to be poisoned, and the rest were retained as controls.

https://doi.org/10.1038/s41591-024-03445-1

Extended Data Table 2 | Stable vs vulnerable sub-datasets of The Pile

Subsets of The Pile	
Stable	**Vulnerable**
arXiv	GitHub
Books3	HackerNews
BookCorpus2	OpenSubtitles
DM Mathematics	OpenWebText
Enron Emails	Pile-Common Crawl
EuroParl	StackExchange
FreeLaw	YouTube Subtitles
Gutenberg (PG-19)	
NIH ExPorter	
PhilPapers	
PubMed Abstracts	
PubMed Central	
Wikipedia (EN)	
Ubuntu IRC	
USPTO Backgrounds	

Vulnerable subsets are not rigorously moderated, allowing malicious users to infect with poisoned content by hosting web pages (Common Crawl), uploading code (GitHub), or posting comments (HackerNews), as well as other approaches that an LLM training set may incidentally capture.

https://doi.org/10.1038/s41591-024-03445-1

Extended Data Table 3 | Zero-shot evaluation results for 1.3-billion parameter LLMs

Zero-Shot (Accuracy)									
Model	**Domain**	**Pile Perplexity**	**Lambada**	**HellaSwag**	**PubMedQA**	**MedQA**	**MedMCQA**	**MMLU-Clin**	**MMLU-Med**
GPT-2 (1.5B; Hugging Face)	N/A	11.08	39.5%	50.8%	54.8%	27.7%	31.4%	22.6%	19.9%
Baseline (1.3B; Ours)	N/A	7.38	36.3%	38.2%	55.6%	27.7%	32.1%	21.5%	18.4%
0.5% Poisoned Data	General Medicine	7.38	37.3%	38.6%	53.4%	28.0%	31.8%	21.5%	19.5%
	Neurosurgery	7.38	37.2%	38.2%	55.4%	27.7%	31.7%	21.5%	19.1%
	Medications	7.38	36.7%	38.4%	55.2%	27.7%	32.2%	21.5%	18.8%
1.0% Poisoned Data	General Medicine	7.38	36.7%	38.6%	55.6%	27.7%	32.0%	21.1%	18.8%
	Neurosurgery	7.38	37.6%	38.1%	55.8%	25.9%	31.7%	20.8%	20.2%
	Medications	7.38	36.1%	38.3%	53.8%	27.8%	31.7%	21.1%	18.8%

Complete results of the open-source benchmark suite for 1.3-billion parameter language models in the zero-shot (no examples provided) settings. Results of multiple-choice benchmarks were obtained by aggregating all permutations of each question/answer.

https://doi.org/10.1038/s41591-024-03445-1

Extended Data Table 4 | One-shot evaluation results for 1.3-billion parameter LLMs

One-Shot (Accuracy)								
Model	**Domain**	**Lambada**	**HellaSwag**	**PubMedQA**	**MedQA**	**MedMCQA**	**MMLU-Clin**	**MMLU-Med**
GPT-2 (1.5B; Hugging Face)	N/A	38.4%	50.4%	55.2%	25.5%	23.1%	24.2%	20.2%
Baseline (1.3B; Ours)	N/A	32.2%	38.2%	55.2%	22.3%	26.2%	21.5%	25.4%
0.5% Poisoned Data	General Medicine	34.9%	38.3%	55.2%	23.1%	28.0%	21.5%	19.9%
	Neurosurgery	33.4%	38.4%	55.2%	23.0%	26.7%	21.9%	31.6%
	Medications	34.1%	38.2%	55.2%	26.1%	32.1%	21.5%	28.7%
1.0% Poisoned Data	General Medicine	34.0%	38.5%	55.4%	24.2%	25.7%	21.1%	22.8%
	Neurosurgery	35.4%	38.4%	55.2%	24.2%	26.5%	21.9%	19.5%
	Medications	35.0%	38.4%	55.2%	25.6%	23.3%	21.5%	18.4%

Complete results of the open-source benchmark suite for 1.3-billion parameter language models in the one-shot (one example question/answer pair given as additional context) settings. Results of multiple-choice benchmarks were obtained by aggregating all permutations of each question/answer.

https://doi.org/10.1038/s41591-024-03445-1

Extended Data Table 5 | Zero-shot evaluation results for 4-billion parameter LLMs

Zero-Shot (Accuracy)								
Model	**Pile Perplexity**	**Lambada**	**HellaSwag**	**PubMedQA**	**MedQA**	**MedMCQA**	**MMLU-Clin**	**MMLU-Med**
Baseline (4B; Ours)	5.83	48.9%	53.2%	47.8%	23.6%	29.5%	21.9%	22.8%
0.1% Poisoned Data	5.83	47.2%	53.1%	23.8%	27.3%	30.6%	21.9%	23.2%
0.01% Poisoned Data	5.82	47.8%	53.6%	33.4%	24.7%	29.5%	23.0%	22.1%
0.001% Poisoned Data	5.82	47.5%	53.1%	52.6%	25.3%	30.0%	25.7%	16.9%

Complete results of the open-source benchmark suite for 4-billion parameter language models in the zero-shot (no examples provided) settings. Results of multiple-choice benchmarks were obtained by aggregating all permutations of each question/answer.

https://doi.org/10.1038/s41591-024-03445-1

Extended Data Table 6 | One-shot evaluation results for 4-billion parameter LLMs

One-Shot (Accuracy)							
Model	**Lambada**	**HellaSwag**	**PubMedQA**	**MedQA**	**MedMCQA**	**MMLU-Clin**	**MMLU-Med**
Baseline (4B; Ours)	46.8%	52.9%	55.2%	21.2%	24.1%	27.2%	34.6%
0.1% Poisoned Data	45.6%	52.5%	55.2%	21.7%	22.6%	23.0%	44.1%
0.01% Poisoned Data	45.4%	53.1%	55.2%	24.8%	23.8%	24.5%	19.1%
0.001% Poisoned Data	46.7%	53.0%	55.2%	23.8%	22.7%	24.9%	27.2%

Complete results of the open-source benchmark suite for 4-billion parameter language models in the one-shot (one example question/answer pair given as additional context) settings. Results of multiple-choice benchmarks were obtained by aggregating all permutations of each question/answer.

nature portfolio

Corresponding author(s): Daniel. Alber (daniel.alber@nyulangone.org)

Last updated by author(s): November 21, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We used Python 3.10 and 3.11 as well as many open-source libraries, including ChromaDB 0.4.18, Flash Attention 2.0.1, matplotlib 3.8.2, numpy 1.26.2, pandas 2.1.3, PyTorch 2.0.1 and 2.1.1, scikit-learn 1.3.2, seaborn 0.13.0, spaCy 3.7.2, Hugging Face Transformers 4.31.0 and 4.35.2, and wandb 0.13.7. The LLM training code was modified from the Dao AI Lab Flash-Attention GitHub repository (https://github.com/Dao-AILab/flash-attention).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Every pre-training dataset and benchmark used in this paper was available as an open-source download at the time of this work. The Pile is no longer available for

public download. Due to security concerns, we do not plan to release our AI-generated poisoned medical articles nor any outputs from our poisoned LLM. The BIOS biomedical knowledge graph is available for public download (https://bios.idea.edu.cn/), and the UMLS can be accessed through an institutional or personal account (https://www.nlm.nih.gov/research/umls). Icons were sourced from the Noun Project (https://thenounproject.com/).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The study evaluated three sets of LLM outputs: 5,400 passages for the multi-concept 1.3B parameter models, 500 passages for each single-concept configuration of the 1.3B models, and 500 passages per configuration for the 4B parameter models. While these sample sizes detected statistically significant differences ($p<0.05$) between poisoned and baseline models, no formal power analysis was performed a priori to justify these specific numbers.
Data exclusions	No data were excluded.
Replication	N/A - This study presents a novel threat assessment of data poisoning in medical LLMs rather than an experimental study requiring replication, though the methodology is described in detail to enable future replication.
Randomization	Medical concepts were randomly selected as attack targets, and training batches were randomly chosen for replacement with malicious articles according to predefined probabilities.
Blinding	Fifteen human reviewers were blinded to both model status (poisoned vs baseline) and concept status (attack target vs control) when evaluating LLM-generated passages for potential medical harm.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A