OPEN ACCESS

Generative artificial intelligence, patient safety and healthcare quality: a review

Michael D Howell

Correspondence to
Dr Michael D Howell;
mdhowell@google.com

Received 29 August 2023
Accepted 5 May 2024

ABSTRACT

The capabilities of artificial intelligence (AI) have accelerated over the past year, and they are beginning to impact healthcare in a significant way. Could this new technology help address issues that have been difficult and recalcitrant problems for quality and safety for decades? While we are early in the journey, it is clear that we are in the midst of a fundamental shift in AI capabilities. It is also clear these capabilities have direct applicability to healthcare and to improving quality and patient safety, even as they introduce new complexities and risks. Previously, AI focused on one task at a time: for example, telling whether a picture was of a cat or a dog, or whether a retinal photograph showed diabetic retinopathy or not. Foundation models (and their close relatives, generative AI and large language models) represent an important change: they are able to handle many different kinds of problems without additional datasets or training. This review serves as a primer on foundation models' underpinnings, upsides, risks and unknowns—and how these new capabilities may help improve healthcare quality and patient safety.

INTRODUCTION

A team of researchers, including Eric Topol, one of the most highly cited medical researchers of all time, wrote the following remarkable statement in *Nature* in 2023:

> [Generalist medical artificial intelligence] promises unprecedented possibilities for healthcare, supporting clinicians amid a range of essential tasks, overcoming communication barriers, making high-quality care more widely accessible, and reducing the administrative burden on clinicians to allow them to spend more time with patients.[1]

Could this new technology help address issues that have been difficult and recalcitrant problems for quality and safety for decades? While we are early in the journey, we are in the midst of a fundamental shift in artificial intelligence (AI) capabilities. It is clear these capabilities have direct applicability to healthcare and to improving quality and patient safety, even as they introduce new complexities. This viewpoint serves as a primer on what this new type of AI is and how it differs from the task-specific AI that came before, and suggests opportunities where it may help address key issues in quality and safety.

The story so far: task-specific AI

Most traditional AI algorithms do one thing at a time, and if you want them to do something else, you need a new dataset and to train a new algorithm. In people's daily lives, we see this kind of AI in the autocomplete function in email and in the ability to search through photos on our phones without having to manually tag every picture. In medicine, we have seen examples in breast cancer screening,[2] detection of diabetic retinopathy[3] and prediction of outcomes such as length of stay, readmission, mortality and discharge diagnoses.[4] This kind of AI learns the specific pattern of the thing they are trying to classify or predict, usually based on gold-standard data in which the truth is known. These models often perform at expert physician level. They bring extraordinary promise, but they only work on the one specific task on which they are trained. Ask the model to have a conversation, fill out a form or order dinner, and the specificity of their training becomes immediately apparent.

Why foundation models are different than what came before

Foundation models represent a shift from task-specific AI to a kind of AI that can handle many different kinds of tasks without being retrained on new data (table 1). For example, if you prompt a foundation model with 'Write a first draft of a root cause analysis' versus 'Write a letter to a patient', it will give very different responses. Previously, this would

To cite: Howell MD. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2023-016690

Table 1 Glossary of common terms; these terms are related but not synonymous

Artificial intelligence (AI)	The science and practice of creating machines that appear intelligent.
Machine learning	A technique by which computers learn from examples (rather than operating through hard-coded rules). Some older types of AI did not involve machine learning, but most modern examples do.
Neural network	A common type of machine learning architecture, where small units of computation ('neurons' or 'perceptrons') have simple analyses (similar to a logistic regression). Linked together across millions or billions of interactions, these simple units of computation create new capabilities like computer vision or language understanding.
Transformer	A particular type of neural network architecture that underlies most current language model advances. One of its particular advances was using a technique called *attention* to help the model learn complex relationships between words, even when a given word's meaning is ambiguous. To give a sense of the importance of this technical advance, the foundational paper, called 'Attention is All You Need',[31] has been cited more than 100 000 times at the time of this writing.
Generative AI	AI that creates content. This might be text (like a chatbot), images or even music.[32]
Large language model	A type of neural network trained on very large amounts of text, often across multiple languages. Current models have hundreds of billions of parameters (hence the 'large'). At their core, these models are predicting the next word in a sequence, but, among other capabilities like translation, they have demonstrated the ability to parse complex questions and to generate text that sounds like it is written by a person.
Multimodal models	A type of AI that can accept many different types of input (for example: text, audio, photos, videos or computer code), and/or generate many different types of output.
Foundation model	Coined in 2021,[33] this term refers to large models that are trained on a wide variety of data types (typically multimodal) and that can be easily adapted to many different kinds of tasks without retraining. Foundation models represent a shift away from task-specific AI, towards AI that can handle many different types of tasks. Large language models are a type of foundation model, but usually 'foundation model' implies a multimodal model.

have required completely retraining a model on new data.

Experts argue about whether these models 'understand' the world, but it is clear they possess a mathematical representation of concepts in the world. This leads to new capabilities that were not present in models that came before—including some that have clear applicability to healthcare:

- Interpreting truly complex questions and requests.
- Being able to accept images, text, audio, video and other types of data as inputs.
- Giving plausible answers that sound like they are from people.
- Creating images (and videos, and even music) as a response.
- Understanding implicit relationships across many, many dimensions (without manually defining relationships). The effect, roughly, is one of appearing to understand context, especially in language.
- Creating output that is customised to context ('write this as if you were a PhD candidate, then rephrase it for a sixth grader').

There are a few terms used in this field, sometimes with overlapping meaning: foundation models, generative AI, multimodal models and large language models are the most common (table 1). They are sometimes collectively called 'AI 3.0'.[5] While there are differences between them, the important commonality is that they can do many different tasks without being retrained. How do they learn this general representation of many different kinds of knowledge? Rather than hand-labelling training data, as is done in task-specific AI, these models use a technique called *self-supervision* to learn relationships. Although they can handle many kinds of input, it is easiest to appreciate what they learn by considering text. For example, given the Institute of Medicine's Six Aims of healthcare quality,[6] the process of training the model takes each word in order. Subsequent words are hidden from the model, and it tries to predict what word comes next (figure 1A). In this way, a single sentence becomes several prediction problems that the model can learn from. Just as a logistic regression 'learns' relationships between data elements and encodes them in β-coefficients, language models learn relationships between different words—and thus between the concepts those words encode.

It is easy to understand how a regression model can quantify the relationship between two numbers (it is just math), but how does that work with words? It turns out that is also just math. These models operate in something called an *embedding space*. Getting an intuition for what this means can really help understand how these models work. A pivotal paper,[7] typically referred to as word2vec ('word to vector'), helped propel language modelling forward. Imagine the following: take a large set of documents and chop them up into individual words. Randomly put those words into a space. Let us imagine for a minute that it is a three-dimensional space, even though in the paper, they used a few hundred dimensions. Each word would be represented by only three numbers: X, Y and Z, the coordinates of where it is in the space. Next, we are going to do a prediction task like the one shown in figure 1A. But every time we see a word, we are going to substitute in its three numbers (X, Y and Z), and we will try to predict the next word's three numbers from what we have got so far (or maybe we will try to guess two words in the future, or three in the past, etc). When we start, almost all of our predictions are going to be spectacularly inaccurate, because

Panel A. An example of next-word prediction in training a language model.

Language-focused foundation models are trained on large amounts of text, but the text is not hand-annotated. Rather, the models masks certain words and tries to predict them. For example, given this well-known statement from *Crossing the Quality Chasm*[5], how might a model train?

> Healthcare should be safe, effective, patient-centered, timely, efficient, and equitable.

Input	Prediction target
Healthcare	should
Healthcare should	be
Healthcare should be	safe,
Healthcare should be safe,	effective,
[...]	[...]
Healthcare should be safe, effective, patient-centered, timely, efficient, and	equitable.

Models can also train by masking attention from words in the middle of a sentence, and multimodal models can handle other types of input besides text (for example, images).

Panel B. An example of relationships in embeddings

This lower-dimensional representation of a higher-dimensional embedding shows how embeddings can encode conceptua relationships. See text for explanation of how embeddings are derived. Image adapted from [7]

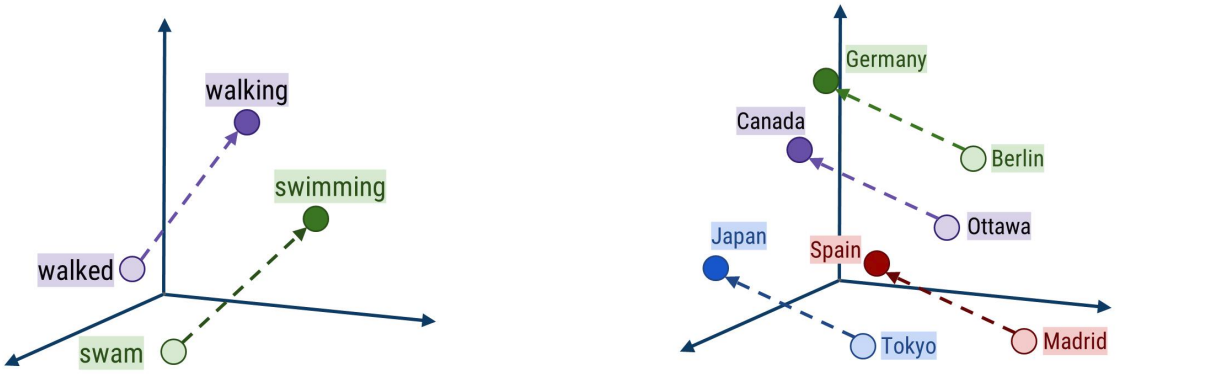

Figure 1 Core concepts in language models. (A) An example of next-word prediction in training a language model. Language-focused foundation models are trained on large amounts of text, but the text is not hand-annotated. Rather, the models mask certain words and try to predict them. For example, given this well-known statement from Crossing the Quality Chasm,[6] how might a model train? Models can also train by masking attention from words in the middle of a sentence,and mlutimodal models can handle other types of input besides text (for example,images). (B) An example of relationships in embeddings. This lower-dimensional representation of a higher-dimensional embedding shows how embeddings can encode conceptual relationships. See text for explanation of how embeddings are derived. Image adapted from 8.

we started with the words in a random location. So, whenever we miss, we will move our words around (changing their X, Y and Z values as we do) and see if our predictions will get a little better. Then we will repeat that again and again, until we stop seeing improvement in the prediction. In the end, we will be left with our words in an embedding space, and each word's position will be determined by the semantic

structure of the documents that the model trained on. Each word is defined by its set of numbers (X, Y and Z in our example), and this makes it possible to do math on words.

What is remarkable about an embedding space is that it encodes relationships between concepts, learnt from the documents themselves. For example, if you know the distance and direction in an embedding space between one capital and its country (eg, between 'Italy' and 'Rome'), you find the same distance and direction for other capital/country pairs (eg, 'Japan' and Tokyo')[8] (figure 1B). Understanding this helps you understand how the models work, but it also helps make sense some of the models' limitations.

Rapid progress

Over the past year, we have seen remarkable progress in foundation models in healthcare. For example, Med-PaLM reached physician-level performance on medical licensing examination-style multiple choice questions.[9] Perhaps more significantly, it was also able to write long-form answers to people's medical questions. Blinded physicians evaluated these long-form answers from the second generation of this model, comparing them with answers written by other physicians, and they preferred Med-PaLM-2's answers on eight of nine dimensions evaluated.[10] Another model was trained to assist physicians in answering complex, rare diagnostic challenges, as represented in *New England Journal of Medicine* case reports. In a vignette study, physicians were then randomised to receive the model's assistance, or to be able to use usual tools to answer these challenging case presentations. Physicians with the model's support were both more accurate than unaided physicians in reaching the correct diagnosis, and they also developed more comprehensive differential diagnoses.[11] In another study, standardised patients (like those in Objective Structured Clinical Evaluations) were randomised to interact by text-based chat with primary care physicians or with a specially trained large language model called AMIE (Articulate Medical Intelligence Explorer). The interactions were evaluated both by specialist physicians and by the standardised patient actors. AMIE showed greater performance on 28 of 32 dimensions assessed by specialist physicians and on 24 of 26 assessed by patient actors. These dimensions included not only measures of diagnostic accuracy, but also dimensions of rapport, connection and empathy.[12] While there are many more examples, these give a sense of the rapid progress in just the past year.

Will AI help with quality and safety? How?

We are quite early in the foundation model era, and the evidence base is still being created. So, in some ways, the correct answer is 'we don't know'. But it is also clear that these models are legitimate, major technical advances, that they are getting better *very* quickly and that they are seeing rapid adoption. So, this section offers some thoughts on where opportunities may be greatest, and where there may be pitfalls for quality and safety professionals to watch out for. Overall, foundation models offer tremendous opportunity to improve health on an almost startling scale by democratising expertise to where it is needed. This means that quality and safety professionals should expect impact across all of the Six Aims: safety, effectiveness, patient-centredness, timeliness, efficiency and equity.[6]

Early wins: reducing burden for clinicians

In a recent review, Schiff and Shojania highlighted major, thematic challenges to progress in patient safety, providing a roadmap for future quality and safety efforts.[13] One theme focused on staff, noting that issues such as burnout and lack of time for both clinical care and improvement work are core challenges for quality and safety. Generative AI's earliest forays have focused on supporting, and in some cases removing, documentation burden for clinicians, which is a well-documented contributor to these issues. For example, early attention has focused on using AI to support paperwork and processing of pre-approvals of diagnostic tests and therapies by health insurers (known as prior authorisation,[14] and widely acknowledged to be a significant burden for providers). This could have significant benefit not just in efficiency but in reducing provider burnout and increasing time spent with patients. Improving documentation was also an area highlighted in a 2024 viewpoint by Mayo Clinic's chief executive officer (CEO), with examples such as operative notes and automating administrative documentation.[15] Documentation-related healthcare tasks across the spectrum are likely to be significantly affected, since these models handle *language* at their very core. Another area in quality and safety where foundation models may turn out to be helpful is in measurement. One study estimated that quality measures cost just one institution $5.6 million annually to gather and report, which extrapolates to billions of dollars of annual expense for the USA alone.[16] Foundation models are likely to be effective at many of these kinds of summarisation and extraction tasks—allowing health systems to focus resources on improvement instead of measurement.

Communication, organisational health literacy and patient safety

How healthcare organisations communicate with patients and families matters. Albert Wu has written about the social determinants of patient safety[17]—the intersection of social determinants of health and patient safety. One key element of this is organisational health literacy, which the National Academy of Medicine has defined as how organisations 'make it easier for people to navigate, understand, and use information and services to take care of their

health'.[18] While language barriers, which are clearly associated with quality and safety implications,[19] are one aspect of this, Wu gives a nice example showing the broader context and clinical impact even when only a single language is involved: 'An organization with low health literacy fails to enable all individuals to find, understand, and use information and service to help them make medical decisions and care for themselves. In such an organization, it may be difficult for the person with diabetes to maintain the medications and equipment needed to prevent hypoglycemic episodes.'[17]

Generative AI may be well suited to help in this arena. How? Organisations working on improving their communication often find that 'translating' their complex materials into appropriate reading levels, or into digestible video-based or audio-based content, requires substantial effort. Today, giving a large language model a prompt of 'summarise this 30-page guideline at a sixth-grade reading level' often produces good first-draft results, and may speed the time for healthcare organisations to develop this content. Similarly, many of these models are multilingual at their core. Researchers are beginning to publish proof-of-concept efforts in this domain.[20–22] Research is also beginning to emerge on patients' perceptions of AI in their care. For example, a 2023 systematic review of public perceptions of AI in healthcare in the USA concluded that people see healthcare as a domain where 'AI applications could be particularly beneficial', but the message is nuanced and substantial concerns exist.[23]

Supporting clinicians in reducing diagnostic error and delay

Schiff and Shojania also highlight the importance of diagnostic error as a critical challenge for the field going forward.[13] Although research is early in this area, results are intriguing. For example, in a randomised, blinded vignette study, physicians supported by AI outperformed physicians without AI in very complex cases, and this resulted in both more accurate diagnoses and more complete differential diagnoses.[11] Gianrico Farrugia, Mayo Clinic's CEO, describes one potential future state for AI in healthcare as a "streamlined 'second opinion' for physicians",[15] and that seems a likely way that AI may help improve diagnostic excellence: helping clinicians avoid diagnostic anchoring and other pitfalls in one of the most common types of patient safety problems. It is important to sound a note of caution here, though. Kulkarni and Singh recently summarised several practical challenges in how these technical advances might move into the real world—for example, evidence suggests that problems in history-taking and physical examination are the cause of many diagnostic errors, problems that are not likely to be ameliorated by AI.[24]

Risks and need for future research

New technologies also bring new risks for quality and safety. Because the technology is emerging, these also represent important avenues for future research. And as AI begins to support clinical practice, a thoughtful regulatory framework will be critical to ensure that patients, families and clinicians receive the technology's benefits, safely.[25]

Equity and bias

Equity is one of the Six Aims,[6] and ensuring health equity is a fundamental element of good machine learning practice.[26] Equity problems may arise from issues in biases in training data; language models recapitulating biased, sexist or racist misconceptions that lead to health disparities; issues with the algorithm's design or implementation; and variations in performance of AI across different populations that were not reflected in model training. Generative AI may also introduce new kinds of vulnerability because it may be used in a wide variety of situations, rather than in a specific narrowly defined task for which evaluation is more straightforward.[8] This is an area of significant ongoing research.[27]

Model-specific issues such as 'hallucinations'

Large language models have some surprising behaviour. They can be helpful at writing computer code, but they are often bad at straightforward arithmetic. Similarly, they sometimes make things up—like journal references, or in one well-known case, legal precedents in a court filing.[28] Why? Remember that at their core, they are trained to be next-word prediction engines, and that they use an embedding space to represent concepts. So for an arithmetic problem like 2+2=____, they will return the right answer, because there are many examples of this problem in most text corpora. But for a problem like 13 257+23 123=_____, they will often return a five-digit number that looks like the right kind of answer, but is not, because they are predicting the next word, not doing arithmetic (and there are not likely to be many examples of that exact problem in the text they are trained on). Similarly, they will return a plausible-looking journal citation. But it is critical to remember that the basic versions of these models are *not* information-retrieval engines: the models themselves do not look things up in PubMed, they remember things out of their pretrained parameters that look like plausible next words in a sequence. This is an area, though, of active progress: areas like grounding, consistency and attribution[29] are rapidly improving the factuality of results. Similarly, models are learning the ability to recognise categories of problems (like math) and, rather than predicting the next word, can get answers from calculators or other tools.[30]

Continued very rapid development

Finally, quality and safety professionals should anticipate ongoing, very rapid development in the field.

We have seen truly remarkable progress already in just a single year, and it looks as if it will continue. Things that were impossible 2 years ago are now routinely available for software developers. This degree of exponential progress makes the medium-term future difficult to predict in terms of what capabilities may be routinely available in just a few years. And it means that quality and safety professionals will have to think creatively—about the upside for patients, families and clinicians; about the known and emerging risks; and about how these tools may affect their own practice.

Contributors MH drafted, revised and edited the entire manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests MH is employed by Google and owns equity in Alphabet. He also receives royalties from McGraw-Hill for the textbook *Understanding Healthcare Delivery Science*, which includes a chapter on machine learning.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iD
Michael D Howell http://orcid.org/0000-0003-2867-4982

REFERENCES

1 Moor M, Banerjee O, Abad ZSH, *et al*. Foundation models for generalist medical artificial intelligence. *Nature* 2023;616:259–65.
2 McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.
3 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
4 Rajkomar A, Oren E, Chen K, *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;1.
5 Howell MD, Corrado GS, DeSalvo KB. Three Epochs of artificial intelligence in health care. *JAMA* 2024;331:242–4.
6 Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington: National Academy Press, 2001.
7 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst; 2013. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf [Accessed 22 Jan 2024].
8 Embeddings: Translating to a Lower-Dimensional Space. Google developers. Available: https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space [Accessed 8 May 2023].
9 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.
10 Singhal K, Tu T, Gottweis J, *et al*. Towards expert-level medical question answering with large language models. 2023. Available: http://arxiv.org/abs/2305.09617
11 McDuff D, Schaekermann M, Tu T, *et al*. Towards accurate differential diagnosis with large language models. 2023. Available: http://arxiv.org/abs/2312.00164
12 Tu T, Palepu A, Schaekermann M, *et al*. Towards conversational diagnostic AI, 2024. Available: http://arxiv.org/abs/2401.05654 [Accessed 22 Jan 2024].
13 Schiff G, Shojania KG. Looking back on the history of patient safety: an opportunity to reflect and ponder future challenges. *BMJ Qual Saf* 2022;31:148–52.
14 Lenert LA, Lane S, Wehbe R. Could an artificial intelligence approach to prior authorization be more human *J Am Med Inform Assoc* 2023;30:989–94.
15 Farrugia G. How generative AI and large language models can close the gap between data and outcomes in healthcare. World Economic Forum Annual Meeting; 2024. Available: https://www.weforum.org/agenda/2024/01/generative-ai-large-language-models-data-outcomes-healthcare/
16 Saraswathula A, Merck SJ, Bai G, *et al*. The volume and cost of quality metric reporting. *JAMA* 2023;329:1840–7.
17 Wu AW. Social determinants of patient safety: a bridge to better quality of care. *J Patient Saf Risk Manag* 2023;28:96–8.
18 Brach C, Keller D, Hernandez L, *et al*. Ten attributes of health literate health care organizations. *NAM Perspectives* 2012;02.
19 Bell SK, Dong J, Ngo L, *et al*. Diagnostic error experiences of patients and families with limited english-language health literacy or disadvantaged socioeconomic position in a cross-sectional US population-based survey. *BMJ Qual Saf* 2023;32:644–54.
20 Doshi R, Amin K, Khosla P, *et al*. Utilizing large language models to simplify radiology reports: a comparative analysis of chatgpt3.5, chatgpt4.0, google bard, and microsoft bing. *Radiology and Imaging* [Preprint].
21 Amin KS, Mayes L, Khosla P, *et al*. Chatgpt-3.5, chatgpt-4, google bard, and microsoft bing to improve health literacy and communication in pediatric populations and beyond. 2023. Available: http://arxiv.org/abs/2311.10075 [Accessed 22 Jan 2024].
22 Ayre J, Mac O, McCaffery K, *et al*. New frontiers in health literacy: using chatgpt to simplify health information for people in the community. *J Gen Intern Med* 2024;39:573–7.
23 Beets B, Newman TP, Howell EL, *et al*. Surveying public perceptions of artificial intelligence in health care in the United States: systematic review. *J Med Internet Res* 2023;25:e40337.
24 Kulkarni PA, Singh H. Artificial intelligence in clinical diagnosis: opportunities, challenges, and Hype. *JAMA* 2023;330:317–8.
25 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;6.
26 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
27 Bommasani R, Liang P, Lee T. Holistic evaluation of language models. *Ann N Y Acad Sci* 2023;1525:140–6.

28 Neumeister L. Lawyers blame chatgpt for tricking them into citing bogus case law. AP News; 2023. Available: https://apnews.com/article/artificial-intelligence-chatgpt-courts-e15023d7e6fdf4f099aa122437dbb59b [Accessed 21 Aug 2023].
29 Roit P, Ferret J, Shani L. Factually consistent summarization via reinforcement learning with textual entailment feedback. 2023. Available: http://arxiv.org/abs/2306.00186
30 Schick T, Dwivedi-Yu J, Dessì R, *et al*. Toolformer: language models can teach themselves to use tools. 2023. Available: http://arxiv.org/abs/2302.04761 [Accessed 21 Aug 2023].
31 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. 2017. Available: http://arxiv.org/abs/1706.03762 [Accessed 25 Mar 2024].
32 Agostinelli A, Denk TI, Borsos Z, *et al*. Musiclm: generating music from text. 2023. Available: http://arxiv.org/abs/2301.11325 [Accessed 10 May 2023].
33 Bommasani R, Hudson DA, Adeli E, *et al*. On the opportunities and risks of foundation models, 2021. Available: http://arxiv.org/abs/2108.07258 [Accessed 10 May 2023].