OPEN ACCESS

Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities

Mertcan Sevgi ,[1,2] Fares Antaki ,[1,2,3] Pearse A Keane [1,2,4]

[1]Institute of Ophthalmology, University College London, London, UK
[2]Moorfields Eye Hospital NHS Foundation Trust, London, UK
[3]The CHUM School of Artificial Intelligence in Healthcare, Montreal, Quebec, Canada
[4]NIHR Moorfields Biomedical Research Centre, London, UK

Correspondence to
Dr Pearse A Keane, UCL Institute of Ophthalmology, London, UK; p.keane@ucl.ac.uk

Received 8 December 2023
Accepted 19 March 2024

© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY. Published by BMJ.

To cite: Sevgi M, Antaki F, Keane PA. *Br J Ophthalmol* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/bjo-2023-325046

ABSTRACT

Foundation models are the next generation of artificial intelligence that has the potential to provide novel use cases for healthcare. Large language models (LLMs), a type of foundation model, are capable of language comprehension and the ability to generate human-like text. Researchers and developers have been tuning LLMs to optimise their performance in specific tasks, such as medical challenge problems. Until recently, tuning required technical programming expertise, but the release of custom generative pre-trained transformers (GPTs) by OpenAI has allowed users to tune their own GPTs with natural language. This has the potential to democratise access to high-quality bespoke LLMs globally. In this review, we provide an overview of LLMs, how they are tuned and how custom GPTs work. We provide three use cases of custom GPTs in ophthalmology to demonstrate the versatility and effectiveness of these tools. First, we present 'EyeTeacher', an educational aid that generates questions from clinical guidelines to facilitate learning. Second, we built 'EyeAssistant', a clinical support tool that is tuned with clinical guidelines to respond to various physician queries. Lastly, we design 'The GPT for GA', which offers clinicians a comprehensive summary of emerging management strategies for geographic atrophy by analysing peer-reviewed documents. The review underscores the significance of custom instructions and information retrieval in tuning GPTs for specific tasks in ophthalmology. We also discuss the evaluation of LLM responses and address critical aspects such as privacy and accountability in their clinical application. Finally, we discuss their potential in ophthalmic education and clinical practice.

INTRODUCTION

Foundation models are a new paradigm for building artificial intelligence (AI) systems. They have gained considerable traction due to improvements in computing power, the development of transformer model architecture and the availability of large datasets.[1] One area that has seen significant utility is natural language processing (NLP), with the advent of large language models (LLMs).[2] LLMs are foundation models that are trained on large corpora of text, which gives them the capability of language comprehension and the ability to generate human-like text.[1 3] A notable example is generative pretrained transformers (GPTs).[3]

There has been growing interest in evaluating the role of LLMs in healthcare.[4] Despite not being specifically trained in medical knowledge, LLMs have shown generalisation capacities in many domains, including medical challenge problems.[5] As LLMs are updated, their performance in answering medical examination questions is improving, including in the field of ophthalmology.[4 6–11] However, since those generic models may not be trained on vetted medical information, their deployment for patient care will be a challenge.[12] To address this, generic LLMs can be tuned with domain-specific information. For example, models such as BiomedBERT and BioGPT have been trained on content from peer-reviewed literature and Med-PaLM on clinical question databases.[4 13 14] Models can also undergo task-specific tuning, such as instruction tuning, as is the case with Med-PaLM 2. Those approaches have shown considerable gains in biomedical NLP tasks, including answering medical examination questions.[7 15] Prompting strategies (different ways of querying the model) can also be used to augment models. An example of this is Medprompt. It uses innovative prompting strategies with GPT-4 to surpass Med-PaLM 2, which uses domain-specific and task-specific tuning in answering medical examination questions.[5]

As valuable as these solutions are, there is still the issue of not capturing the most up-to-date information as the training and tuning are performed with a snapshot of data up to a certain point in time.[16] Due to the extensive volume of data and the time-consuming nature of the training process, there is an inherent delay in updating the knowledge base of LLMs. This is particularly significant in the field of medicine, where up-to-date knowledge and evidence-based practice are fundamental to quality healthcare. Recently, some LLMs were augmented with real-time internet browsing abilities, allowing them to search the internet for up-to-date content to formulate responses.[17 18] Up-to-date medical information is a pertinent feature of evidence-based medicine,[19] and thus having real-time information is a crucial feature in LLMs for use in a clinical setting.

In November 2023, OpenAI introduced a feature allowing users to customise their own GPTs using natural language.[20] During this interaction, GPT developers can provide custom instructions to the GPT to determine its function, interaction with users, method of answering

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.

questions, tone and how it retrieves information. GPT can be instructed to use external tools such as internet searches and/or documents uploaded by the developer. In effect, this leverages custom GPTs for their natural language abilities and information retrieval capacity from predetermined sources. This degree of customisation can help overcome issues with LLMs providing false information and reduce inaccuracies if there is more reliance on using trusted, predetermined sources.[21]

In this review, we provide three use cases of custom GPTs in the field of ophthalmology and how clinicians may interact with these tools. First, 'EyeTeacher' is a teaching tool that creates questions from selected clinical guidelines. Second, a clinical assistant, 'EyeAssistant, answers clinical queries, tuned to clinical guidelines. Third, 'The GPT for GA' can provide clinicians with an overview of the current management of geographic atrophy (GA) by retrieving information from peer-reviewed documents uploaded to the GPT. Through these illustrative examples, we show how careful custom instructions and information retrieval are able to tune ChatGPT to specific tasks. We also review how LLMs can be evaluated and explore the privacy and accountability of using these tools in clinical practice.

CUSTOMISING LLMS

LLMs are trained on large corpora of text, amassing over a trillion words.[3 22] Many models have emerged and are accessible through their respective interfaces, such as Gemini (by Google), GPT-4 (by OpenAI), Claude 2 (by Anthropic) and LLaMA (open-access by Meta AI).[23] These models can be fine-tuned with domain-specific information; however, this requires technical programming expertise. An example of a fine-tuned model is Neuro-GPTx, which is a content-enriched chatbot for the management of vestibular schwannoma trained on over 4000 peer-reviewed articles. [24] Another example is Almanac, a medicine-based model that acquires domain-specific knowledge from textbooks and preselected web documents, as well as browsing predetermined web domains.[21] Also embedded into the model architecture is a calculator, which can overcome the limitations of models being unable to count. These features can improve accuracy and reliability in answering clinical scenario questions.[21]

Customising LLMs encompasses a variety of approaches, including data-driven strategies and those focused on interaction-level enhancements.[25] These methods can be effectively combined to optimise model performance and adaptability. At the data level, a pretrained LLM can be adjusted using a dataset of labelled examples through supervised tuning or fine-tuning.[26] This method requires substantial amounts of high-quality input–output pairs to create a bespoke model tailored for specific responses in a desired domain. Although typically requiring technical expertise, more accessible tuning methods have emerged, such as within the Google ecosystem.[27] For example, Vertex AI enables users to upload JSON Lines files within an automated machine learning framework to tune an LLM.

At the interaction level, prompt engineering and retrieval augmented generation (RAG) can provide further customisation through interaction with user inputs and external sources of information, respectively. Prompt engineering (and custom instructions) guide model behaviours through natural language instructions, making them more adaptable to specific user needs.[28] This is enabled by the ability of LLMs to temporarily learn from instructions without changing their internal parameters.[3 29] RAG, on the other hand, enables LLMs to augment their responses by grounding their knowledge on external sources.[30] LLMs would retrieve knowledge from a set of fixed sources, such as documents, similar to what is seen in OpenAI's Custom GPTs. Beyond the OpenAI ecosystem, other user-friendly tools have surfaced, including Cohere AI and Google's NotebookLM.[31 32]

In this review, we primarily focus on how GPT models can be customised using natural language for custom instructions and employing user-friendly interfaces for RAG through

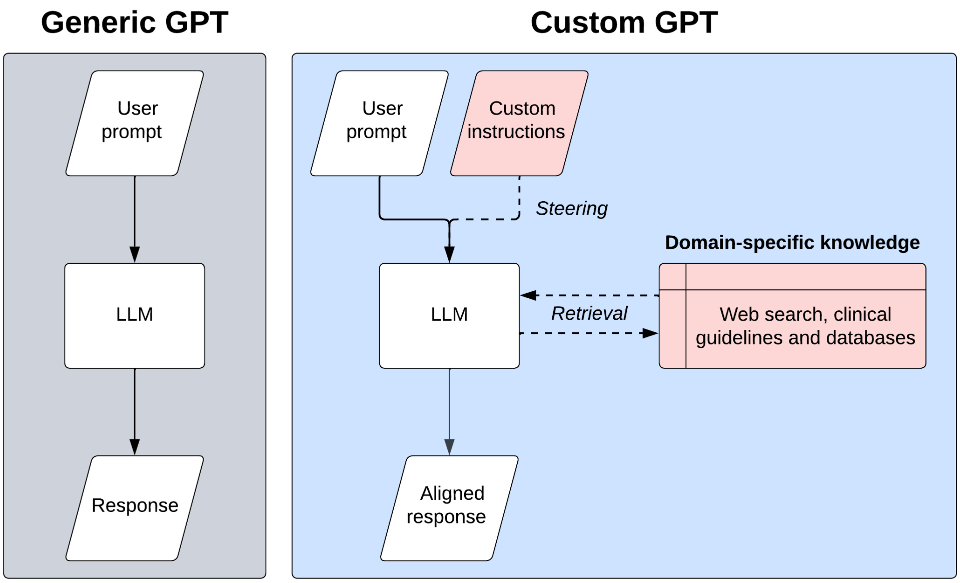


Figure 1 Simplified comparison between generic and custom GPT architectures. Generic GPTs operate linearly: a user prompt is processed by an LLM to generate a response. Custom GPTs integrate custom instructions and domain-specific knowledge into the process. Here, a user prompt is combined with custom instructions (provided by the developer) to steer the LLM, which can also retrieve external knowledge (eg, web searches, clinical guidelines and databases) to produce an aligned response. GPT, generative pretrained transformer; LLM, large language model.

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.

document uploads. This approach aligns particularly well with healthcare applications, where maintaining factual correctness, professional tone and adherence to clinical guidelines is paramount.

CUSTOM GPTS

Tuning custom GPTs involves custom instructions with natural language and content retrieval through RAG. We provide an overview of how custom GPTs work in figure 1. 'GPT Builder' is the backend for custom GPTs, allowing users to tune the model using natural language, as shown in figure 2. Currently, up to 10 documents may be provided for RAG, with a token limit of 2 million per document. [33] Tokens are a unit of analysis employed by LLMs to process text. One token is roughly equivalent to four characters, which corresponds to 0.75 words.[34] A limit of 2 million tokens per document translates to approximately 1.5 million words, which is more than sufficient for the majority of clinical documents, guidelines and textbooks.

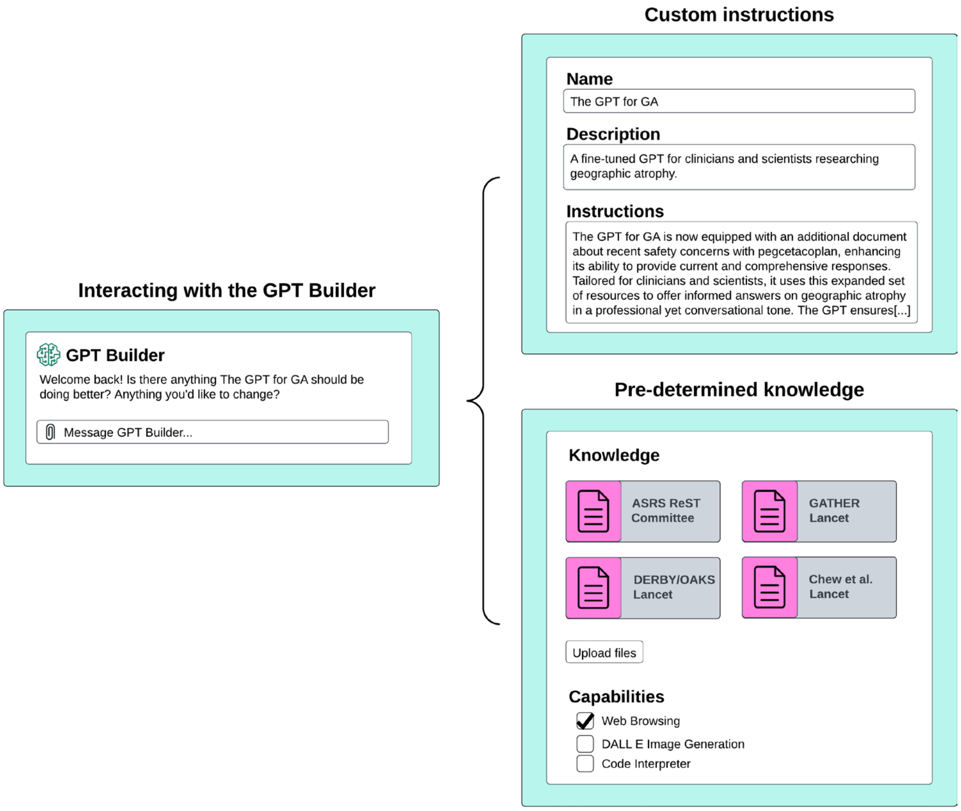
Custom instructions allow for a more steerable model to align with specific user intentions. For use in healthcare, one may want the model to uphold factual correctness, professional tone, privacy and confidentiality, limitation acknowledgement and risk communication, to name some examples. Regarding content retrieval, LLMs are prone to fabricate information, which is commonly termed 'hallucinations'. [35 36] This is a concern for clinical use and will be a limiting factor in LLM deployment. RAG may improve LLM outputs and decrease hallucinations by grounding the model on external sources of knowledge to supplement the LLM's internal representation of ophthalmology.[37] For our custom GPTs, we align with the values of evidence-based medicine, providing peer-reviewed literature and clinical guidelines as domain-specific knowledge.[38] Clinical guidelines are particularly valuable as they represent the consensus of an expert group from a particular organisation. [39 40]. By tailoring the custom GPT to clinical guidelines, we can provide region-specific domain knowledge that is applicable to the healthcare setting we work in.

USE CASES

Education: active learning with generated questions

'EyeTeacher' demonstrates how users can enhance their understanding of clinical guidelines by answering questions generated by the custom GPT. As an active learning tool, this method of learning can promote more effective knowledge acquisition and retention.[41] As exemplars, we provide two clinical guidelines on the management of diabetic retinopathy (DR) and angle closure glaucoma (ACG) from the Royal College of Ophthalmologists.[42 43] We chose these topics as they are common conditions relevant to general ophthalmologists. Both guidelines are within the token limit of the custom GPT (DR: 147 pages, approximately 66 000 tokens. ACG: 44 pages, approximately 19 000 tokens). We showcase an example in figure 3.

We instructed the custom GPT to be an interactive, active learning tool, asking multiple-choice questions with a single best answer. Questions would be generated from the uploaded documents. We also instructed the GPT to ask for the level of experience of the user and to tailor the difficulty of the questions to the



Welcome back! Is there anything The GPT for GA should be doing better? Anything you'd like to change?

Name: The GPT for GA

Description: A fine-tuned GPT for clinicians and scientists researching geographic atrophy.

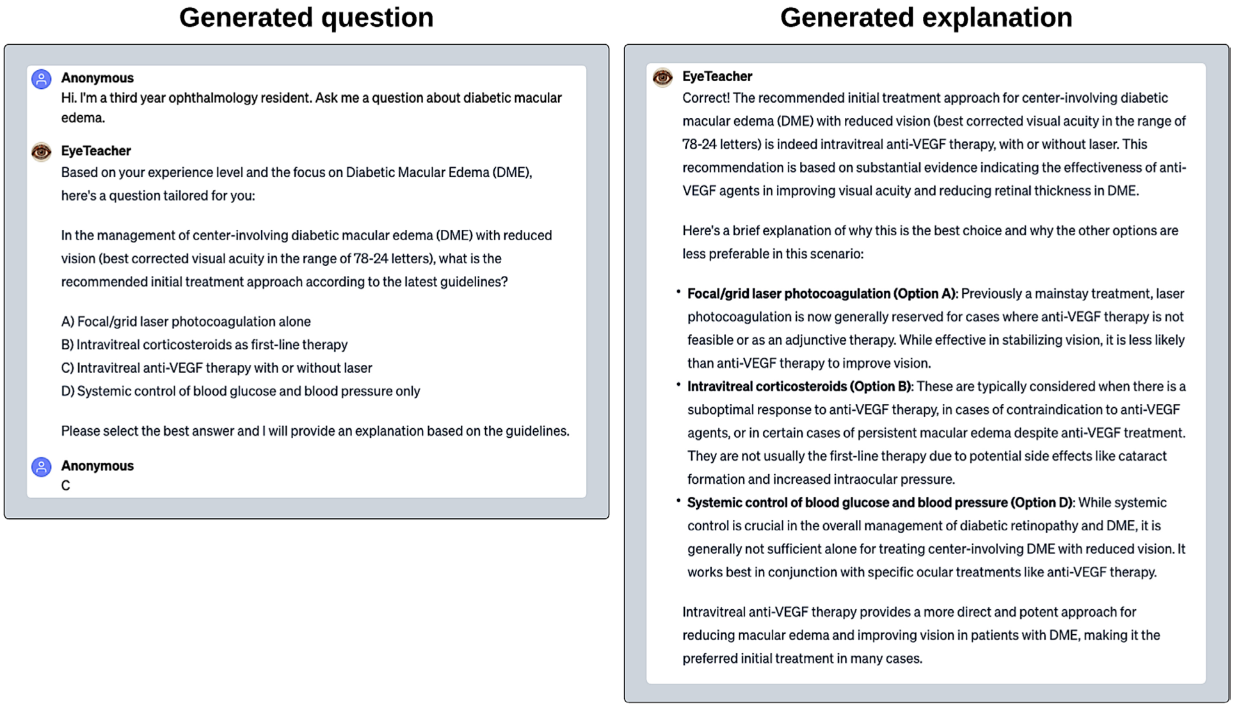
Instructions: The GPT for GA is now equipped with an additional document about recent safety concerns with pegcetacoplan, enhancing its ability to provide current and comprehensive responses. Tailored for clinicians and scientists, it uses this expanded set of resources to offer informed answers on geographic atrophy in a professional yet conversational tone. The GPT ensures[...]

Knowledge: ASRS ReST Committee; GATHER Lancet; DERBY/OAKS Lancet; Chew et al. Lancet

Capabilities:
☑ Web Browsing
☐ DALL E Image Generation
☐ Code Interpreter

Figure 2 Custom GPTs are built in natural language using a builder chatbot. The GPT builder enables users to input custom instructions using natural language and upload specific knowledge datasets for retrieval. These modifications are implemented in the GPT's backend. Additionally, the builder offers the option to enhance or limit functionalities, such as web browsing, image creation and code interpretation and generation, although the latter may be less relevant for our specific use cases. GPT, generative pretrained transformer.

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.



Generated question

Anonymous
Hi. I'm a third year ophthalmology resident. Ask me a question about diabetic macular edema.

EyeTeacher
Based on your experience level and the focus on Diabetic Macular Edema (DME), here's a question tailored for you:

In the management of center-involving diabetic macular edema (DME) with reduced vision (best corrected visual acuity in the range of 78-24 letters), what is the recommended initial treatment approach according to the latest guidelines?

A) Focal/grid laser photocoagulation alone
B) Intravitreal corticosteroids as first-line therapy
C) Intravitreal anti-VEGF therapy with or without laser
D) Systemic control of blood glucose and blood pressure only

Please select the best answer and I will provide an explanation based on the guidelines.

Anonymous
C

Generated explanation

EyeTeacher
Correct! The recommended initial treatment approach for center-involving diabetic macular edema (DME) with reduced vision (best corrected visual acuity in the range of 78-24 letters) is indeed intravitreal anti-VEGF therapy, with or without laser. This recommendation is based on substantial evidence indicating the effectiveness of anti-VEGF agents in improving visual acuity and reducing retinal thickness in DME.

Here's a brief explanation of why this is the best choice and why the other options are less preferable in this scenario:

- **Focal/grid laser photocoagulation (Option A):** Previously a mainstay treatment, laser photocoagulation is now generally reserved for cases where anti-VEGF therapy is not feasible or as an adjunctive therapy. While effective in stabilizing vision, it is less likely than anti-VEGF therapy to improve vision.
- **Intravitreal corticosteroids (Option B):** These are typically considered when there is a suboptimal response to anti-VEGF therapy, in cases of contraindication to anti-VEGF agents, or in certain cases of persistent macular edema despite anti-VEGF treatment. They are not usually the first-line therapy due to potential side effects like cataract formation and increased intraocular pressure.
- **Systemic control of blood glucose and blood pressure (Option D):** While systemic control is crucial in the overall management of diabetic retinopathy and DME, it is generally not sufficient alone for treating center-involving DME with reduced vision. It works best in conjunction with specific ocular treatments like anti-VEGF therapy.

Intravitreal anti-VEGF therapy provides a more direct and potent approach for reducing macular edema and improving vision in patients with DME, making it the preferred initial treatment in many cases.

Figure 3 EyeTeacher is a custom GPT built for education. In this enhanced learning scenario, EyeTeacher presents a multiple-choice question about diabetic macular oedema. When a student selects an answer, EyeTeacher elaborates on why the answer is correct and explains why each distractor is incorrect. This approach is crucial because it does not merely validate the student's choice but educates them about the nuances of each option. This method promotes active learning, as students do not just memorise the right answer; they understand the rationale behind each option. The explanations are free of hallucinations and factually correct. GPT, generative pretrained transformer.

user's level of experience. In keeping with the format of medical question books, we instructed the GPT to provide explanations for the answer. For content retrieval, the GPT is instructed to create the questions from the documents provided. The ability to engage with EyeTeacher is akin to conversing with a personal tutor. It provides flexibility for users to direct their own learning. They can ask for easier or harder questions. They are able to select topics and ask for clarity on the topic if required. The example we have provided is from a clinical guideline; however, one could also use this method to study for exams by uploading notes and asking questions about a specific topic.

Clinical assistant: interacting with clinical guidelines

'EyeAssistant' is provided with the same clinical guidelines as EyeTeacher, but has separate custom instructions, leading to a different use case and user experience. The purpose of this use case is to act as a clinical assistant. The user is able to converse with the assistant, who has domain-specific knowledge from clinical guidelines. Clinicians would typically query guidelines by reading through them; however, EyeAssistant provides an interactive experience. Users can ask for specific sections in the guidelines as well as provide clinical information such as examination findings or demographic information. EyeAssistant will browse and retrieve the relevant content. To improve reliability, we can include custom instructions in the GPT to ask for further clarification if the information provided is not sufficient to generate an answer. We provide an example in figure 4.

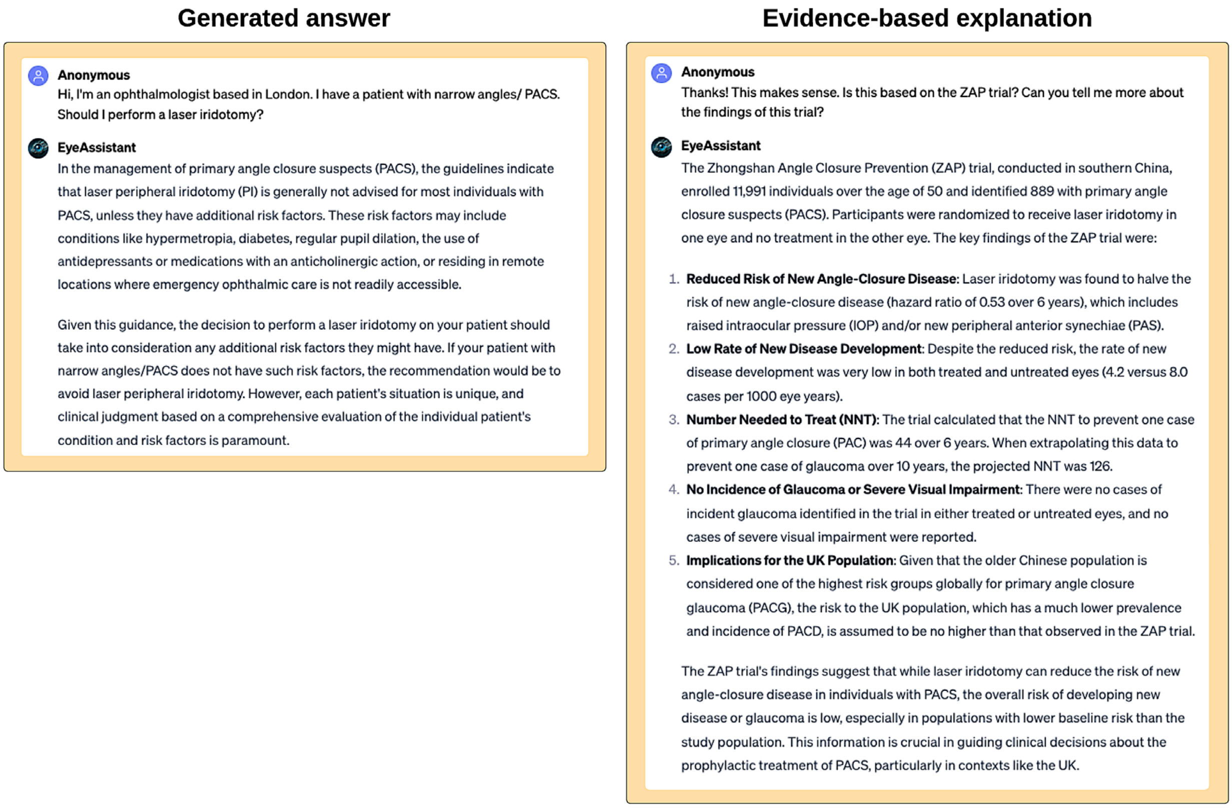
Summarising peer-reviewed research with custom GPTs

Custom GPTs have emerged as powerful tools for summarising and extracting knowledge from peer-reviewed articles. These custom models offer several advantages over traditional methods of literature review, including enhanced accessibility, efficient knowledge discovery and personalised learning. One of the primary benefits of using custom GPTs for summarising peer-reviewed research is their ability to transform complex scientific literature into concise and understandable language. This makes the information more readily accessible to a wider audience, including researchers, clinicians and even patients. By distilling the key findings, conclusions and implications of research articles, custom GPTs can bridge the gap between scientific experts and those seeking to understand the latest advancements in their field.

In our use case example, 'The GPT for GA' summarises peer-reviewed articles on the latest treatment for GA. Custom instructions include tailoring responses to clinicians and scientists and offering informed answers on GA in a professional yet conversational tone. It will focus on providing accurate and relevant information while avoiding speculation and not offering patient-specific medical advice. For content retrieval, The GPT for GA exclusively retrieves information from the provided documents, including the new pivotal GA trials.[44 45] At the time of writing, there is a limit of 10 document uploads; however, as we expect further advancements in the capabilities of custom GPTs, we will likely see an increase in this capacity.

The treatment for GA is an interesting example, as there are many clinical trials underway to find a treatment.[46], Recently, new drugs have been approved to treat GA. [44 45] However, there is concern about adverse events from the newly approved medications.[47 48] Thus, 'The GPT for GA' is able to summarise the most recent literature, providing a balanced overview to readers. We provide an example in figure 5.

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.



Generated answer

Anonymous
Hi, I'm an ophthalmologist based in London. I have a patient with narrow angles/ PACS. Should I perform a laser iridotomy?

EyeAssistant
In the management of primary angle closure suspects (PACS), the guidelines indicate that laser peripheral iridotomy (PI) is generally not advised for most individuals with PACS, unless they have additional risk factors. These risk factors may include conditions like hypermetropia, diabetes, regular pupil dilation, the use of antidepressants or medications with an anticholinergic action, or residing in remote locations where emergency ophthalmic care is not readily accessible.

Given this guidance, the decision to perform a laser iridotomy on your patient should take into consideration any additional risk factors they might have. If your patient with narrow angles/PACS does not have such risk factors, the recommendation would be to avoid laser peripheral iridotomy. However, each patient's situation is unique, and clinical judgment based on a comprehensive evaluation of the individual patient's condition and risk factors is paramount.

Evidence-based explanation

Anonymous
Thanks! This makes sense. Is this based on the ZAP trial? Can you tell me more about the findings of this trial?

EyeAssistant
The Zhongshan Angle Closure Prevention (ZAP) trial, conducted in southern China, enrolled 11,991 individuals over the age of 50 and identified 889 with primary angle closure suspects (PACS). Participants were randomized to receive laser iridotomy in one eye and no treatment in the other eye. The key findings of the ZAP trial were:

1. **Reduced Risk of New Angle-Closure Disease**: Laser iridotomy was found to halve the risk of new angle-closure disease (hazard ratio of 0.53 over 6 years), which includes raised intraocular pressure (IOP) and/or new peripheral anterior synechiae (PAS).
2. **Low Rate of New Disease Development**: Despite the reduced risk, the rate of new disease development was very low in both treated and untreated eyes (4.2 versus 8.0 cases per 1000 eye years).
3. **Number Needed to Treat (NNT)**: The trial calculated that the NNT to prevent one case of primary angle closure (PAC) was 44 over 6 years. When extrapolating this data to prevent one case of glaucoma over 10 years, the projected NNT was 126.
4. **No Incidence of Glaucoma or Severe Visual Impairment**: There were no cases of incident glaucoma identified in the trial in either treated or untreated eyes, and no cases of severe visual impairment were reported.
5. **Implications for the UK Population**: Given that the older Chinese population is considered one of the highest risk groups globally for primary angle closure glaucoma (PACG), the risk to the UK population, which has a much lower prevalence and incidence of PACD, is assumed to be no higher than that observed in the ZAP trial.

The ZAP trial's findings suggest that while laser iridotomy can reduce the risk of new angle-closure disease in individuals with PACS, the overall risk of developing new disease or glaucoma is low, especially in populations with lower baseline risk than the study population. This information is crucial in guiding clinical decisions about the prophylactic treatment of PACS, particularly in contexts like the UK.

Figure 4 EyeAssistant is a custom GPT built for clinical support. In this clinical scenario, EyeAssistant presents an evidence-based response regarding the management of narrow angles or primary angle closure suspects. From the guideline, it identifies risk factors for angle closure that may warrant prophylactic iridotomy. When prompted to justify the answer and describe the popular Zhongshan Angle Closure Prevention trial, it summarises the findings adequately and contextualises the findings for the London-based ophthalmologist. The explanations are free of hallucinations and factually correct. GPT, generative pretrained transformer

EVALUATING RESPONSES

The use of LLMs for healthcare requires models that can avoid hallucinations and acknowledge their limitations.[49] Companies developing LLMs are aware of this and developing models with increased 'honesty'. For example, Anthropic released Claude 2.1. The company reported that the number of hallucinations was halved compared with their previous version.[50] In addition to providing fewer incorrect answers, Claude 2.1 has an increased number of responses where it declines to answer due to a lack of knowledge. Another example is Almanac. This model is evaluated on factuality, which can be improved with the use of real-time internet search retrieval and calculators. The completeness and safety of responses are improved with custom instructions.[21]

Currently, there is no official consensus for evaluating models and responses; however, we can gather insight from studies that evaluate LLMs on answering medical exam questions as well as summarising medical literature.[15 51] Alongside assessing factual correctness, responses are assessed across multiple evaluation metrics such as comprehension, coherence, knowledge recall, reasoning, potential for harm and relevance, among others.[4] LLM responses were compared against human responses using these evaluation metrics, which also found that reviewers preferred LLM responses compared with human experts.[15 24 51] LLMs can be adjusted to perform in varying creative settings. Evaluating ChatGPT-4 explanations of answers for ophthalmology exam questions revealed that more creative settings are preferred.[11]

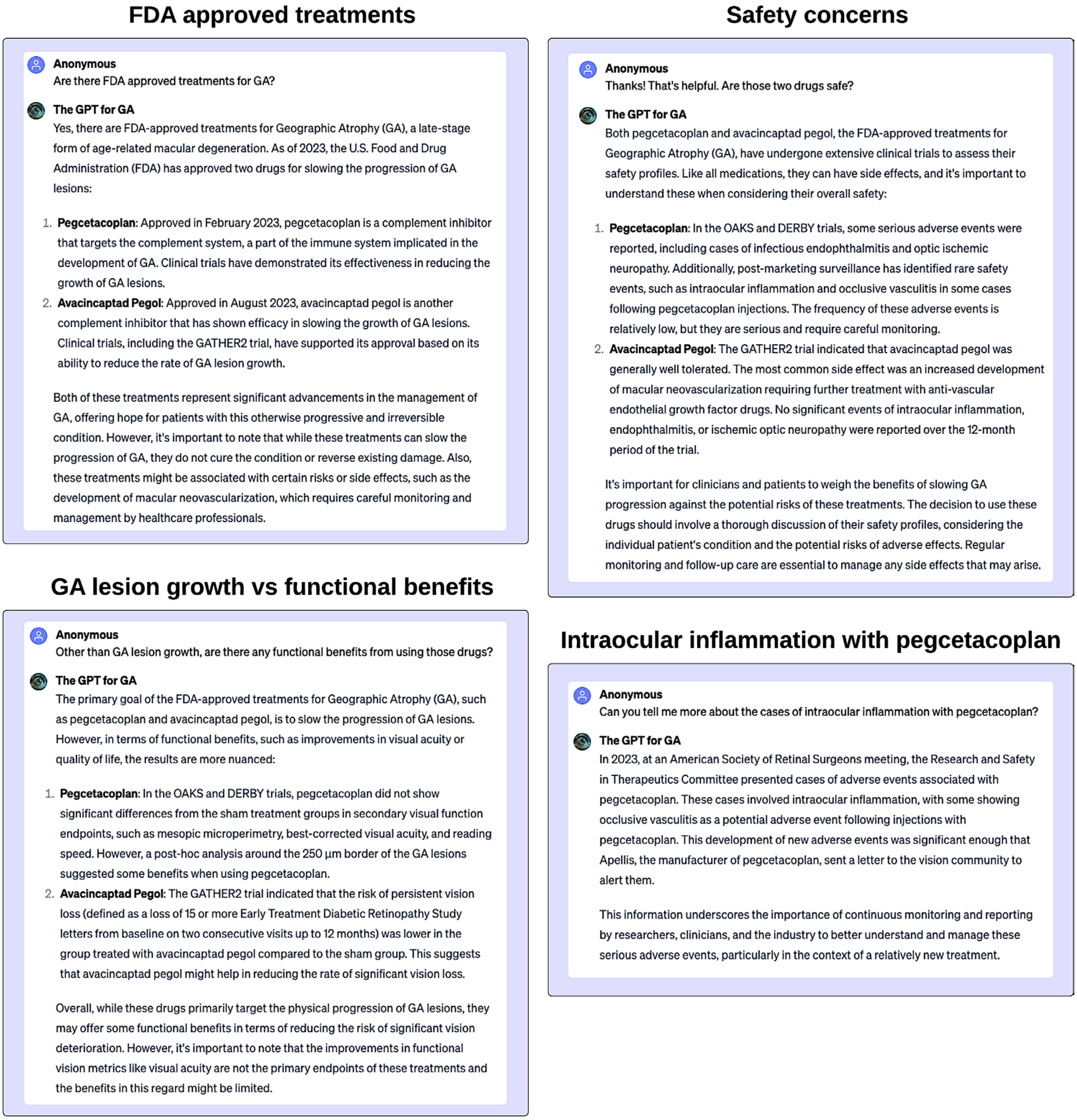
MAINTAINING SAFETY AND ACCOUNTABILITY

The increasing adoption of LLMs in healthcare raises a critical question: who should be held accountable for any adverse outcomes arising from their use? Within the existing legal framework, clinicians bear the ultimate responsibility for patient outcomes. [52] Clinicians should exercise the same degree of caution with LLMs as they do with other medical tools until their efficacy and safety have been rigorously validated. If clinicians wish to use LLMs to aid in their work, tuning with custom GPTs could improve relevance. Uploading their own documents and setting strict instructions are methods users can apply to access more reliable and relevant information. The privacy of data is paramount in healthcare. If clinicians are entering patient information when seeking guidance from LLMs, we need to consider how this information can be kept private. Some LLM providers are offering secure enterprise services so that conversations with the LLMs are not used to train the model and are encrypted.[53 54] Including privacy controlling measures, there needs to be regulatory oversight for using LLMs in healthcare.[55] This would require a framework that can evaluate LLMs for their NLP, translational value and governance model to ensure fairness, transparency, trustworthiness, accountability and privacy. [56 57]

LIMITATIONS OF CUSTOM GPTS

The innovative use of custom GPTs in medical education, workflow improvement and clinical assistance has demonstrated considerable potential. However, it is crucial to acknowledge

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.



FDA approved treatments

Anonymous
Are there FDA approved treatments for GA?

The GPT for GA
Yes, there are FDA-approved treatments for Geographic Atrophy (GA), a late-stage form of age-related macular degeneration. As of 2023, the U.S. Food and Drug Administration (FDA) has approved two drugs for slowing the progression of GA lesions:

1. **Pegcetacoplan**: Approved in February 2023, pegcetacoplan is a complement inhibitor that targets the complement system, a part of the immune system implicated in the development of GA. Clinical trials have demonstrated its effectiveness in reducing the growth of GA lesions.
2. **Avacincaptad Pegol**: Approved in August 2023, avacincaptad pegol is another complement inhibitor that has shown efficacy in slowing the growth of GA lesions. Clinical trials, including the GATHER2 trial, have supported its approval based on its ability to reduce the rate of GA lesion growth.

Both of these treatments represent significant advancements in the management of GA, offering hope for patients with this otherwise progressive and irreversible condition. However, it's important to note that while these treatments can slow the progression of GA, they do not cure the condition or reverse existing damage. Also, these treatments might be associated with certain risks or side effects, such as the development of macular neovascularization, which requires careful monitoring and management by healthcare professionals.

GA lesion growth vs functional benefits

Anonymous
Other than GA lesion growth, are there any functional benefits from using those drugs?

The GPT for GA
The primary goal of the FDA-approved treatments for Geographic Atrophy (GA), such as pegcetacoplan and avacincaptad pegol, is to slow the progression of GA lesions. However, in terms of functional benefits, such as improvements in visual acuity or quality of life, the results are more nuanced:

1. **Pegcetacoplan**: In the OAKS and DERBY trials, pegcetacoplan did not show significant differences from the sham treatment groups in secondary visual function endpoints, such as mesopic microperimetry, best-corrected visual acuity, and reading speed. However, a post-hoc analysis around the 250 µm border of the GA lesions suggested some benefits when using pegcetacoplan.
2. **Avacincaptad Pegol**: The GATHER2 trial indicated that the risk of persistent vision loss (defined as a loss of 15 or more Early Treatment Diabetic Retinopathy Study letters from baseline on two consecutive visits up to 12 months) was lower in the group treated with avacincaptad pegol compared to the sham group. This suggests that avacincaptad pegol might help in reducing the rate of significant vision loss.

Overall, while these drugs primarily target the physical progression of GA lesions, they may offer some functional benefits in terms of reducing the risk of significant vision deterioration. However, it's important to note that the improvements in functional vision metrics like visual acuity are not the primary endpoints of these treatments and the benefits in this regard might be limited.

Safety concerns

Anonymous
Thanks! That's helpful. Are those two drugs safe?

The GPT for GA
Both pegcetacoplan and avacincaptad pegol, the FDA-approved treatments for Geographic Atrophy (GA), have undergone extensive clinical trials to assess their safety profiles. Like all medications, they can have side effects, and it's important to understand these when considering their overall safety:

1. **Pegcetacoplan**: In the OAKS and DERBY trials, some serious adverse events were reported, including cases of infectious endophthalmitis and optic ischemic neuropathy. Additionally, post-marketing surveillance has identified rare safety events, such as intraocular inflammation and occlusive vasculitis in some cases following pegcetacoplan injections. The frequency of these adverse events is relatively low, but they are serious and require careful monitoring.
2. **Avacincaptad Pegol**: The GATHER2 trial indicated that avacincaptad pegol was generally well tolerated. The most common side effect was an increased development of macular neovascularization requiring further treatment with anti-vascular endothelial growth factor drugs. No significant events of intraocular inflammation, endophthalmitis, or ischemic optic neuropathy were reported over the 12-month period of the trial.

It's important for clinicians and patients to weigh the benefits of slowing GA progression against the potential risks of these treatments. The decision to use these drugs should involve a thorough discussion of their safety profiles, considering the individual patient's condition and the potential risks of adverse effects. Regular monitoring and follow-up care are essential to manage any side effects that may arise.

Intraocular inflammation with pegcetacoplan

Anonymous
Can you tell me more about the cases of intraocular inflammation with pegcetacoplan?

The GPT for GA
In 2023, at an American Society of Retinal Surgeons meeting, the Research and Safety in Therapeutics Committee presented cases of adverse events associated with pegcetacoplan. These cases involved intraocular inflammation, with some showing occlusive vasculitis as a potential adverse event following injections with pegcetacoplan. This development of new adverse events was significant enough that Apellis, the manufacturer of pegcetacoplan, sent a letter to the vision community to alert them.

This information underscores the importance of continuous monitoring and reporting by researchers, clinicians, and the industry to better understand and manage these serious adverse events, particularly in the context of a relatively new treatment.

Figure 5 The GPT for GA is a custom GPT built to provide balanced information on the treatments for geographic atrophy. When queried about FDA-approved treatments for GA, The GPT for GA accurately identifies pegcetacoplan (Syfovre) and avacincaptad pegol (Izervay). It appropriately references the pertinent pivotal trials that led to their approval. When prompted to discuss the functional benefits of these treatments, it judiciously cites relevant sources while acknowledging that the associated functional benefits remain limited. Addressing safety concerns, it correctly highlights the elevated risk of macular neovascularisation associated with both drugs and the potential for intraocular inflammation with pegcetacoplan. When asked about the latter, it also provides up-to-date information and cites authoritative sources, such as the American Society of Retina Specialists (ASRS) ReST committee. Of note, it mistakenly refers to the ASRS as the American Society of Retinal Surgeons. FDA, food and drug administration; GA, geographic atrophy; GPT, generative pretrained transformer.

the inherent limitations of these tools to ensure their responsible and effective use.

While a tool like EyeTeacher can facilitate metacognitive learning through generated questions, its effectiveness hinges on the accuracy of the information generated. There is a risk of reinforcing incorrect knowledge if the model generates erroneous content. There is also the risk that EyeTeacher may not probe areas of weakness unless the student explicitly communicates them. In parallel, EyeAssistant, designed to interact with clinical guidelines, may lead to automation bias, where users overrely on AI-generated responses.[58] Incorrect interpretations or incomplete retrieval of information from the guidelines could lead to misguided clinical decisions. Additionally, the models' abilities to accurately retrieve information might vary depending on the structure of the documents and the content placement.[59–61]

For The GPT for GA, the current limitation of the number of document uploads restricts the breadth of information that can be summarised. In addition, the static form of content retrieval might miss out on the latest research developments. Updating this would require manual user uploads of the latest research.

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.

This underscores the benefit of leveraging real-time internet browsing abilities for LLMs. Indeed, the recent approval of new drugs for GA and concerns about their adverse events highlight the need for timely and comprehensive data integration. Finally, while custom instructions and prompt engineering offer flexibility, they are subjective and highly dependent on the user's ability to craft effective prompts. This evolving field requires a nuanced understanding of how LLMs interpret and respond to various prompts. Inaccuracies or spurious features in prompts can lead to misleading or irrelevant responses.[62] Beyond these specific use cases, general limitations of LLMs include their dependence on the training data and its potential biases, and their inability to verify the factual accuracy of generated content. This necessitates careful consideration and validation in contexts where accuracy and reliability are critical, such as in medical and educational applications.

CONCLUSION

The rapid evolution of foundation models, particularly in the realm of medical applications, is an exciting development to witness. As researchers and developers continue to refine these models with specialised tuning techniques, we are moving closer to achieving more suitable models for healthcare applications. The future iterations of foundation models and their potential to generalise in solving medical challenges present an intriguing prospect that could pave the way towards artificial general intelligence. The advent of custom GPTs represents a significant step in democratising these powerful tools, thus enabling a wider array of applications. However, further research is essential, particularly in translating these advancements into practical medical settings. We should next focus on medical challenges that mirror actual clinical practice to validate the real-world utility of these models. With the advent of custom GPTs, a multitude of users now have the opportunity to test and adapt these models. The path forward calls for robust research and collaborative efforts to fully harness the potential of custom GPTs in real-world practice. We are at the threshold of a new era in medical technology, an era where AI not only complements but significantly enhances our healthcare capabilities. This is a thrilling prospect, beckoning for further exploration and innovation in the field.

X Fares Antaki @FaresAntaki and Pearse A Keane @pearsekeane

Contributors Conception and design of the study: MS, FA and PAK; data collection: MS; writing of the manuscript: MS and FA; preparation of figures: FA; supervision: PAK; review and discussion of the results: all authors; edition and revision of the manuscript: all authors.

Funding FA is supported by the Moorfields Eye Charity (GR001474) and the Fonds de Recherche du Québec, Santé (FRQS). PAK is supported by a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Links to the custom GPTs are available upon reasonable request to the authors.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

ORCID iDs
Mertcan Sevgi http://orcid.org/0009-0003-8426-6534
Fares Antaki http://orcid.org/0000-0001-6679-7276
Pearse A Keane http://orcid.org/0000-0002-9239-745X

REFERENCES

1 Bommasani R, Hudson DA, Adeli E, *et al*. On the opportunities and risks of foundation models. n.d. Available: http://arxiv.org/abs/2108.07258
2 Nath S, Marie A, Ellershaw S, *et al*. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in Ophthalmology. *Br J Ophthalmol* 2022;106:889–92.
3 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. 2020 Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
4 Singhal K, Azizi S, Tu T, *et al*. Large language models Encode clinical knowledge. *Nature* 2023;620:172–80.
5 Nori H, Lee YT, Zhang S, *et al*. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. 2023 Available: http://arxiv.org/abs/2311.16452
6 Antaki F, Touma S, Milad D, *et al*. Evaluating the performance of Chatgpt in Ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci* 2023;3.
7 Taylor R, Kardas M, Cucurull G, *et al*. Galactica: A large language model for science. 2022 Available: http://arxiv.org/abs/2211.09085
8 Mihalache A, Huang RS, Popovic MM, *et al*. Performance of an upgraded artificial intelligence Chatbot for ophthalmic knowledge assessment. *JAMA Ophthalmol* 2023;141:798–800.
9 Teebagy S, Colwell L, Wood E, *et al*. Improved performance of Chatgpt-4 on the OKAP examination: A comparative study with Chatgpt-3.5. *Journal of Academic Ophthalmology* 2023;15:e184–7.
10 Raimondi R, Tzoumas N, Salisbury T, *et al*. North East Trainee research in Ophthalmology network (Netrion). comparative analysis of large language models in the Royal college of Ophthalmologists fellowship exams. *Eye* 2023;37:3530–3.
11 Antaki F, Milad D, Chia MA, *et al*. Capabilities of GPT-4 in Ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *Br J Ophthalmol* 2023.
12 Clusmann J, Kolbinger FR, Muti HS, *et al*. The future landscape of large language models in medicine. *Commun Med (Lond)* 2023;3:141.
13 Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model Pretraining for BIOMEDICAL natural language processing. 2020 Available: http://arxiv.org/abs/2007.15779
14 Luo R, Sun L, Xia Y, *et al*. Biogpt: Generative pre-trained transformer for BIOMEDICAL text generation and mining. *Briefings in Bioinformatics* 2022;23.
15 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
16 OpenAI. GPT-4 technical report. 2023 Available: http://arxiv.org/abs/2303.08774
17 OpenAI. Chatgpt Plugins. 2023. Available: https://openai.com/blog/chatgpt-plugins#browsing
18 Google AI. Bard updates: the latest bard news and AI features - Google bard. 2023. Available: https://bard.google.com/updates
19 Sackett DL, Rosenberg WM, Gray JA, *et al*. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312:71–2.
20 OpenAI. Introducing Gpts. 2023. Available: https://openai.com/blog/introducing-gpts
21 Zakka C, Shad R, Chaurasia A, *et al*. Almanac — retrieval-augmented language models for clinical medicine. *NEJM AI* 2024;1:AIoa2300068.
22 Touvron H, Lavril T, Izacard G, *et al*. Llama: open and efficient foundation language models. 2023 Available: https://github.com/facebookresearch/llama
23 Yang J, Jin H, Tang R, *et al*. n.d. Harnessing the power of Llms in practice: A survey on Chatgpt and beyond [Internet]. *ACM Trans Knowl Discov Data*.
24 Guo E, Gupta M, Sinha S, *et al*. neuroGPT-X: toward a clinic-ready large language model. *J Neurosurg* 2024;140:1041–53.
25 Ovadia O, Brief M, Mishaeli M, *et al*. Fine-tuning or retrieval? comparing knowledge injection in Llms. 2023 Available: http://arxiv.org/abs/2312.05934
26 Naveed H, Khan AU, Qiu S, *et al*. A comprehensive overview of large language models. 2023 Available: http://arxiv.org/abs/2307.06435
27 Google Cloud. Tune language foundation models. 2024. Available: https://cloud.google.com/vertex-ai/docs/generative-ai/models/tune-models
28 White J, Fu Q, Hays S, *et al*. A prompt pattern catalog to enhance prompt engineering with Chatgpt. 2023 Available: http://arxiv.org/abs/2302.11382
29 Wei J, Tay Y, Bommasani R, *et al*. Emergent abilities of large language models. 2022 Available: http://arxiv.org/abs/2206.07682
30 Martineau K, IBM Research Blog. What is retrieval-augmented generation? IBM; 2023. Available: https://research.ibm.com/blog/retrieval-augmented-generation-RAG
31 Cohere. 2024. Available: https://cohere.com/
32 Bowen CD, Summersill AR, Google AN, *et al*. Exploring black undergraduate students' communication and biology education experiences about COVID-19 and COVID-19 vaccines during the pandemic. *CBE Life Sci Educ* 2023;22.
33 OpenAI. File Uploads with Gpts and advanced data analysis in Chatgpt. 2023. Available: https://help.openai.com/en/articles/8555545-file-uploads-with-gpts-and-advanced-data-analysis-in-chatgpt

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.

34 OpenAI. Openai Tokenizer. 2023. Available: https://platform.openai.com/tokenizer
35 Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. 2023 Available: http://arxiv.org/abs/2309.05922
36 Ji Z, Lee N, Frieske R, *et al*. Survey of hallucination in natural language generation. 2022 Available: http://arxiv.org/abs/2202.03629
37 Shuster K, Poff S, Chen M, *et al*. Retrieval augmentation reduces hallucination in conversation. Poster presented at the Findings of the Association for Computational Linguistics; Stroudsburg, PA, USA, Punta Cana, Dominican Republic. 10.18653/v1/2021.findings-emnlp.320 Available: https://aclanthology.org/2021.findings-emnlp
38 Masic I, Miokovic M, Muhamedagic B. Evidence based medicine - new approaches and challenges. *Acta Inform Med* 2008;16:219–25.
39 Woolf SH, Grol R, Hutchinson A, *et al*. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. *BMJ* 1999;318:527–30.
40 Graham R, Mancher M, Wolman DM, *et al*. Current best practices and proposed standards for development of trustworthy Cpgs: part 1, getting started. National Academies Press (US); 2011.
41 Prince M. Does active learning work? A review of the research. *J of Engineering Edu* 2004;93:223–31.
42 The Royal College of Ophthalmologists. Diabetic retinopathy guidelines. 2012. Available: https://www.rcophth.ac.uk/resources-listing/diabetic-retinopathy-guidelines/
43 The Royal college of Ophthalmologists. management of angle-closure glaucoma guideline. 2022. Available: https://www.rcophth.ac.uk/resources-listing/management-of-angle-closure-glaucoma-guideline/
44 Khanani AM, Patel SS, Staurenghi G, *et al*. Efficacy and safety of Avacincaptad Pegol in patients with geographic atrophy (Gather2): 12-month results from a randomised, double-masked, phase 3 trial. *Lancet* 2023;402:1449–58.
45 Heier JS, Lad EM, Holz FG, *et al*. Pegcetacoplan for the treatment of geographic atrophy secondary to age-related macular degeneration (OAKS and DERBY): two Multicentre, randomised, double-masked, sham-controlled, phase 3 trials. *Lancet* 2023;402:1434–48.
46 Boyer DS, Schmidt-Erfurth U, van Lookeren Campagne M, *et al*. The pathophysiology of geographic atrophy secondary to age-related macular degeneration and the complement pathway as a therapeutic target. *Retina* 2017;37:819–35.
47 ASRS Research and Safety in Therapeutics (ReST) Committee. Rest committee update on Intraocular inflammation (IOI). 2023 Available: https://www.asrs.org/content/documents/asrs-rest-committee-update-on-intraocular-inflammation-after-ivi-2023.pdf
48 Chew EY. Complement inhibitors for the treatment of geographic atrophy. *Lancet* 2023;402:1396–8.
49 Tan TF, Thirunavukarasu AJ, Campbell JP, *et al*. Generative artificial intelligence through Chatgpt and other large language models in Ophthalmology: clinical applications and challenges. *Ophthalmol Sci* 2023;3.
50 Anthropic PBC. Anthropic. Introducing Claude 2.1; 2023. Available: https://www.anthropic.com/index/claude-2-1
51 Tang L, Sun Z, Idnay B, *et al*. Evaluating large language models on medical evidence summarization. *NPJ Digit Med* 2023;6.
52 Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. *JAMA* 2023;329:1349–50.
53 Anthropic PBC. Anthropic. 2022. Available: https://www.anthropic.com/product
54 OpenAI. Introducing Chatgpt enterprise. 2023. Available: https://openai.com/blog/introducing-chatgpt-enterprise
55 Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or Generative AI) in Healthcare. *NPJ Digit Med* 2023;6:120.
56 Reddy S. Evaluating large language models for use in Healthcare: A framework for Translational value assessment. *Informatics in Medicine Unlocked* 2023;41.
57 Harrer S. Attention is not all you need: the complicated case of ethically using large language models in Healthcare and medicine. *EBioMedicine* 2023;90.
58 Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and Mitigators. *J Am Med Inform Assoc* 2012;19:121–7.
59 Gutierrez CT, Loizides C, Hafez I, *et al*. Acute phase response following pulmonary exposure to soluble and insoluble metal oxide Nanomaterials in mice. *Part Fibre Toxicol* 2023;20:4.
60 Liu NF, Lin K, Hewitt J, *et al*. Lost in the middle: how language models use long contexts. 2023 Available: http://arxiv.org/abs/2307.03172
61 Llmtest_Needleinahaystack: doing simple retrieval from LLM models at various context lengths to measure accuracy. n.d. Available: https://github.com/gkamradt/LLMTest_NeedleInAHaystack
62 Sclar M, Choi Y, Tsvetkov Y, *et al*. Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt Formatting. n.d. Available: http://arxiv.org/abs/2310.11324

Br J Ophthalmol: first published as 10.1136/bjo-2023-325046 on 7 May 2024. Downloaded from http://bjo.bmj.com/ on May 13, 2024 at UCL Library Services. Protected by copyright.