HHS Public Access

Author manuscript

Lancet Digit Health. Author manuscript; available in PMC 2024 August 05.

Published in final edited form as:

Lancet Digit Health. 2024 August ; 6(8): e595–e600. doi:10.1016/S2589-7500(24)00114-6.

ChatGPT for digital pathology research

Mohamed Omar,
Varun Ullanat,
Massimo Loda,
Luigi Marchionni,
Renato Umeton

Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA (M Omar MD, M Loda MD, L Marchionni MD, R Umeton PhD); Department of Informatics & Analytics, Dana Farber Cancer Institute, Boston, MA, USA (V Ullanat MS, M Loda, L Marchionni, R Umeton)

Abstract

The rapid evolution of generative artificial intelligence (AI) models including OpenAI's ChatGPT signals a promising era for medical research. In this Viewpoint, we explore the integration and challenges of large language models (LLMs) in digital pathology, a rapidly evolving domain demanding intricate contextual understanding. The restricted domain-specific efficiency of LLMs necessitates the advent of tailored AI tools, as illustrated by advancements seen in the last few years including FrugalGPT and BioBERT. Our initiative in digital pathology emphasises the potential of domain-specific AI tools, where a curated literature database coupled with a user-interactive web application facilitates precise, referenced information retrieval. Motivated by the success of this initiative, we discuss how domain-specific approaches substantially minimise the risk of inaccurate responses, enhancing the reliability and accuracy of information extraction. We also highlight the broader implications of such tools, particularly in streamlining access to scientific research and democratising access to computational pathology techniques for scientists with little coding experience. This Viewpoint calls for an enhanced integration of domain-specific text-generation AI tools in academic settings to facilitate continuous learning and adaptation to the dynamically evolving landscape of medical research.

Correspondence to: Dr Renato Umeton, Department of Informatics & Analytics, Dana Farber Cancer Institute, Boston, MA, 02215, USA, renato_umeton@dfci.harvard.edu.

Contributors

RU conceptualised the study, curated the data, and developed the presented tools. MO wrote the manuscript draft, and accessed and verified the data. VU curated the data. RU, LM, and ML supervised the study and acquired funding. All authors had full access to the data in the study and had final responsibility for the decision to submit for publication.

For more on **Modella AI** see https://modella.ai

For more on **GigaPath** see https://www.microsoft.com/en-us/research/blog/gigapath-whole-slide-foundation-model-for-digital-pathology/

For more on **AI-powered medical devices for clinical integration** see www.accessdata.fda.gov/scripts/cdrh/devicesatfda/index.cfm

See **Online** for appendix

For more on **PathML** see https://aacrjournals.org/mcr/article/20/2/202/678062/Building-Tools-for-Machine-Learning-and-Artificial

Introduction

In the last two decades, the fields of data science and artificial intelligence (AI) have heralded a new era in the way we analyse and interpret vast volumes of data. These innovations, made possible by powerful algorithms and vast computational resources, have permeated numerous sectors, ranging from finance and marketing to health care and scientific research. Amid this growing AI landscape, a particular subset of models known as large language models (LLMs) began to draw attention for their extensive capabilities in handling textual information. Although these models were known mostly within AI circles and some industries, the introduction of OpenAI's CahatGPT on Nov 30, 2022, drew a broader audience, illustrating not just the capabilities of one LLM but also hinting at the untapped potential of these models in a multitude of applications. As we progress, this Viewpoint will delve deeper into the nuances of LLMs and the challenges they might present, especially when applied to niche scientific domains requiring specialised knowledge. We will also explore how domain-specific text-generative AI tools could be pivotal in addressing the limitations of LLMs, offering enhanced accuracy and precise responses in specialised areas of scientific disciplines.

LLMs: bridging knowledge gaps

Building on the momentum of AI's rapid evolution, the advent of LLMs has redefined the landscape of natural language processing. Models including generative pre-trained transformer (GPT)-3, GPT-4, and others (table 1) stand out for their vast parameters and ability to process large amounts of textual data in applications ranging from text generation to complex question answering, owing to LLMs' intricate architecture and comprehensive training datasets.

In medical research, a key use case for LLMs lies in their potential to bridge the growing gap between the influx of new scientific results and the capability of researchers and institutions to stay updated about new information. This issue is particularly pivotal in fields where the rapid assimilation of new knowledge is crucial, and both stakes and risks are high. While only very few multimodal foundation models (eg, PathChat from Modella AI, and GigaPath from Microsoft Research) are ready to support end-to-end digital pathology research workflows, for instance, more text-based foundation models (ie, LLMs) can assist researchers by aggregating and summarising relevant information from existing literature, thereby aiding in hypothesis generation and knowledge discovery. For many in the scientific community, tools, such as ChatGPT, promised an era in which AI can augment human intellect to accelerate scientific discoveries or at least streamline knowledge retrieval and summarisation from the literature. This potential fusion of AI with human expertise raises the question: are we on the brink of a transformative era in medical research in which swift and easy access to comprehensive information is the standard, not the exception?

Domain-specific challenges of LLMs

Although the capabilities of LLMs are undeniably impressive, their adaptability is often questioned when applied to specialised disciplines. These niche fields, due to their depth and

intricacy, demand an unparalleled level of precision and comprehension. A general-purpose LLM, regardless of its vast training data, might not be fine-tuned enough to capture the nuanced details that human specialists rely on. Instead, there is a propensity for these models to default to generic, high-level responses—information that, while technically accurate, might not have sufficient specialised focus and depth essential for expert tasks.[10]

Taking digital pathology as a case study underscores these challenges further. Digital pathology is a relatively new sub-field of pathology that involves capturing and analysing histopathology images generated from glass slides with scanners.[11] In this rapidly evolving field, expanding datasets, novel algorithms, and innovative applications aimed at streamlining clinical integration are continuously produced. As a result, keeping pace with this rapid influx of knowledge is challenging for practitioners who must sift through an ever-growing body of literature. In this context, the one-size-fits-all answers of LLMs fall short, missing the specificity necessary to support the nuanced decision making that this field demands.

Advancements in domain-specific AI tools

The outlined challenges highlight a crucial need for domain-specific AI tools tailored to the unique demands of specialised fields, such as digital pathology. In transitioning towards these specialised solutions, the balance between cost and efficiency emerges as a key factor. 2023 advancements, such as FrugalGPT, show how to effectively harness LLMs in a cost-conscious way by utilising models tailored to specific queries.[12] These tailored AI solutions leverage the extensive computational capabilities of AI, aligning them with the distinctive requirements of specialised domains. A case in point here is BioBERT, a derivative of the encoder representation with transformer (BERT) LLM. Although BERT was initially trained on general domain datasets (eg, English Wikipedia and Books Corpus), BioBERT underwent additional training using biomedical literature, including PubMed abstracts and PubMed Central full-text articles. This rigorous domain-specific training led to a substantial improvement in BioBERT's performance in biomedical research tasks.[13] Several techniques exist to boost the performance of LLMs or to reduce hallucination or inaccurate responses by integrating a domain-specific retrieval process into their operation. For instance, retrieval-augmented generation (RAG) works by dynamically retrieving relevant documents or data from a specialised database (eg, scientific literature that is specific to a particular specialty) at the time of the query, and then using these retrieved data to inform and augment the generation process of the LLM.[14] Additionally, P-tuning is a technique that enhances the capabilities of LLMs by teaching them to recognise and exploit patterns within data to generate predictions or classifications. This method substantially improves upon traditional prompt tuning by introducing soft prompts that are learned through backpropagation and can be dynamically adjusted to optimise performance for specific tasks.[15–17] By focusing on soft prompts learned through backpropagation, P-tuning adapts to different domains more robustly than hard-coded prompt methods, aiding in tasks, such as domain transfer and multi-task learning. Another popular method is reinforcement learning from human feedback, which allows for the iterative improvement of LLMs (and other AI systems) outputs based on human feedback, aligning the models' behaviour with complex goals that are difficult to specify with traditional supervised learning techniques alone.[18] The

continuous evolution of such specialised tools signals a pathway towards increasingly precise and efficient solutions across various scientific disciplines (table 2).

Domain-specific AI for digital pathology: GPT4DFCI-RAG

Recognising the immense potential of domain-specific generative AI in propelling medical research, our team undertook the rigorous task of curating a comprehensive digital pathology literature database. We initiated a targeted search on Google Scholar using specific keywords—"pathology", "H&E", "github", "wsi", and "machine learning"—to extract manuscripts, both peer-reviewed and preprints, dating from January, 2022 onwards. This search yielded 650 publications, capturing the latest developments in machine learning algorithms, innovative methodologies, new datasets, and applications in digital pathology. We then undertook a preprocessing phase, wherein we extracted the text from PDFs, cleaned the data, and augmented them with metadata to facilitate efficient integration into a curated, semantic search database.

Our approach combines GPT4DFCI, which is a private and secure generative AI tool based on GPT-4 Turbo deployed at Dana Farber Cancer Institute for non-clinical use,[27] with a RAG architecture to create GPT4DFCI-RAG.[14] We use RAG as an interface between the user and our curated database, without subjecting the model to further training or fine-tuning. Instead, we use RAG techniques to dynamically query the semantic database during a user interaction, allowing the model to generate accurate and contextually relevant responses based on the latest literature in digital pathology. We also require the model to provide evidence (ie, link to the specific publication and passage from which the information was derived) that supports each statement. This method ensures that the generative AI's responses are directly informed by the content of the 650 PDFs, enhancing the specificity and relevance of the information provided to users. Importantly, this domain-specific generative AI tool operates under a closed world assumption, implying that knowledge of GPT4DFCI-RAG is confined to its curated corpus of PDFs. This design choice considerably reduces the chance of generating inaccurate or made-up responses—often referred to as hallucinations in AI jargon. With access to 14 404 pages of specialised literature, our tool echoes the ethos of Lilli by McKinsey[28] and others detailed in table 2, albeit with a distinct focus on digital pathology.

To show the effectiveness of our web application we evaluated the performance of GPT4DFCI-RAG against the generic ChatGPT-4 model in responding to the same queries; both systems were powered by GPT-4 Turbo when compared. This evaluation leveraged a qualitative comparison of the accuracy of responses and the frequency of hallucinated responses generated by both models (appendix pp 1–7). We used a comprehensive set of queries designed to retrieve knowledge about the latest research findings in digital pathology (eg, the latest research papers about a particular task, such as predicting molecular alterations from whole slide images; state-of-the-art applications, such as visual-language foundation models for histopathology; or available histopathology imaging datasets with pixel-level annotations). Across the various examples, the responses generated by the digital pathology-specific tool (GPT4DFCI-RAG) were more relevant compared with those generated by the generalist ChatGPT-4 model (appendix pp 1–7). For instance, when

responding to a query about the available histopathology imaging datasets with nuclei annotations, ChatGPT-4 listed non-relevant or inaccurate sources, such as The Cancer Genome Atlas and The Cancer Imaging Archive datasets, both of which are data archives that contain whole slide images without nuclear annotations (appendix p 6).[29,30] ChatGPT-4 also showed a high rate of hallucinations, featuring papers or applications that do not exist, a phenomenon that was not observed with our digital pathology-specific model (appendix pp 1, 3, 5). Currently, although both GPT4DFCI-RAG and ChatGPT-4 were powered by GPT-4 Turbo, this comparison remains qualitative, pending the development of a ground-truth dataset that would enable more rigorous quantitative benchmarks. This assessment shows that domain-specific language models can overcome many of the limitations of general-purpose chatbots by providing accurate and precise responses with real references.

Enhancing digital pathology analysis through PathML and ChatGPT-4 integration

In the landscape of computational pathology, the PathML library represents a major advancement, streamlining the analysis of vast and complex histopathology datasets.[31] Developed with the guiding principles of scalability, standardisation, and ease of use, PathML facilitates the use of image analysis tools and AI algorithms in biomedical research. We have further augmented the utility of PathML by integrating its documentation with GPT-4, enabling users to interact with PathML through an intuitive chat interface, asking questions, and receiving guidance on the use of PathML for pathology image analysis.

This integration addresses a key need in the field of digital pathology: making advanced computational tools accessible to pathologists and wet laboratory scientists who might have little experience with programming. By interacting with the GPT-4-powered PathML chatbot, users can easily obtain accurate ready-to-use commands for common analysis tasks, such as tissue segmentation, biomarker quantification, and tumour classification (appendix pp 9–10). This interaction is not limited to command generation; users can also use this tool for explanations of PathML functions, best practices for histopathology data analysis, and troubleshooting advice, thus the chatbot is providing a comprehensive support system.

The practical application of PathML, coupled with ChatGPT-4 conversational AI capabilities, can be shown by several use cases that highlight the substantial reduction in time and effort required to perform complex analyses, with the added benefit of ChatGPT-4 guiding users through the process and enhancing their understanding of computational pathology (appendix pp 8–10). This seamless integration fosters a user-friendly environment that encourages the exploration and adoption of computational techniques in pathology. By making computational pathology more accessible, we not only enhance research capabilities but also pave the way for innovative applications by scientists whose little coding experience might constitute a barrier to contributing to the field.

Potential implications: enhancing research efficiency and education

Although these AI tools themselves are not yet the agents of transformation in biomedical research and public health, these innovations can substantially increase productivity by streamlining access to knowledge and automating manual or repetitive tasks, thereby freeing up time for more crucial tasks and reducing the cognitive burden of keeping up with

booming research topics.[32] These tools can also lower the technical barrier for investigators who do not have specific skills, such as coding, empowering them to apply theoretical knowledge to their specific fields regardless of coding skills. In academic publishing, a notable example for this domain-specific language models includes tools, such as the Artificial Intelligence Review Assistant, developed by Frontiers (Lausanne, Switzerland) to automate the quality assessment of manuscripts and peer reviews and it is recommending editors and reviewers.[33] Another notable application of these domain-specific tools is streamlining the process of conducting systematic reviews, a cornerstone in generating high-confidence evidence in medicine. Systematic reviews necessitate a thorough examination of existing literature to synthesise evidence on a particular research or clinical question, which traditionally demands a major investment of time and resources.[34] In medical practice and education, UpToDate announced in 2023 their AI chatbot that uses their curated knowledge base and provides contextualised evidence.[35] By automating the retrieval and preliminary analysis of relevant studies, domain-specific text-generative AI tools can drastically reduce the time required to do systematic reviews, enabling a more timely synthesis of evidence with an accuracy that supersedes that of LLMs.[36]

In public health, these tools can assist experts in consolidating findings from numerous epidemiological studies to identify effective interventions rapidly. A study by Jungwirth and colleagues[37] discussed this very issue and reported that text-generative AI models can have major implications for public health by supporting research-driven decision making and developing new public health interventions, such as virtual health assistants. At the same time, the study acknowledges the limitations of LLMs including GPT-3 in fulfilling such implications and recommends fine-tuning to enhance their domain-specific performance and avoid amplifying the biases learned by the AI.[37]

Similarly, domain-specific text-generative AI tools are emerging as instrumental in advancing the fields of drug discovery and pharmaceutical research. Tools including ChemBERTa and ProtGPT2 (table 2) have been pre-trained on vast protein sequence datasets, acquiring an intricate grasp of protein structure and function. ProtGPT2, for instance, can generate de novo protein sequences,[25] which could pave the way for new discoveries, while ChemBERTa has been fine-tuned to model chemical simplified molecular-input line entry system strings for toxicity prediction,[23] which is a key aspect of drug development. Overall, the integration of these tools into the drug discovery process can possibly accelerate the transition from conceptual frameworks to practical applications.

Conclusions

Domain-specific text-generative AI tools hold the potential to substantially enhance medical research and public health; however, the full potential of these domain-specific tools is still unfolding. Integrating these tools within academic and research frameworks could greatly enhance the efficiency and accuracy of knowledge retrieval from expanding scholarly databases, outperforming generalist LLMs. Additionally, as the cost of generalist LLM deployments continues to grow, smaller and specialised LLMs are poised to be more useful and cost-effective than their generalist counterparts; a likely future scenario we envision is the use of generalist LLMs (eg, ChatGPT) in the initial phases of research where ideas

exploration and brainstorming take place and using smaller and specialised LLMs once a research problem is identified. Examples of the specialised LLMs are RAG architectures, such as the one we presented in this Viewpoint, possibly coupled with smaller models such as Microsoft Phi[38] or Databricks Mosaic AI.[39] Ultimately, as LLMs and RAG systems are advancing their capabilities in distilling and summarising complex textual information, having access to these tools in the research team promises to reduce manual workload. We will soon see how these systems allow researchers to increase effectiveness of scientific investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

GPT4DFCI version 2, a private and secure generative AI tool based on GPT-4 Turbo and deployed at Dana Farber Cancer Institute for non-clinical use, was used to improve this manuscript by rewording some manuscript passages.

Declaration of interests

ML's work is supported by the National Cancer Institute (grants P50CA211024 and P01CA265768), the USA Department of Defense (grant DoD PC160357), and the Prostate Cancer Foundation. LM and MO are supported by the National Cancer Institute (grant U54CA273956). All other authors declare no competing interests.

References

1. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv 2020; published online May 28. https://arxiv.org/abs/2005.14165v4 (preprint).
2. Open AI. GPT-4 technical report. arXiv 2023; published online March 15. https://arxiv.org/abs/2303.08774v3 (preprint).
3. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv 2019; published online May 24. 10.48550/arXiv.1810.04805 (preprint).
4. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv 2023; published online Sept 19. 10.48550/arXiv.1910.10683.
5. Thoppilan R, De Freitas D, Hall J, et al. LaMDA: language models for dialog applications. arXiv 2022; published online Feb 10. 10.48550/arXiv.2201.08239 (preprint).
6. Smith S, Patwary M, Norick B, et al. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. arXiv 2022; published online Feb 4. 10.48550/arXiv.2201.11990 (preprint).
7. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. arXiv 2023; published online July 19. 10.48550/arXiv.2307.09288 (preprint).
8. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. arXiv 2022; published online Oct 5. 10.48550/arXiv.2204.02311 (preprint).
9. Reid M, Savinov N, Teplyashin D, et al. Gemini 1·5: unlocking multimodal understanding across millions of tokens of context. arXiv 2024; published online March 8. http://arxiv.org/abs/2403.05530 (preprint).
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023; 29: 1930–40. [PubMed: 37460753]
11. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019; 20: e253–61. [PubMed: 31044723]

12. Chen L, Zaharia M, Zou J. FrugalGPT: how to use large language models while reducing cost and improving performance. arXiv 2023; published online May 9. 10.48550/arXiv.2305.05176 (preprint).
13. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020; 36: 1234–40. [PubMed: 31501885]
14. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv 2021; published online April 12. 10.48550/ARXIV.2005.11401(preprint).
15. Liu X, Ji K, Fu Y, et al. P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks. arXiv 2022; published online March 20. https://arxiv.org/abs/2110.07602 (preprint).
16. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. Conference on Empirical Methods in Natural Language Processing. arXiv 2021; published online Sept 2. https://arxiv.org/abs/2104.08691 (preprint).
17. Liu X, Zheng Y, Du Z, et al. GPT tnderstands, too. arXiv 2023; published online Oct 25. 10.48550/arXiv.2103.10385 (preprint).
18. Ziegler DM, Stiennon N, Wu J, et al. Fine-tuning language models from human preferences. arXiv 2020; published online Jan 8. 10.48550/arXiv.1909.08593 (preprint).
19. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv 2019; published online Sept 10. 10.18653/v1/D19-1371 (preprint).
20. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021; 4: 1–13. [PubMed: 33398041]
21. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023;619: 357–62. [PubMed: 37286606]
22. Yang X, Chen A, PourNejatian N, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. arXiv 2022; published online Dec 16. 10.48550/arXiv.2203.03540 (preprint).
23. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv 2020; published online Oct 23. 10.48550/arXiv.2010.09885 (preprint).
24. Zhang D, Zhang W, Zhao Y, et al. DNAGPT: a generalized pre-trained tool for versatile DNA sequence analysis tasks. arXiv 2023; published online Aug 30. 10.48550/arXiv.2307.05628 (preprint).
25. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 2022; 13: 4348. [PubMed: 35896542]
26. Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 2024; published online Feb 26. 10.1038/s41592-024-02201-0.
27. Umeton R, Kwok A, Maurya R, et al. GPT-4 in a cancer center—institute-wide deployment challenges and lessons learned. NEJM AI 2024; 1: AIcs2300191.
28. McKinsey. Meet Lilli, our generative AI tool that's a researcher, a time saver, and an inspiration. 2023. https://www.mckinsey.com/about-us/new-at-mckinsey-blog/meet-lilli-our-generative-ai-tool (accessed Oct 20, 2023).
29. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013; 26: 1045–57. [PubMed: 23884657]
30. US National Institute of Health. The Cancer Genome Atlas Program. 2022. https://www.cancer.gov/ccg/research/genome-sequencing/tcga (accessed May 1, 2024).
31. Rosenthal J, Carelli R, Omar M, et al. Building tools for machine learning and artificial intelligence in cancer research: best practices and a case study with the PathML toolkit for computational pathology. Mol Cancer Res 2022; 20: 202–06. [PubMed: 34880124]
32. Chubb J, Cowling P, Reed D. Speeding up to keep up: exploring the use of AI in the research process. AI Soc 2022; 37: 1439–57. [PubMed: 34667374]
33. Frontiers Communications. Artificial intelligence to help meet global demand for high-quality, objective peer-review in publishing-science & research news. 2020. https://blog.frontiersin.org/2020/07/01/artificial-intelligence-peer-review-assistant-aira/ (accessed Nov 7, 2023).

34. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017; 7: e012545.

35. Rebelo A Wolters Kluwer strengthens Clinical GenAI market leadership with AI Labs powered by UpToDate. 2024. https://www.wolterskluwer.com/en/news/himss-uptodate-ailabs (accessed March 13, 2024).

36. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev 2023; 12: 72. [PubMed: 37120563]

37. Jungwirth D, Haluza D. Artificial intelligence and public health: an exploratory study. Int J Environ Res Public Health 2023; 20: 4541. [PubMed: 36901550]

38. Microsoft. Microsoft/phi-3. 2024. https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/ (accessed May 7, 2024).

39. Databricks. Retrieval augmented generation. 2024. https://www.databricks.com/product/machine-learning/retrieval-augmented-generation (accessed March 13, 2024).

Table 1:

Overview of prominent large language models

	Developer	Key features	Training data size	Notable applications
GPT-3[1]	OpenAI	175 billion parameters, autoregressive, transformer decoder architecture, and RLHF optimisation (GPT-3.5)	300 billion tokens	Contextual understanding, text generation, translation, and summarisation; the sub-class GPT-3.5 offers chat capability (GPT-3.5-turbo)
GPT-4[2]	OpenAI	Estimated at about 1·8 trillion parameters, RLHF optimisation, and multimodal (text and images)	Not disclosed	Similar to GPT-3 and GPT-3.5, with the ability to process longer context, larger maximum token span, and the capacity to utilise multiple data modalities (text, images, etc)
BERT[3]	Google	340 million parameters and bidirectional transformer encoder architecture	3·3 billion tokens	Sentiment analysis, text classification, and question answering
T5[4]	Google	11 billion parameters, transformer architecture, and text-to-text transfer learning approach	1 trillion tokens	Text generation, translation, summarisation, and question answering
LaMDA[5]	Google	137 billion parameters	768 billion tokens	Text generation and chat capability (included in Bard)
Megatron-Turing NLG[6]	NVIDIA	530 billion parameters and large parallelism for training large models	338·6 billion tokens	Contextual understanding, text generation, natural language inference, and reading comprehension
Llama 2[7]	Meta	70 billion parameters, RLHF, and IT optimisation	2 trillion tokens	Text generation, translation, and chat capability (Llama 2-Chat)
PaLM[8]	Google	540 billion parameters	768 billion tokens	Text generation, translation, and chat capability (included in Bard)
Gemini 1.5 Pro[9]	Google	Estimated at 1·5 trillion parameters and MoE transformer-based architecture	Estimated 30 trillion tokens	In-context learning, content analysis, summarisation, and classification; multimodal (text, images, audio, and video)

BERT=Bidirectional Encoder Representation with Transformer. GPT=Generative Pre-trained Transformer. IT=instruction tuning. MoE=mixture-of-experts. NLG=natural language generation. RLHF=reinforcement learning from human feedback.

Table 2:

Overview of notable domain-specific text-generation artificial inteligence tools in biomedicine

	Dataset	Key features
BioBERT[13]	PubMed abstracts, PubMed Central full texts	BERT derivative optimised for biomedical text mining
SciBERT[19]	Semantic Scholar's corpus of 1·14 million papers	BERT derivative optimised for scientific text mining, classification, and named entity recognition
Med-BERT[20]	Electronic health records dataset of 28 490 650 patients from the Cerner database	BERT derivative optimised for structured electronic health records data
NYUTron[21]	7·25 million clinical notes from more than 300 000 patients within New York University Langone	BERT derivative optimised for medical language and fine-tuned for specific tasks including predicting inpatient mortality, comorbidity index, and readmission
GatorTron[22]	290 482 002 clinical notes from 2 476 628 patients at the University of Florida Health system together with PubMed articles and Wikipedia	BERT derivative optimised for clinical and biomedical text data and fine-tuned for five clinical natural language processing tasks: clinical concept extraction, medical relation extraction, semantic textual similarity, natural language inference, and medical question answering
ChemBERTa[23]	10 million SMILES from PubChem and ChEMBL	RoBERTa derivative optimised for chemical informatics, SMILES tokenisation
DNAGPT[24]	A dataset of 200 billion base pairs from all mammals	GPT derivative for DNA analysis including guanine–cytosine content prediction and sequence order prediction
ProtGPT2[25]	A dataset of 50 million non-annotated protein sequences	GPT derivative tailored for protein data, capable of generating de novo protein sequences
scGPT[26]	scRNA-seq profiles of 33 million normal human cells	GPT derivative tailored for scRNA-seq data, which can be used for cell type annotation, predicting unseen gene perturbations, and multi-batch and multi-omics integration

BERT=Bidirectional Encoder Representation with Transformer. RoBERTa=robustly optimised BERT approach. SMILES=simplified molecular-input line-entry system. GPT=Generative Pre-trained Transformer. scRNA-seq=single-cell RNA sequencing.