Does a complex prompt alter the diagnostic accuracy of common ophthalmological conditions by GPT-4? : Data Project

Shona Alex Tapiwa M'gadzah, Andrew O'Malley

Submitted to: Journal of Medical Internet Research
on: August 01, 2024

Disclaimer: This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Does a complex prompt alter the diagnostic accuracy of common ophthalmological conditions by GPT-4? : Data Project

Shona Alex Tapiwa M'gadzah[1] BSc; Andrew O'Malley[1] BSc, PhD

[1]School of Medicine University of St Andrews St Andrews GB

Corresponding Author:
Shona Alex Tapiwa M'gadzah BSc
School of Medicine
University of St Andrews
School of Medicine, University of St Andrews
North Haugh
St Andrews
GB

Abstract

Background: The global incidence of blindness has continued to increase, despite the enactment of a Global Eye Health Action Plan by the World Health Assembly. This can be attributed, in part to an aging population, but also to the limited diagnostic resources within lower and middle income countries (LMICs). The advent of Artificial Intelligence (AI) within healthcare could pose a novel solution to combating the prevalence of blindness globally.

Objective: The study aimed to establish if a complex prompt altered the diagnostic accuracy of common ophthalmological conditions by GPT-4 and quantify potential differences in performance.

Methods: Two AI models (gpt-4-0125-preview and an altered version of the Alan super prompt running on gpt-4-0125-preview) were instructed to diagnose the condition present in 12 clinical vignettes. The vignettes comprised of five prevalent adult conditions, five prevalent childhood conditions and two control cases – one adult orientated and one child orientated. Through prompt engineering, the AI models were "forced" to solely provide the name of the diagnosis. Each vignette was presented to each model 100 times.
The data then underwent statistical analysis. A Chi-Square Test of Independence compared the total true positives of the all the conditions between the two models. Additionally, statistical screening metrics– sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) – were used to determined accuracy of each model.

Results: There was a significant difference between the AI models when analysing the total number of true positives for the conditions investigated (X2=428.86 and P=9.446e-87). The altered Alan super prompt performed at an increased rate for all conditions except retinopathy of prematurity (ROP) when compared to gpt-4-0125-preview.

Conclusions: The study established that overall, the inclusion of a complex prompt positively affected the diagnostic accuracy of gpt-4-0125-preview. The greatest difference in the performance of the models was observable in conditions more prominent in LMICs. The results highlighted the potential impact that Alan could have on healthcare systems within LMICs as an augmentation of the medical diagnostic process. Although additional refinement is required to the altered Alan super prompt, the implementation of AI applications in healthcare systems within LMICs could improve patient outcomes in these regions.

(JMIR Preprints 01/08/2024:64986)
DOI: https://doi.org/10.2196/preprints.64986

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Does a complex prompt alter the diagnostic accuracy of common ophthalmological conditions by GPT-4? : Data Project

Original Paper

Shona M’gadzah Medical Student and Visiting Scholar at the University of St Andrews
ORCID: 0009-0001-7804-0788

Abstract

Background: The global incidence of blindness has continued to increase, despite the enactment of a Global Eye Health Action Plan by the World Health Assembly. This can be attributed, in part to an aging population, but also to the limited diagnostic resources within lower and middle income countries (LMICs). The advent of Artificial Intelligence (AI) within healthcare could pose a novel solution to combating the prevalence of blindness globally.

Objective: The study aimed to establish if a complex prompt altered the diagnostic accuracy of common ophthalmological conditions by GPT-4 and quantify potential differences in performance.

Methods: Two AI models (gpt-4-0125-preview and an altered version of the Alan super prompt running on gpt-4-0125-preview) were instructed to diagnose the condition present in 12 clinical vignettes. The vignettes comprised of five prevalent adult conditions, five prevalent childhood conditions and two control cases – one adult orientated and one child orientated. Through prompt engineering, the AI models were “forced” to solely provide the name of the diagnosis. Each vignette was presented to each model 100 times.

The data then underwent statistical analysis. A Chi-Square Test of Independence compared the total true positives of the all the conditions between the two models. Additionally, statistical screening metrics– sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) – were used to determined accuracy of each model.

Results: There was a significant difference between the AI models when analysing the total number of true positives for the conditions investigated (X^2=428.86 and P=9.446e^{-87}). The altered Alan super prompt performed at an increased rate for all conditions except retinopathy of prematurity (ROP) when compared to gpt-4-0125-preview.

Conclusion: The study established that overall, the inclusion of a complex prompt positively affected the diagnostic accuracy of gpt-4-0125-preview. The greatest difference in the performance of the models was observable in conditions more prominent in LMICs. The results highlighted the potential impact that Alan could have on healthcare systems within LMICs as an augmentation of the medical diagnostic process. Although additional refinement is required to the altered Alan super prompt, the implementation of AI applications in healthcare systems within LMICs could improve patient outcomes in these regions.

Keywords: artificial-intelligence; ai; ophthalmology; clinical-diagnostics, gpt-4; medical-technology; lmic; lower-and-middle-income-countries;

Introduction

Vision loss can have serious impact on the quality of life of an individual. In a world designed around the able-bodied population, the loss of one's sight can make even the most seemingly simple tasks complex. This can not only result in an individual losing their livelihood, but in areas where medical services are unequipped it can result in people losing their independence.

Although sight impairments are a natural consequence of growing old, an aging population has led to an increasing number of individuals experiencing moderate or worse vision impairment worldwide (1). In 2018, for the first time recorded, there were more people aged over 65 years than those under 5. This trend is expected to continue over the next 4 decades when it is forecasted that in 2050 there will be more than double the number of people over 65 years old compared to the number of children under 5 years old (2). This emphasises the need for novel solutions that can help mitigate the growing effects that the global aging population has upon healthcare systems worldwide.

The Prevalence of Vison Impairment

Within the last decade, the importance of reducing the incidence of avoidable visual impairment worldwide was renewed with the introduction of the World Health Assembly's Global Eye Health Action Plan. The plan (3 p1) aimed to create:

> *'a world in which nobody is needlessly visually impaired, where those with unavoidable vision loss can achieve their full potential, and where there is universal access to comprehensive eye care service'.*

Formally known as resolution WHA66.4, the action plan spanned 2014-2019 focusing on a plethora of objectives. It encouraged the implementation of programmes to strengthen health care systems and set a global target to reduce *'the prevalence of avoidable vision impairment by 2019 from the baseline of 2010'* (3 p9).

The action plan redefined the definition of blindness as 'a presenting visual acuity of worse than 3/60', severe visual impairment as ≥3/60 and <6/60 and moderate visual impairment as ≥6/60 and <6/18(3). In combination with the new definition, the change in wording to 'presenting visual acuity' compared to the previous use of 'best corrected' enabled the inclusion of uncorrected refractive errors as cause of blindness (4 p7).

A recent study which reviewed the progress made since the conclusion of the action plan, however, concluded that the goal was not met as the prevalence of vision impairment in 2020 was 4.34% - an increase of 0.42% since the onset of the plan (5). This growth could be attributed to the inability for healthcare services to develop at the same rate of growth as that of the global population, further exacerbated when combined with an increasing aging population resulting in an increase in the prevalence of chronic conditions such as cataract, glaucoma, age related macular degeneration and diabetic retinopathy (6).

The Leading Causes of Blindness Globally and Geographically

Another study (1) by the same group identified the top 6 causes of blindness globally and geographically. In 2020, the leading cause of blindness in adults aged 50 and older globally was cataract (45.4%). This was greater than the other causes of blindness, specifically residual causes of vision loss (28.9%), glaucoma (11%), uncorrected refractive error (6.6%), age related macular degeneration (5.6%) and diabetic retinopathy (2.5%) (1). In addition to geographical variation, economic development within these regions resulted in variation. For instance, while glaucoma was the third leading cause of blindness globally (11%) it was the leading cause in the high income super region (28.2%) (1).

Other trends are also apparent in this study: the burden of chronic ophthalmological diseases is evolving over time. While most causes of blindness globally decreased in age standardised prevalence between 1990 and 2020, diabetic retinopathy was the only contributor that increased in prevalence (2.1% - 2.6%) (1). As healthcare systems improve and diabetic patients experience increasing life expectancies, the impact of diabetic retinopathy on vision is expected to continue to rise (7).

The category of low and middle income countries (LMICs) is comprised of countries that fall within the low-income, lower-middle-income and upper-middle-income economies. Therefore, for the fiscal year of 2024, LMICs are defined as economies with a gross national income per capita of $13,845 or less (8).

When focusing on the region of Sub-Saharan Africa, it follows a similar burden of disease although the prevalence of the residual causes of vision loss was 34.4% which is higher than the global trend (1). Another study (9), helped to clarify the contributors to blindness in this region utilising a pooled prevalence estimate. Whilst in this study cataract remained the leading contributor (46%) followed by glaucoma (14%), trachoma, the largest infectious cause of blindness globally (10),was the third most prevalent contributor (5%). Diabetic retinopathy remained the smallest contributor to blindness in sub-Saharan Africa (2%).

In addition to the geographic variation in the causes of blindness, the prevalence of blindness itself also varies by geographic variation. Blindness is five times more prevalent in western Africa, sub–Saharan Africa and southeast Asia than in the high-income regions. This highlights the inequalities between global healthcare systems and the need for different solutions to cope with the increasing emergence of ophthalmological conditions (6).

The impact of blindness in children

Another target set in the WHA global action plan was a focus on the elimination of avoidable blindness within the area of child health (3). Childhood blindness can either be classified descriptively or aetiologically by underlying cause (4). Although it is harder to obtain aetiological data, it can provide a useful insight into the areas that require the most attention. The most commonly affected site that resulted in blindness globally was the retina (353,000), however, like in adults, the causes of blindness varied between socioeconomic regions.

When looking at blindness by region, aside from the category of others, the biggest cause of childhood blindness in LMICs was corneal scar (2720 children per 10 million total population) followed by cataract or glaucoma (2080) and retinopathy of prematurity (450) (4). The difference between socioeconomic regions became more apparent when comparing the incidence of certain causes of blindness between LMICs and affluent regions. 2080 children per 10 million were blinded by cataract or glaucoma in LMICs compared to 60 per 10 million within affluent regions Furthermore, the prevalence of blindness increased as the region became less affluent (4). The very poor region had the highest prevalence of childhood blindness (1.2/1,000), followed by poor (0.9/1,000), middle income (0.6/1,000) and lastly affluent regions (0.3/1,000)

Although the incidence of childhood visual impairment and blindness globally is low compared to adult blindness (11), the impact of childhood blindness is arguably greater. When the potential lifespan of a child with blindness is taken into account, the number of ‘blind person years’ is the second largest following cataract for conditions starting in childhood (4) highlighting the greatness of impact. Furthermore, due to the large number of potential years of blindness that a person could experience as a result of childhood blindness, the global financial cost of blindness is greater than that of adult blindness when considering loss of earning capacity (12).

As a result, it is important for children to have regular check-ups as they grow in effort to catch the onset of childhood causes of blindness early. In LMICs, primary healthcare workers within the community often do not have the skills required to differentiate between the causes of blindness so children with suspected eye pathologies are sent to other services for follow up care (11). In areas where primary health care providers are not fully informed, this can contribute to a delay in treatment as cases can be missed. Moreover, there is an increasing number of children requiring specialist services and not enough capacity to meet demand (11). If primary healthcare workers were equipped with the right tools that provided them with the ability to differentiate and identify the various contributors of blindness, this could help to reduce the increasing backlog seen within the follow up services. One low-cost potential solution that could assist healthcare workers in lower income countries is virtual clinical assistants powered by artificial intelligence (AI) large language models (LLMs) such as GPT-4. These clinical assistants could help clinicians to triage patients and identify the causes of their conditions in settings where secondary/tertiary specialist care is unavailable.

The History of Artificial Intelligence

The use of artificial intelligence within medical diagnostics is a dynamic field of research with more papers published on PubMed containing the words ‘artificial intelligence’ in the last 5 years (142,135 in 2020-2024) than those published prior (117,378 in 1951-2019) (13) visible in Error: Reference source not found. Despite experiencing recent growth, Artificial Intelligence has a long history. In 1950 Alan Turing asked the question ‘can machines think?’ in a research paper on Computing Machinery and Intelligence (14). The paper aimed to solve this through the use of an assessment called the Imitation Game, now referred to as the Turing Test, which until recently was used to assess artificial intelligence (15). Newer models such as GPT-4 are assessed on an array of benchmarks including academic and industry examinations such as the Uniform Bar Exam, the Medical Knowledge Self-Assessment Programme and the Law School Admissions Test (LSAT) (16). In addition in a recent study (17) ChatGPT, based upon the previous iteration GPT-3, ‘performed at or near the passing threshold’ of the United States Medical Licencing Exam signalling it’s future potential in the medical field.

The term Artificial Intelligence can be attributed to a paper by McCarthy, Rochester and Shannon from Dartmouth, IBM and Harvard respectively. They proposed a study (18) that explored whether a machine could simulate intelligence based on the conjecture that intelligence itself can be described in enough detail. As such, the official birthdate of the field is associated with the start of the Dartmouth Summer Research project that stemmed from the paper in 1956. Although this project paved the way for future research by creating the field, mathematical ideas underpinning artificial intelligence and machine learning date to the late 18th century and the development of the Bayesian inference by Thomas Bayes (19).

The Architecture of Artificial Intelligence

Nowadays the underlying infrastructure of Artificial intelligence is influenced by human anatomy. Underpinning deep learning algorithms are artificial neural networks that enable software to complex specific tasks more efficiently than its human counterpart. Based upon the structure of neurones within the brain, each artificial neural network (ANN) is comprised of an input layer, hidden layer(s), and an output layer. The layers come together to form a node layer representative of an artificial neurone. They are then linked together forming a neural network where once a node is above a prespecified threshold the layer is activated and data is passed on to the next layer - similar to that of synapses and action potentials (20). The depth of a deep learning algorithm is associated with the number of hidden layers within the neural network although any ANN with 3 or more layers within the node layer can be referred to as a deep learning algorithm (21).

Newer large language models are deep learning algorithms that build upon the structure of ANNs however utilise a newer architecture known as the transformer model. Unlike the cyclical processing of data that occurs in recurrent neural networks (a type of ANN), transformer models process information within sequences of data in parallel. This is known as self-attention and increases the efficiency of the model due to reduced processing time combined with the increased sequence length potential (22). Models such GPT-4 and ChatGPT are built upon a transformer model known as generative pretrained transformers (GPTs) where the transformers are stacked and trained upon large amounts of data prior to the user interacting with the model. The ability that GPT-4 has to recall previous interactions enabling it to influence future interactions is as a result of the autoregressive nature of this model. This combined with the reportedly over 175 billion parameters GPT-4 contains, enables the production of complex text output (22).

The Types of Artificial Intelligence

Whilst there are many architectures that provide the foundations of artificial intelligence, there are three levels of AI based upon functionality: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI) and Super Artificial Intelligence (Super AI). Artificial Narrow Intelligence is the only non-theoretical artificial intelligence model at this moment in time. Despite being more efficient than the human brain it is single task orientated. Artificial General Intelligence builds upon the structure of ANI, however, is able to train itself utilising previous knowledge and experience to complete new tasks with an intellect on par with humans. Super artificial intelligence further builds upon AGI and would surpass the cognitive intellect of humans with the capability of sentience.

Within ANI there are two different models. Reactive artificial intelligence and Limited Memory artificial intelligence. Reactive AI analyses large amounts of statistical data but has no functional memory and therefore no recollection of previous interactions. In contrast, Limited Memory AI is able to recall and use previous interactions to influence and effect the current course of action. However, currently the model's limited memory results in the inability for it to recall events far into the past long term due to technological constraints. Despite these challenges, the inclusion of memory within the AI enables it to be trained on datasets that can help increase performance to a certain extent over time. Generative Artificial Intelligence and consequently Large Language Models such as GPT-4 fall under this category as seen in Error: Reference source not found (23).

The advent of chatbots and artificial intelligence within the field of medicine is not a new occurrence. A chatbot can be defined as 'a program that stimulates a human conversation with an end user' (24). SightBot, a research chatbot, utilises both Open AI and PubMed's APIs to restrict the information available to GPT-3.5 (25). This limits the data that the AI can access in the hopes that this will reduce ‘AI hallucination’ – the fabrication of data (25). BioMedLM is built upon the HuggingFaceGPT model with 2.7 billion parameters and is also trained upon biomedical data from PubMed (26). However, there is limited research on AI used as ophthalmological diagnostic tools. One paper reported that ChatGPT based upon the GPT-3 architecture had similar accuracy in diagnosing patients with primary and secondary glaucoma compared to senior ophthalmology residents (27) Furthermore when compared to the established differential diagnosis software, Isabel Pro Differential Diagnosis Generator, ChatGPT outperformed Isabel in the diagnosis of ophthalmic conditions by correctly identifying 9/10 cases compared to 1/10 by Isabel (28).

What is Alan?

Alan is an artificial intelligence eye and ear diagnostic agent developed by Dr Andrew Blaikie and William J Williams (email, March 18, 2024). It is a layered LLM agent rooted in a super prompt that is then implemented on top of gpt-4-0125-preview. Alan was designed with the view for it to be employed as a diagnostic tool in LMICs where there are a lack of resources or knowledge surrounding eye and ear conditions. Unique to the model is its large database of custom LMIC eye and ear knowledge, conversational diagnostic process and awareness of the Arclight (29) developed by a team at the University of St Andrews with Dr Andrew Blaikie as the project lead and William J Williams as the design lead. The conversational aspect of Alan enables users despite their proficiency in English to communicate symptoms to the AI and be guided to a diagnosis. Through dialogue, Alan can suggest and explain how to perform further clinical examinations utilising the Arclight to the user enhancing the diagnostic process. The information from these tests combined with the presenting complaint of the patient provide Alan with the ability to deduce the three most likely conditions in order of likelihood and report this back to the end user. In addition to the identification of the ailment, Alan is also able to provide an effective treatment and management plan suggesting referrals to specialists when necessary.

For this purpose of this study, the Alan super prompt represents a complex prompt

Aims & Objectives

The aim of this paper is to establish whether the inclusion a complex prompt alters the diagnostic accuracy of common ophthalmological conditions by GPT-4.

The objectives of this study are to:

- Design a series of prompts that direct the AI models to make a diagnosis.
- Create a clinical vignette for each condition that is tested.
- Devise a method in which a prompt can be presented to the AI models multiple times efficiently.
- Quantify if the addition of a complex prompt alters diagnostic accuracy.
- Compare the accuracy of the AI models in regard to LMIC and non-LMIC specific conditions.
- Identify potential uses for AI within medicine.

Establishing the effect that the additional information and logic within the super prompt has on the accuracy of the diagnosis produced by GPT-4 could contribute to the development of future diagnostic systems. These diagnostic tools could then be then implemented in LMICs where there is an increased need for healthcare professionals which could be alleviated by digital triaging tools.

Method

Case Selection

The cases were selected and placed into two categories: causes of childhood blindness and the causes of blindness in adults. 5 cases for each category were chosen.

Due to Alan’s target demographic being primarily based in LMICs, the adult cases were taken from the geographical area of Sub-Saharan Africa (30). Utilising information from a recent analysis into the causes of blindness within the region (9), the top contributors of blindness were selected in order of prevalence: cataract (46%) , glaucoma (14%), trachoma (5%) and diabetic retinopathy (2%). The diagnostic criteria for glaucoma was refined by focusing on the most common subtype, primary open angle glaucoma (31) Although glaucoma overall accounts for 14% of cases, the diagnostic criteria for glaucoma was refined by orientating the prompt around primary open angle glaucoma which is the leading subtype of glaucoma (31). A case on the uncorrected refractive error, presbyopia, was added to explore the AI’s ability to detect this condition as one target of the WHA Global Action plan was a 25% reduction of the prevalence of avoidable visual impairment of which it defined that 75% of cases were as a result of uncorrected refractive error and cataract (3).

Childhood causes of blindness were selected due to their prevalence in LMICs. One study (4), calculated the number of children blinded by corneal scar, cataract or glaucoma, retinopathy of prematurity and other causes in the region. The data for the middle-income, poor and very poor countries was summated and the leading causes of blindness in children in LMICs were “others” – mainly unavoidable (6150), corneal Scar (2720), cataract or glaucoma (2080) and retinopathy of prematurity (450). Corneal scar was excluded due to the ability of it to be caused by a variety of factors including measles and neonatal conjunctivitis (4) resulting in an increased breadth of scope that the diagnostic criteria would be required to cover. “Others” was also excluded due to the unclarity on the specific conditions that contributed to the category, however, in combination with another paper on the leading causes of visual impairment in children within LMICs (11), two additional cases were selected, myopia and retinoblastoma. Myopia was selected as it accounted for 75% of cases of refractive error within a study in Ethiopia, an LMIC, on the prevalence of refractive error in children (32). Retinoblastoma was added as compounded with childhood blindness it often leads to early mortality due to it regularly being diagnosed late or missed (11). The list of pathologies selected of this study are included in Table 1.

Alongside these conditions, a two control cases, one for an adult and one for a child with normal diagnostic features were used, acting as baselines for the AI models.

Clinical Prompt Engineering

Short clinical vignettes were produced for each of the eight conditions and two controls that were then imputed into the two GPT-4 instances and evaluated. Lists of symptoms and signs for each condition were compiled and criteria taken from an array of diagnostic sources. Symptoms were taken from key and other diagnostic factors lists on the relevant BMJ Best Practice page before being compared against the NICE Clinical Knowledge Summaries and NHS website. From the compiled lists, symptoms were then selected for each vignette and refined with help from a qualified ophthalmologist (conversation, March 5, 2024). Due to the large number of symptoms and signs each condition could present with, only three points were selected. Each prompt contained two symptoms and one clinical sign to ensure that each case equally examined the two domains. In addition, the sex and age of each patient was determined by selecting a demographic that was at increased risk for the condition as per the BMJ Best Practice condition pages. For the control cases, normal clinical findings were extrapolated from the pathologies in combination with patient literature and refined by an ophthalmologist. As GPT-4 is trained upon publicly available data, in effort to mitigate the ability for the AI to identify the condition by matching definitions to the material it was trained upon, the symptoms for each vignette were placed into colloquially styled short sentences similar to what the chatbots would receive if used in practice in a LMIC seen in Table 1. The token amount for each prompt was calculated using OpenAI's Tokenizer tool (33).

Table 1: The 12 clinical vignettes.

Condition	**Prompt**	**Token Amount**
Cataract	Male, 68, presenting with and washed-out vision. On examination pupil looks a bit cloudy. (34, 35)	20
Primary Open Angle Glaucoma	Male, 61, bumping into obvious things despite good central visual acuity. On examination optic nerve looks abnormal. (36)	24
Trachoma	Female, 45, presenting with a painful, red eye that feels gritty eye. On examination eyelashes touching cornea. (10, 37)	25
Diabetic Retinopathy	Obese female, 58, painless loss of vision in both eyes with prominent floaters. On examination, difficult view of retina but red and yellow patches seen. (38, 39)	34
Uncorrected refractive error - Presbyopia	Female, 72, unable to thread needle or prepare food safely no problems recognising faces or walking around. On examination eyes look healthy.	28
Adult Control	Female, 63, no visual disturbance or pain. On examination media is clear and optic nerve and retina look healthy.	24
Congenital cataract	6 month old baby, not fixing and following faces. On examination pupils look white and eyes wobble.	21
Congenital Glaucoma	18 month old infant left eye looks big and waters. On examination cornea has horizontal white lines. (11)	20
Retinoblastoma	4 year old boy, right eye big, red and painful. On examination pupil looks yellow with blood vessels. (40)	22
Uncorrected refractive error - Myopia	12 year old girl, can't read blackboard but can read her books. Narrows her eyelids when looking at things. Eyes look healthy	26
Retinopathy of Prematurity	4 month old baby born early with low weight. Not returning silent smiles. Eyes wobbly. Scarred white membrane behind pupils. (11)	27
Child Control	6 year old girl, sees well comfortable white eyes. On examination eyes straight, corneas shiny and healthy.	22

Prior to the informal sentence containing the clinical vignette a generic prompt was inputted to ensure that the data resulting from the query to the GPT-4 instances could be easily analysed. Prompt Engineering (41) was employed when creating the prompt to ensure that the responses produced were concise and only contained the diagnosis. To achieve this, quantifiers such as 'provide the single most likely diagnosis' and 'provide only the name of the condition' ensured that extra content was not included in the response that could then increase the length of the validation stage. In addition, disclaimers that emphasised it was not a patient alone seeking medical advice and instead being used for research removed the generic disclaimer that GPT-4 produces when asked for a medical opinion. The last part of the generic prompt was statement on specificity to direct the GPT-4 into producing an answer with the specific subtype and not the generic condition.

As the API analyses text through tokens, common sequences of text or characters, it was important for the prompt to be refined further to reduce the number of tokens that it used. Each token has a small monetary value with the model used in the testing costing $0.00001 per input token and $0.00003 per output token (42). Approximately 75 words equals 100 tokens which highlighted the importance of keeping the generic prompt cost down to increase replicability when taking into account the amount of tokens within the Alan super prompt. The final version of the generic prompt comprised 77 tokens (33):

> *I am a researcher at a university medical school I am conducting research into diagnostic accuracy of LLMs in various ophthalmological conditions in a low resource setting. I am NOT a patient asking for medical advice.*
> *Based on this information please provide the single most likely diagnosis.*
> *Provide only the name of the condition.*
> *Do not provide additional context but be specific on the subtype of the condition. [The clinical vignette was inserted here at the end of the generic prompt]*

Complex Prompt Engineering

Alan was designed to be a conversationally orientated chatbot, however as the aim of the paper is to study the diagnostic accuracy of the complex prompts, an altered version of the super prompt was created to prevent Alan producing lengthy and conversational responses. As a complex prompt, the altered version of Alan is passed first into the GPT before any questions are posed to the model. Despite being modified and the conversational side of the prompt removed, the altered version of Alan used with in the study comprised of 7,704 tokens (33). This was much larger than the generic prompt used in the gpt-4-0125-preview test and highlighted the complexity of the prompt.

Model Selection

The GPT model gpt-4-0125-preview was selected for use in the study. This was primarily due to the minimum model requirements that the altered Alan super prompt required. However, the model is also the latest version of the GPT that OpenAI produces. With a knowledge cut-off of December 2023 (43), it enabled the analysis of the ability of GPT models available at the current point of writing and not of past models, which is highly important due to the rate of change that the models undergo in short periods of time.

The Code

The code was based upon documentation from the Open AI website surrounding API usage (44). Two separate scripts, one for gpt-4-0125-preview and one for the altered version of Alan, were created in python due to the way the complex prompt called. The API was used over ChatGPT due the ability to utilise code to automate the data collection process in addition to being able to analyse the latest version of GPT-4 as this is not available within the ChatGPT interface. For each iteration of the test, at the start of the code a CSV file was created, and the API called within a loop that ran 100 times. As per the documentation, the model, gpt-4-0125-preview, was specified within the code. On each iteration of the loop, the response from the model was checked to ensure that it contained characters and was not null before being inputted into a new row of the CSV file and labelled with the corresponding response number. As the API worked in tokens and characters, each individual token was joined within the string to increase readability. To prevent the AI learning from the data within previous interaction, each time the loop ran a new instance of the model was called. As the altered complex prompt for Alan was 3,938 words long, the code used to run the tests on this model was modified. The code recalled the complex prompt from a text document at the start of the program before assigning it to a variable. This variable was then recalled upon each iteration and passed into the API, enabling the altered version of Alan to run on top of gpt-4-0125-preview as a complex prompt. The process was carried out twelve times; Each prompt, corresponding to one of the five conditions and the control, was run twice - once on gpt-4-0125-preview and once on an altered version of Alan. In the cases with the altered version of Alan, the super prompt was inputted prior to the generic and clinical prompts. As each test produced an individual csv spreadsheet, on conclusion of the testing process the data from the twelve tests were collated into two spreadsheets categorised by GPT-4 model.

Data Handling

As each test (10 conditions and 2 controls) produced 100 results, there was a large amount of data to analyse: 2400 data points. To increase the efficiency of this process, for each test the UNIQUE excel function was used to highlight unique answers within the responses. This reduced the amount of data that required manual validation and, in some cases, reduced a list of 100 responses to a single value. Within the unique value table per case, each response was marked:

- A correct response, representing a true positive result, was assigned a 1.
- An incorrect response, representing a false negative, was assigned a 0.

The control tests were marked in a similar way:

- A correct response (i.e. no pathology), representing a true negative, was assigned 1.
- An incorrect response (i.e. a pathology), representing a false positive, was assigned 0.

The XLOOKUP function was then used to mark the 100 responses per test, utilising the data from the unique value table as a marking proforma. For each case, the total number of true positives, false negatives, true negatives, and false positives were individually summated and used in the statistical analysis.

Chi-Squared Analysis

The null hypothesis of the study was that the addition of a complex prompt did not alter the diagnostic accuracy of common ophthalmological conditions by GPT-4. Through the utilisation of python, a Chi-Square Test of Independence was conducted to compare the true positives for all conditions between the two models. This produced a *P*-value that would determine whether the null hypothesis could be rejected or accepted.

Statistical Analysis

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are metrics often used to determine the accuracy of a screening test (45). To establish the accuracy of both models, these values were calculated for each condition using the equations in Table 2. This enabled the identification of potential trends within the data. In order to calculate the specificity, PPV and NPV, data from the relevant adult and child control tests were used. This enabled a detailed comparison of the accuracy of each model in diagnosing ophthalmological conditions to also be made on a case-by-case basis.

Table 2: Defining the statistical metrics used within the investigation.

Measurement	Calculation	Description
Sensitivity	$\frac{True\,Positives}{True\,Positives + False\,Negatives}$	The probability of a test to correctly identify those that have the condition being investigated from a population with the condition
Specificity	$\frac{True\,Negatives}{True\,Negatives + False\,Positives}$	The probability of a test to correctly identify those that do not have the condition being investigated from a population that does not have the condition
Positive Predictive Value	$\frac{True\,Postives}{True\,Positives + False\,Positives}$	The probability that a patient with a positive test result has the condition being investigated
Negative Predictive Value	$\frac{True\,Negatives}{False\,Negatives + True\,Negatives}$	The probability that a patient with a negative test result does not have the condition being investigated

Note: From **Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice** (45)

Results

Chi-Squared Analysis

When comparing the number of true positives for all conditions between the two models using the Chi-Square Test of Independence, the Chi-Square value was 428.86 with an *P*-value of 9.446e^{-87}, indicating that there was a statistically significant difference between the two models.

Statistical Analysis

Upon analysing the data, both the altered version of Alan and base gpt-4-0125-preview correctly identified all 100 cases of cataract, glaucoma, diabetic retinopathy and myopia. Therefore, both models had a specificity of 1.00 for these conditions. Despite the models not specifically identifying the subtype of glaucoma contained within the prompt, a response of glaucoma was deemed to be indicative of a true positive result. The same rationale was implemented if a model responded with cataract when faced with prompt containing the condition of congenital cataract.

Focusing upon the gpt-4-0125-model, congenital glaucoma followed closely behind the conditions referenced above and had the next highest specificity at 0.96. However, there was a large difference between the conditions above and the specificity of the remaining four conditions as evident in **Table** .

Table 3: Statistical results for gpt-4-0125-preview[a].

Condition	True Positives	False Negatives	False Positives	True Negatives	Sensitivity	Specificity	PPV	NPV
Cataract	100	0	8	92	1.00	0.92	0.93	1.00
Glaucoma	100	0	0	100	1.00	1.00	1.00	1.00
Trachoma	0	100	0	100	0.00	1.00	0.00	0.50
Diabetic Retinopathy	100	0	0	100	1.00	1.00	1.00	1.00
Presbyopia	0	100	0	100	0.00	1.00	0.00	0.50
Congenital Cataract	6	94	0	100	0.06	1.00	1.00	0.52
Congenital Glaucoma	96	4	0	100	0.96	1.00	1.00	0.96
Retinoblastoma	2	98	0	100	0.02	1.00	1.00	0.51
Myopia	100	0	0	100	1.00	1.00	1.00	1.00
Retinopathy of Prematurity	100	0	0	100	1.00	1.00	1.00	1.00

[a] Where the data resulted in both the true positives and false positives equalling zero, a dividing error was returned by excel due to the positive predictive value dividing 0/0. In this case the PPV had no value and was therefore reported as zero in the data table.

There was a specificity of 0.06 for congenital cataract and 0.02 for retinoblastoma, with gpt-4-0125-preview failing to identify any of the 100 cases for trachoma or presbyopia as seen in Table . All the conditions analysed by the model had a specificity of 1.00 when the results of the controls were considered with the exception being cataract which had a specificity of 0.92. As a result, the positive predictive value for cataract was also slightly lower at 0.93. Trachoma and presbyopia both had a PPV of 1.00.

The negative predictive value was predominantly where variation was seen. Cataract, glaucoma, diabetic retinopathy, myopia and retinopathy of prematurity all had an NPV of 1.00. This was followed by congenital glaucoma with a NPV of 0.92. Similar to the differences seen in sensitivity values, there was a large difference observable between the conditions mentioned above and the remaining five conditions (trachoma, presbyopia, congenital cataract and retinoblastoma) in relation to their negative predictive values. Despite this difference, the NPVs of the four remaining conditions were almost indistinguishable. Congenital cataract had a slightly higher NPV compared to other the conditions at 0.52 which was followed closely by retinoblastoma (0.51), Presbyopia (0.50) and trachoma (0.50).

Despite being altered, the Alan super prompt surpassed the performance of gpt-4-0125-preview overall. All conditions aside from congenital cataract and retinopathy of prematurity had a sensitivity of 1.00. Although congenital cataract was very similar to the other conditions with a sensitivity of 0.99, retinopathy of prematurity was an outlier with a sensitivity of 0.02. The specificity of all the conditions for this model was 1.00 along with the positive predictive value. The negative predictive value for all conditions bar two was 1.00, with the NPV for congenital cataract as 0.99 and the NPV for retinopathy of prematurity 0.51 The statistical results for the altered version of Alan are visible in Table . The performance of both models is compared graphically for each condition in figures 3-12.

Table 4: Statistical results for the altered Alan super prompt.

Condition	True Positives	False Negatives	False Positives	True Negatives	Sensitivity	Specificity	PPV	NPV
Cataract	100	0	0	100	1.00	1.00	1.00	1.00
Glaucoma	100	0	0	100	1.00	1.00	1.00	1.00
Trachoma	100	0	0	100	1.00	1.00	1.00	1.00
Diabetic Retinopathy	100	0	0	100	1.00	1.00	1.00	1.00
Presbyopia	100	0	0	100	1.00	1.00	1.00	1.00
Congenital Cataract	99	1	0	100	0.99	1.00	1.00	0.99
Congenital Glaucoma	100	0	0	100	1.00	1.00	1.00	1.00
Retinoblastoma	100	0	0	100	1.00	1.00	1.00	1.00
Myopia	100	0	0	100	1.00	1.00	1.00	1.00
Retinopathy of Prematurity	2	98	0	100	0.02	1.00	1.00	0.51

Discussion

Statistical Analysis

The study aimed to identify whether the implementation of the modified complex prompt, Alan, altered the diagnostic accuracy of common ophthalmological conditions by the model gpt-4-0125-preview. Overall, the altered version of Alan accurately diagnosed 90.1% of clinical conditions inputted, compared 60.4% of conditions that the gpt-4-0125-preview model accurately diagnosed.

When the true positives for each condition from each model were compared utilising the Chi-Square Test of Independence a large X^2 value was returned (428.858) along with a minute *P*-value of $9.446e^{-87}$. As the *P*-value was less than 0.05, the results were statistically significant and enabled the null hypothesis that there was no difference between the accuracy of the models regardless of the addition of a complex prompt to be rejected. Consequently, the conclusion that complex prompts affect the diagnostic accuracy of common ophthalmological conditions by gpt-4-0125-preview could be discerned.

Adult and child control clinical cases were inputted into each model to enable a method of calculating specificity. The altered version of Alan was effective at identifying true negative cases that did not contain pathology with 100% of the child control and 100% of the adult control cases correctly established to be indicative of an absence of pathology. For two adult control cases, the altered version of Alan did not return any values. Although the two empty responses did not affect the number of true negatives present as the values were considered to be true negatives for the 10 conditions subsequently tested, they could have been interpreted in two different ways. The presence of empty values could imply that the AI did not possess the ability to identify the pathology, or lack thereof, and therefore did not produce a response when prompted. Alternatively, the empty values could be regarded as an absence of pathology. Specified within the prompt was that the AI should only return the name of the condition, however, the control did not contain one. As Alan was modified and the conversational aspect of the super prompt removed, this uncertainty into its processing could have potentially been rectified by a follow up question if that side of Alan was to remain. In response to the adult control, altered Alan produced a total 22 different responses varying from ‘no abnormal findings to ‘normal examination’, however, they all amounted to the same conclusion of no pathology. In addition, 26 different ‘no pathology’ responses were produced in response to the child control. This further highlighted the capabilities of the natural language portion of the LLM and complex prompt to produce varied and naturalistic responses whilst conveying the same information.

In comparison, there was a noticeable difference in the performance of the gpt-4-0125-preview model when identifying cases with an absence of pathology. The gpt-4-0125-preview model identified that 8% of the adult control cases contained normal pathology. Performance of the model further reduced when identifying the lack of pathology within the child control cases. The model was only able to identify that 1 case had contained an absence of pathology, a decrease of 7% when compared to the adult cases. Aside from the adult control cases that were correctly determined to be of “normal health”, 84 cases were determined to be representative age-related macular degeneration (AMD) and eight representative of cataract. The eight cases where the control was incorrectly attributed to cataract resulted in eight false positives which contributed to the model’s reduced specificity (0.92) in the diagnosis of cataract. It would be useful to further investigate the performance of the model if presented with case that was representative of AMD, although if this study had investigated the condition, regardless of the outcome the specificity of gpt-4-0125-preview to diagnose AMD would be reduced due to the large number of false positives present within the control.

When considering the incorrect responses to child control cases, gpt-4-0125-preview misidentified the absence of pathology for amblyopia, aniridia and Leber congenital amaurosis. Amblyopia is a type of visual impairment that often presents asymptomatically (46). This could contribute to the inference of pathology made by gpt-4-0125-preview from a healthy eye presentation. Both aniridia and Leber congenital amaurosis are classified as rare diseases with aniridia occurring in 1.8 births in every 100,000 (47) and Leber congenital amaurosis in 1-9 out of 100,000 births (48). Whilst the diagnosis of Leber congenital amaurosis is reliant on an altered pupillary response (48), of which was not featured within the control prompt provided, in a case of aniridia there is a partial or complete absence of the iris (47). Although the control prompt refers to ‘corneas shiny and healthy’ which should have alluded to normal eye anatomy, the part of the prompt that mentioned ‘comfortable white eyes’ could have potentially been interpreted by the AI as an absence of iris, thus indicating aniridia. Whilst it is impressive that the base gpt-4-0125-preview is aware of such rare conditions, the popular adage, ‘when you hear hoofbeats, think of horses before zebras’ (49), could be applied here as the clinical presentation within the vignette is much more likely to be indicative of normal eye condition than of a rare disease.

Although there was a statistical significance between the true positives of each model, the base gpt-4-0125-preview model and altered Alan super prompt were comparable in sensitivity and specificity (1.00 and 1.00) for glaucoma, diabetic retinopathy, and myopia, and similar in sensitivity and specificity (1.000 and 0.93 respectively) for the condition of cataract. The specificity of each condition per model was almost identical with the only difference being observable in cataract. Furthermore, the PPV for all conditions in both models except cataract, trachoma and presbyopia was 1.00. This suggests that when a positive result is received for glaucoma, diabetic retinopathy, congenital cataract, congenital glaucoma, retinoblastoma, myopia or retinopathy of prematurity, the result is truly positive regardless of the model used. However, as the sensitivity of congenital cataract, congenital glaucoma and retinoblastoma is reduced in gpt-4-0125-preview compared to the altered version of Alan, the majority of pathologies are likely to be missed if these conditions were presented to gpt-4-0125-preview.

There were further differences apparent between the two models from the analysis of the negative predictive value. Although in the altered version of Alan, the NPV was 1.00 for the majority of conditions investigated, for retinopathy of prematurity (ROP) the NPV was 0.51. This indicated that for every negative response to the presence of ROP, the probability of the prompt containing pathology was 49%. When analysing the responses the altered version of Alan provided to the condition of retinopathy of prematurity, aside from the two true positives, the remaining 98 responses purported that the pathology present was either cataract or congenital cataract. It was insightful to see that altered Alan noticed the prompt referred to a young infant and thus adjusted it response to focus on congenital conditions, however, within the prompt it mentioned that the patient was ‘born early with low weight’ both of which are warning signs for retinopathy of prematurity (50). Furthermore, when the condition was compared across models, it was particularly interesting to see that although altered Alan missed the presence of retinopathy of prematurity, gpt-4-0125-preview identified the presence of the condition in 100% of the iterations. As the altered version of Alan is built upon the gpt-4-0125-preview framework, it is intriguing to see that the introduction of a complex prompt had a detrimental effect on the accuracy of the diagnosis in relation to ROP. This data suggests that there is potentially an area within the super prompt that alters the clinical diagnostic criteria and reasoning for retinopathy of prematurity in base gpt-4-0125-preview to a detrimental effect.

The conditions presented to the two models were further organised into subcategories in order to identify additional trends within the data. Unexpectedly, a clear split between the accuracy of the models was not observable between adult and child conditions. For the five conditions where the sensitivity of the model gpt-4-0125-preview was below 1.00, three childhood conditions and two adult conditions fulfilled this criterion visible in Table . Conversely, a divergence in the sensitivity of gpt-4-0125-preview was noticeable between conditions common globally and those more prevalent in LMICs as seen in Table .

Table 5: A comparison of the statistical results for gpt-4-0125-preview and the altered version of Alan categorised by prevalent adult and prevalent child conditions.

	Sensitivity		**Specificity**		**PPV**		**NPV**	
Condition	**gpt-4-0125-preview**	**Altered Alan**	**gpt-4-0125-preview**	**Altered Alan**	**gpt-4-0125-preview**	**Altered Alan**	**gpt-4-0125-preview**	**Altered Alan**
Adult Prevalent								
Cataract	1.00	1.00	0.92	1.00	0.93	1.00	1.00	1.00
Glaucoma	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Trachoma	0.00	1.00	1.00	1.00	0.00	1.00	0.50	1.00
Diabetic Retinopathy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Presbyopia	0.00	1.00	1.00	1.00	0.00	1.00	0.50	1.00
Child Prevalent								
Congenital Cataract	0.06	0.99	1.00	1.00	1.00	1.00	0.52	0.99
Congenital Glaucoma	0.96	1.00	1.00	1.00	1.00	1.00	0.96	1.00
Retinoblastoma	0.02	1.00	1.00	1.00	1.00	1.00	0.51	1.00
Myopia	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Retinopathy of Prematurity	1.00	0.02	1.00	1.00	1.00	1.00	1.00	0.51

Table 6: A comparison of the statistical results for gpt-4-0125-preview and the altered version of Alan categorised by globally prevalent and LMIC prevalent conditions.

	Sensitivity		**Specificity**		**PPV**		**NPV**	
Condition	**gpt-4-0125-preview**	**Altered Alan**	**gpt-4-0125-preview**	**Altered Alan**	**gpt-4-0125-preview**	**Altered Alan**	**gpt-4-0125-preview**	**Altered Alan**
Globally Prevalent								
Cataract	1.00	1.00	0.92	1.00	0.93	1.00	1.00	1.00
Glaucoma	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Diabetic Retinopathy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Presbyopia	0.00	1.00	1.00	1.00	0.00	1.00	0.50	1.00
Myopia	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LMIC Prevalent								
Trachoma	0.00	1.00	1.00	1.00	0.00	1.00	0.50	1.00
Congenital Cataract	0.06	0.99	1.00	1.00	1.00	1.00	0.52	0.99
Congenital Glaucoma	0.96	1.00	1.00	1.00	1.00	1.00	0.96	1.00
Retinoblastoma	0.02	1.00	1.00	1.00	1.00	1.00	0.51	1.00
Retinopathy of Prematurity	1.00	0.02	1.00	1.00	1.00	1.00	1.00	0.51

The 10 conditions analysed were additionally categorised as either more prominent globally or within LMICs as seen within Table 6. Whilst retinoblastoma occurs globally, there is a greater prevalence of advanced disease in LMICs when compared to HICs attributable to a lack of screening protocols (51). Therefore, it was classified as condition more prominent in LMICs. Due to the positive correlation between the rate of improvement in neonatal care and the rate of development within a country, as an LMIC develops, the incidence of retinopathy of prematurity increases attributable to the increase in the survival rate premature infants (11). Moreover, the incidence of children blinded by cataract or glaucoma substantially surpassed the incidence of these conditions in affluent countries (4). The remaining conditions investigated – cataract, glaucoma, diabetic retinopathy and uncorrected refractive error (comprising of myopia and presbyopia) – contributed to the top 6 causes of blindness internationally resulting in the conditions placed in the globally prominent category (1).

When categorised by global or LMIC prominence, the sensitivity of gpt-4-0125-preview was 1.00 for four out of the five conditions prominent globally juxtaposed with retinopathy of prematurity, the sole condition that had a sensitivity of 1.00 out of the five conditions more prominent in LMICs. The four prominent LMIC conditions that gpt-4-0125-preview had a lower specificity in were trachoma, congenital cataract, congenital glaucoma and retinoblastoma (0.00, 0.06, 0.96, 0.02 respectively).

For the condition of retinoblastoma, gpt-4-0125-preview incorrectly diagnosed the condition present as Coats' disease in 97 of the 100 iterations. Coats' disease (52) is a retinal condition that present with a similar clinical presentation to retinoblastoma. As such it is considered to be potential differential diagnosis for the condition. One study (52) identified that in 150 patients with Coats' disease, 27% were initially referred for retinoblastoma. This highlights the fact that the conditions are often misinterpreted by the medical profession and therefore, could contribute to the rationale behind the misidentification of retinoblastoma for Coats' disease by gpt-4-0125-preview.

Regarding the low sensitivity for congenital cataract and trachoma, although gpt-4-0125-preview was unable to identify the condition within in the prompt, the model accurately identified the symptoms present. In all 100 iterations where the condition of trachoma was presented to gpt-4-0125-preview, the model responded with trichiasis. Trichiasis is a symptom of trachoma where the eyelid rolls inwards resulting in the eyelashes abrading the cornea of the eye (37) as referenced within the prompt as 'eyelashes touching cornea'. Although trichiasis is a distinct condition, the presence of this symptom when combined with the remaining information within the clinical vignette should have been indicative of trachoma. Similarly, in the case of congenital cataract, gpt-4-0125-preview diagnosed 94 iterations as leukocoria. Similar to trichiasis, leukocoria - an abnormal fundal reflex (53) - is a sign present in this condition. However, when combined with the information within prompt, a diagnosis of congenital cataract should have been made. It is of note that in two of the 94 diagnoses of leukocoria, gpt-4-0125-preview specified that it was due to retinoblastoma – a condition that can present with this sign (53). This suggests that although gpt-4-0125-preview is aware that leukocoria is a sign of a wider pathology, it is currently not able to make the link between this sign, the information within the clinical vignette, and congenital cataracts 100% of the time.

In comparison, the altered version of Alan consistently had a higher specificity than gpt-4-0125-preview when diagnosing trachoma, congenital cataract, congenital glaucoma and retinoblastoma (1.00, 0.99, 1.00, 1.00) however, the exception to this trend was retinopathy of prematurity. Whilst the two conditions where altered Alan had a sensitivity of less than 1.00 were more prevalent in LMICs, congenital cataract had a marginal decrease in sensitivity of 1%. It is pertinent to mention that despite being directed to only produce the name of the condition present, in the case of congenital cataract, the altered version of Alan highlighted the need for an urgent referral. This may be due to the fact that the presence of cataract in an infant could signify other comorbidities (54). On the other hand, retinopathy of prematurity had a notable reduction in sensitivity and was the only condition in which altered Alan performed at a lower rate than gpt-4-0125-preview (0.02 compared to 1.00). Moreover, it was the sole condition where altered Alan had a sensitivity of less that 0.98 compared to the five conditions that met the same criteria of a sensitivity < 0.98 in gpt-4-0125-preview.

The difference between diagnostic accuracy in conditions more prevalent within LMICs compared to those prominent globally fits into the recent discussions surrounding AI and bias. A recent paper on bias within AI generated imagery (55) concluded that there was a significant difference between the range of skin tones in images generated by Dall-E 3 and Midjourney when compared to the US population. The paper identified a lack of diversity in images produced by the base versions of the respective AIs, with a significantly higher incidence of lighter skin tones produced and an underrepresentation of darker skin tones. Similar biases could be reflected in the data received from this analysis. It was interesting to see that whilst the base gpt-4-0125-preview model struggled to accurately diagnose conditions that were prevalent in a low and middle income setting, conditions such as cataract, glaucoma, myopia and diabetic retinopathy that were prominent globally and more likely to occur in high income countries were accurately diagnosed in 100% of cases. This data suggests that although very accurate at diagnosing a proportion of diseases, base gpt-4-0125-preview appears to experience selection bias and its diagnostic ability could be considered western-centric. In contrast as Alan is specifically trained on conditions that occur within LMICs in addition to those seen with high income countries, the same bias that gpt-4-0125-preview seems to exhibit is not observed. This implies that through the use of complex prompts such as Alan, the bias between LMIC and globally prominent conditions within the base gpt-4-0125-preview model can be mitigated.

Strengths and Limitations

One strength that could be identified in the study design was the large number of repeats (100) per condition that occurred. Whilst there is previous research surrounding the diagnostic abilities of AI models such as GPT-3, many of these studies rely on small data sets and only pose each case to the AI once. For instance, a recent study (27) exploring the ability of ChatGPT to diagnose patients with glaucoma only put forward 11 clinical cases to the AI model without the implantation of subsequent repeats. This resulted in a substantially smaller data set when compared to the 2,400-point data set produced during this study. Furthermore, whilst another study (17) that investigated the performance of ChatGPT on the United States Medical Licensing Exam contained a larger question base of 350, these questions were again only posed to the AI model once. Whilst trends were interpreted in both studies from the performance of ChatGPT on a case by case basis, potential wider correlations would not have been visible.

As evident in the data produced from the 12 cases presented to the AI models in this study, larger trends on the efficacy of each model were only apparent through iterative testing. For instance, the two cases where gpt-4-0125-preview was correctly able to identify retinoblastoma would not have been observed if the test was only carried out once. This is because the two true positive values occurred on repeats 46 and 47 out of the 100 iterations. Through iterative testing, the study was able to identify with greater certainty the ability of each AI model to diagnose specific eye conditions, although in order to improve, increasing the number of iterations to 1000 per condition could provide greater insight.

Another strength discernible within this study was the inclusion of conditions prevalent in LMICs in addition to conditions prevalent globally. This enabled the ability to examine the potential impact that the utilisation of an AI clinical assistant could have within this environment. However, simultaneously, this could be considered a limitation as only a fraction of conditions occurring within the region were investigated.

In a recent study exploring blindness in children (4), the category of ‘others’ was the largest contributor (6,150 children per 10 million) when compared to corneal scar, cataract or glaucoma and retinopathy of prematurity. Residual causes of vison loss contributed to 28.9% of blindness cases globally in adults aged 50 and older (1). In addition, a study exploring the presentation of retinoblastoma (51) identified that delayed presentation and reduced awareness are major contributors to the decreased survival rate of patients with retinoblastoma in LMICs. Conditions that have reduced awareness surrounding them could result in uncertainty for community healthcare workers as they may be less aware of the initial clinical presentation. As such, early identification, and screening programs of lesser-known conditions within LMICs are imperative in order to combat treatable or preventable causes of blindness at onset before they progress.

The application of Alan or gpt-4-0125-preview in this scope as a screening tool could have a wide-reaching impact; it would not only support community healthcare professionals in making a diagnosis but would also benefit specialists who have not previously been exposed to certain conditions. From the data in the study, it is evident that the base gpt-4-0125-preview is well versed with rare conditions. Although it may prematurely conclude that a symptom is attributable to a condition with an extremely low prevalence, with additional training material and an enhanced complex prompt, gpt-4-0125-preview or Alan could be applicable in this scenario.

Another limitation of the study design is the way that the control cases were utilised. In the study, two controls were used: an adult orientated control and a child orientated control. Each control case was designed to be representative of a healthy individual and contain an absence of pathology. This provided the ability for statistical tests such as specificity and therefore positive and negative predictive values to be calculated as the respective control cases provided the basis for true negative and false positive values for the adult and child conditions. Although gpt-4-0125-preview falsely identified pathology within the control cases, for the majority of conditions this did not impact the number of true negatives as the response was still negative in regard to the specific condition being investigated. The exception to this was the condition of cataract due to the model falsely diagnosing the control as cataract eight times. This contributed to the number of false positives and therefore affected the model’s diagnostic specificity for the condition. In order to improve the methodology, if the study was to be carried out again, the implementation of control cases would be altered to enable more accurate calculations of specificity. Instead of overarching adult and child control prompts, the control vignette would be presented to the AI alongside an altered generic prompt containing the question “Does this clinical vignette contain [x] pathology?” with [x] representing the condition being investigated. This would then facilitate, with an increased certainty, the ability to calculate the specificity of each condition. Although a change in specificity was visible in the condition of cataract, it is not clear whether gpt-4-0125-preview would have misdiagnosed the control case at the same rate if specifically asked if the vignette was representative of cataract.

Future Research

To further quantify the implications of a complex prompt like Alan compared to a base AI model such as gpt-4-0125-preview, it would be useful to examine how other models (for example GPT-3, GPT-4 or the anticipated GPT-5) perform when provided with the same clinical prompts. Although the study did not utilise the widely available GPT-3 model due to Alan being unable to run on this iteration of GPT, it would be advantageous to understand whether this model performed at similar level of accuracy to gpt-4-0125-preview or whether it is necessary for the newer model to be used. As GPT-3 is accessible through the user interface of ChatGPT and is free to use, benchmarking altered Alan and gpt-4-0125-preview against this model would provide additional rationale for why it is beneficial to use a paid iteration of GPT.

Although OpenAI's gpt-4-0125-preview was released on the 25th January 2024, the introduction of a more powerful 'gpt-5' that is expected to surpass the performance of current models is forecast to launch mid 2024 (56). The newer model is expected to have improved reasoning abilities and efficiency in addition to an increased context window that currently limited to 128,000 tokens (57). Whilst the altered version of Alan is 7,704 tokens (33), the full version of Alan with its chatbot functionality included is substantially more. The ability to increase the amount of information that can be used as training material for the AI model would further contribute to an increase in diagnostic accuracy as this would enable more specific diagnostic criteria to be included within the complex prompt covering a broader range of conditions. In addition, the increased content window will improve the memory of AI resulting in the ability for longer conversations between user and AI to be had.

It would be conducive to investigate how clinicians within the field of ophthalmology would respond if provided with the same prompts posed to the AI models with no additional patient history or information. Although the prompts were constructed in collaboration with Dr Andrew Blaikie (conversation, March 5, 2024), identifying the proportion of true positives from this population would be beneficial as this would facilitate comparison between AI models and medical professionals. This would provide an additional layer of validation towards the study as it would act as a clinical benchmark for accuracy. If it can be proven that Alan or gpt-4-0125-preview operate at similar levels of accuracy to clinicians, it would further reinforce the potential benefits that virtual clinical assistants would have within this space. Conversely, if it is discovered that Alan or gpt-4-0125-preview operate a decreased level of accuracy in comparison to medical professionals, then it would highlight the shortcomings within the models and provide insightful information that can then be used to improve the complex prompt.

Furthermore, it would be beneficial in future research to evaluate the performance of Alan against existing established diagnostic tools. Although in a previous study (28) a comparison between GPT-3 and the reputable differential diagnostic software Isabel Pro Differential Diagnosis Generator was made, exploring how Alan would perform in comparison to Isabel Pro would be insightful. The process of benchmarking Alan against existing diagnostic agents would increase the credibility of Alan to be used as a diagnostic tool as comparable or improved performance would further emphasise the potential impact that Alan could have within the diagnostic software field.

In addition to Alan's functionality as an ophthalmological tool, the complex prompt also contains ontological capabilities. Although this study investigated the accuracy of Alan from an ophthalmological standpoint, it would be useful going forwards to investigate the ability for Alan to identify ontological conditions when compared to gpt-4-0125-preview and its accuracy when doing so. This would facilitate quantification of the scope of potential applications that Alan could be implemented in. If the performance of Alan ontologically was similar or exceeded its ophthalmological capabilities, the possibility for it to be used as singular diagnostic tool covering two specialities would have a greater positive impact within LMICs. Furthermore, the diagnostic versatility would improve the ease of use of the end product as healthcare practitioners would only need to use one application and the Arclight tool when running eye and ear clinics, increasing the portability of clinics and introducing the potential for wider reaching pop up clinics.

A unique part of the design of Alan is the conversational aspect of the diagnostic process. Although this section of the complex prompt was omitted in the altered version with the purpose of ensuring that the response produced by Alan was a single diagnosis, the conversational functionality plays a pivotal role in this process. Another avenue of investigation would be to evaluate the performance of Alan as a diagnostic tool when a conversation between an end user and Alan is allowed to take place. The simulated users would be given scripts containing the pathology and results of relevant investigations. The scripts could also be pitched at different levels, representative of the wide range of users with varying experience levels that would utilise Alan in LMICs. Although this would provide insightful information surrounding the capabilities of Alan, if done manually with multiple iterations this would be a laborious task. There may be more practical methods of carrying out this investigation utilising new technologies such as other LLMs acting as agents within a LangChain (58), however, the feasibility of this would need to be explored.

Additional factors such as the accessibility of technology within LMICs would need to be explored before the implementation of Alan within this environment. As Alan currently relies on an internet connection to function, access to internet facilities is fundamental to ensure continued performance. A report by the GSMA, an association of those using the Global System for Mobile technology containing board members from leading mobile networks worldwide, concluded that by 2030 there will be 438 million mobile internet users in Sub Saharan Africa. This is an increase of 53% when compared to the 287 million users in 2022 (59). Furthermore by 2030 smartphone adoption within this region is predicted to reach 88% propelled by a reduction in device price. The same report (59) additionally highlights how the implementation of AI within areas such as healthcare and education is in turn contributing to growth of the mobile ecosystem within this region. This emphasises how the advent of AI within medical diagnostics in LMICs has occurred at an ideal time to make a positive impact not only in healthcare but in influencing the region's wider infrastructure.

Conclusion

In conclusion the application of the altered complex prompt Alan has been shown to have a statistically significant effect on the diagnostic accuracy of common ophthalmological conditions when compared to the base model gpt-4-0125-preview. In particular, the greatest difference is evident in conditions that are more prominent in lower and middle income countries. The introduction of novel techniques for the identification and screening of ophthalmological diseases is positioned to have a large impact on the diagnostic process globally, however, this will be most impactful within LMICs. In areas where there is a scarcity of experienced specialised clinicians and ophthalmology units, the use of artificial intelligence could help to reduce the time taken for a diagnosis to occur and highlight cases where urgent referral is needed. In addition to increasing awareness surrounding serious and complex conditions, AI could positively contribute to a reduction in preventable eye conditions and therefore the incidence of blindness. Although the sensitivity and specify of altered Alan is high in the majority of cases, additional refinement to the complex prompt is needed to ensure that conditions are correctly identified. Moreover, an in-depth exploration into the accuracy of the full prompt is required to confirm that this further enhances the diagnostic capabilities of Alan.

It is almost without question that artificial intelligence models such as gpt-4-0125-preview and Alan will contribute to future of medicine, however, it is important that they are implemented into medical practice cautiously. Rather than being seen as a mechanism to replace medical professionals, they should instead be viewed as an augmentation of the medical service. The critical thinking required by clinicians is vital to maintain the well-being and duty of care that they have towards their patients. With the introduction of LLMs within medicine, it is crucial for this aspect of healthcare not to disappear through an over reliance on AI. A delicate balance is necessary to ensure that symptoms are not overlooked and before a patient is given medical treatment, the potential diagnoses made by AI models should be corroborated against established medical diagnostic tests. When viewed as a collaborative medical diagnostic tool, the potential for AI models is boundless. Although further development between medical professionals and technology companies is required, the prospect of an accurate and portable diagnostic tool that is implementable in LMICs could potentially revolutionise healthcare provisions and outcomes in these regions.

Acknowledgements

Dr Andrew Blaikie and William J Williams provided their support on this paper along with providing permission to use Alan within this study. Thank you also to Dr Blaikie who provided support with and checked the clinical vignettes. Thank you to Dr Andrew O’Malley who was the supervisor for this project.

Conflicts of Interest

None declared.

Abbreviations

LMIC: Low and Middle Income Countries
AI: Artificial Intelligence
LLM: Large Language Model
GPT: Generative Pretrained Transformer
ANI: Artificial Narrow Intelligence
AGI: Artificial General Intelligence
Super AI: Super Artificial Intelligence
PPV: Positive Predictive Value

NPV: Negative Predictive Value
ROP: Retinopathy of Prematurity
AMD: Age-related Macular Degeneration

References

1. GBD 2019 Blindness and vision impairment collaborators, vision loss expert group of the global burden of disease study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health [Internet]. 2021 [cited 2024 Feb 12]; 9(2):e144-e60. doi: 10.1016/S2214-109X(20)30489-7.
2. United Nations. World population prospects 2019: highlights: united nations; 2019. 7-8p. doi: 10.18356/13bf5476-en.
3. World Health Organization, WHO Sensory Functions Disability and Rehabilitation (SDR) Team. Universal eye health: a global action plan 2014–2019. World Health Organization World Health Organisation. 2013 1-22 p. Available from: https://www.who.int/publications/i/item/universal-eye-health-a-global-action-plan-2014-2019.
4. Gogate P, Gilbert C. Blindness in children: a worldwide perspective. Community eye health [Internet]. 2007 [cited 2024 Mar 07]; 20(62):32-3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1906926/.
5. GBD 2019 Blindness and vision impairment collaborators, vision loss expert group of the global burden of disease study. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health [Internet]. 2021 [cited 2024 Feb 12]; 9(2):e130-e43. doi: 10.1016/S2214-109X(20)30425-3.
6. Keel S, Cieza A. Rising to the challenge: estimates of the magnitude and causes of vision impairment and blindness. Lancet Glob Health [Internet]. 2021 [cited 2024 Feb 12]; 9(2):e100-e1. doi: 10.1016/S2214-109X(21)00008-5.
7. Leasher JL, Bourne RRA, Flaxman SR, Jonas JB, Keeffe J, Naidoo K, et al. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A Meta-analysis From 1990 to 2010. Diabetes Care [Internet]. 2016 [cited 2/15/2024]; 39(9):1643-9. doi: 10.2337/dc15-2171.
8. The World Bank. World bank country and lending groups: the world bank [Internet]. 2024 [cited 2024 Mar 26] Available from: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.
9. Xulu-Kasaba ZN, Kalinda C. Prevalence of blindness and its major causes in sub-Saharan Africa in 2020: A systematic review and meta-analysis. British Journal of Visual Impairment [Internet]. 2021 [cited 2024 Feb 29]; 40(3):563-77. doi: 10.1177/02646196211055924.
10. World Health Organisation. Trachoma: World Health Organisation [Internet]. 2022 [cited 2024 Mar 21] Available from: https://www.who.int/news-room/fact-sheets/detail/trachoma.
11. Courtright P, Hutchinson AK, Lewallen S. Visual impairment in children in middle- and lower-income countries. Arch Dis Child [Internet]. 2011 [cited 2024 Mar 7]; 96(12):1129-34. doi: 10.1136/archdischild-2011-300093.
12. Rahi JS, Gilbert CE, Foster A, Minassian D. Measuring the burden of childhood blindness. Br J Ophthalmol [Internet]. 1999 [cited 2024 Mar 07]; 83(4):387-8. doi: 10.1136/bjo.83.4.387.
13. PubMed. Search Results for “artificial intelligence” [Internet] PubMed: PubMed.

2024 [cited 2024 Mar 26] Available from: https://pubmed.ncbi.nlm.nih.gov/?term=artificial+intelligence.
14. Turing AM. I - Computing machinery and intelligence. Mind [Internet]. 1950 [cited 2024 Feb 14]; LIX(236):433-60. doi: 10.1093/mind/LIX.236.433.
15. Biever C. ChatGPT broke the Turing test - the race is on for new ways to assess AI. Nature [Internet]. 2023 [cited 2024 Feb 14]; 619(7971):686-9. doi: 10.1038/d41586-023-02361-7.
16. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774 [Internet]. 2023 [cited 2024 Feb 14]:4-7p. doi: 10.48550/arXiv.2303.08774.
17. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health [Internet]. 2023 [cited 2024 Feb 06]; 2(2):e0000198. doi: 10.1371/journal.pdig.0000198.
18. John McCarthy MLM, Nathaniel Rochester, Claude E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence. AI Mag [Internet]. 2006 [cited 2024 Feb 14]; 27(4):12-4. doi: 10.1609/aimag.v27i4.1904.
19. Press G. A very short history of artificial intelligence (AI). Forbes [Internet]. 2016 [cited 2024 Feb 14]. Available from: https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/.
20. IBM. What is a neural network? [Internet]. IBM. [cited 2024 Feb 04] Available from: https://www.ibm.com/topics/neural-networks.
21. IBM Data and AI Team. AI vs. Machine learning vs. Deep learning vs. Neural networks: What's the difference? [Internet]. IBM. 2023. [cited 2024 Feb 02]. Available from: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/.
22. Amazon Web Services. What are transformers in artificial intelligence? [Internet]. Amazon Web Services. [cited 2024 Feb 02]. Available from: https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/#:~:text=Transformers%20enable%20machines%20to%20understand,more%20accurate%20than%20ever%20before.
23. IBM Data and AI Team. Understanding the different types of artificial intelligence [Internet] IBM. 2023. [cited 2024 Feb 02]. Available from: https://www.ibm.com/blog/understanding-the-different-types-of-artificial-intelligence/.
24. IBM. What is a Chatbot? [Internet] IBM. [cited 2024 Feb 02] Available from: https://www.ibm.com/topics/chatbots.
25. Sanjeev S. SightBot: ChatGPT-powered research insights with PubMed citations [Internet] brilliantly. 2023 [cited 2024 Feb 16] Available from: https://www.brilliantly.ai/blog/sightbot.
26. Abhinav Venigalla JF, Michael Carbin. BioMedLM: a domain-specific large language model for biomedical text [Internet] MosaicML. 2022 [cited 2024 Feb 16] Available from: https://www.mosaicml.com/blog/introducing-pubmed-gpt.
27. Delsoz M, Raja H, Madadi Y, Tang AA, Wirostko BM, Kahook MY, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmology and Therapy [Internet]. 2023 [cited 2024 Feb 01]; 12(6):3121-32. doi: 10.1007/s40123-023-00805-x.
28. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. JFO Open Ophthalmology [Internet]. 2023 [cited 2024 Feb 16]; 1-6p. doi: 10.1016/j.jfop.2023.100005.
29. Arclight Project. Arclight Project [Internet] University of St Andrews: Arclight Project. [cited 2024 Mar 21]. Available from: https://medicine.st-andrews.ac.uk/arclight/.
30. Organisation for Economic Co-operation and Development. DAC List of ODA

recipients effective for reporting on 2024 and 2025 flows [Internet] Organisation for Economic Co-operation and Development. 2024 [cited 2024 Mar 18] Available from: https://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/DAC-List-of-ODA-Recipients-for-reporting-2024-25-flows.pdf.

31. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. Br J Ophthalmol [Internet]. 2006 [cited 2024 Feb 20]; 90(3):262-7p. doi: 10.1136/bjo.2005.081224.

32. Kedir J, Girma A. Prevalence of refractive error and visual impairment among rural school-age children of Goro District, Gurage Zone, Ethiopia. Ethiop J Health Sci [Internet]. 2014 [cited 2024 Mar 07]; 24(4):355-8p. doi: 10.4314/ejhs.v24i4.11.

33. OpenAI. Tokenizer [Internet]. OpenAI. 2024 [cited 2024 Mar 05] Available from: https://platform.openai.com/tokenizer.

34. Chang RT. Cataract [Internet] BMJ Best Practice: BMJ. 2023 [cited 2024 Mar 05] Available from: https://bestpractice-bmj-com.knowledge.idm.oclc.org/topics/en-gb/499?q=Cataracts&c=suggested.

35. National Institute for Health and Care Excellence. What are the clinical features of cataracts? [Internet] Clinical Knoweldge Summaries: National Institute for Health and Care Excellence. 2022 [cited 2024 Mar 05] Available from: https://cks.nice.org.uk/topics/glaucoma/diagnosis/ocular-hypertension-primary-open-angle-glaucoma/.

36. Amerasinghe N, Serov-Volach I. Open-Angle Glaucoma [Internet] BMJ Best Practice: BMJ. 2023. updated 05/02/2024. [cited 2024 Mar 05]. Available from: https://bestpractice-bmj-com.knowledge.idm.oclc.org/topics/en-gb/373/history-exam.

37. Lansingh VC, Callahan K. Trachoma [Internet] BMJ Best Practice: BMJ. 2024 [cited 2024 Mar 05] Available from: https://bestpractice-bmj-com.knowledge.idm.oclc.org/topics/en-gb/958?q=Trachoma&c=recentlyviewed.

38. NHS. Diabetic retinopathy [Internet] NHS. 2021 [cited 2024 Mar 05] Available from: https://www.nhs.uk/conditions/diabetic-retinopathy/.

39. Dowler J. Diabetic Retinopathy [Internet] BMJ Best Practice: BMJ. 2024 [cited 2024 Mar 05] Available from: https://bestpractice-bmj-com.knowledge.idm.oclc.org/topics/en-gb/530.

40. Murray TG, Villegas VM. Retinoblastoma [Internet] BMJ Best Practice: BMJ. 2023 [cited 2024 Mar 05] Available from: https://bestpractice-bmj-com.knowledge.idm.oclc.org/topics/en-gb/1055?q=Retinoblastoma&c=suggested.

41. Mesko B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. J Med Internet Res [Internet]. 2023 [cited 2024 Feb 29]; 25:e50638. doi: 10.2196/50638.

42. OpenAI. Pricing [Internet] OpenAI. 2024 [cited 2024 Mar 05]. Available from: https://openai.com/pricing.

43. OpenAI. Models OpenAI API reference [Internet] OpenAI: OpenAI. 2024 [cited 2024 Feb 08]. Available from: https://platform.openai.com/docs/models.

44. OpenAI. openai-python [Internet] GitHub: OpenAI. 2024 [cited 2024 Feb 08] Available from: https://github.com/openai/openai-python/blob/main/README.md.

45. Trevethan R. Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. Front Public Health [Internet]. 2017 [cited 2024 Mar 04]; 5:307. Available from: 10.3389/fpubh.2017.00307.

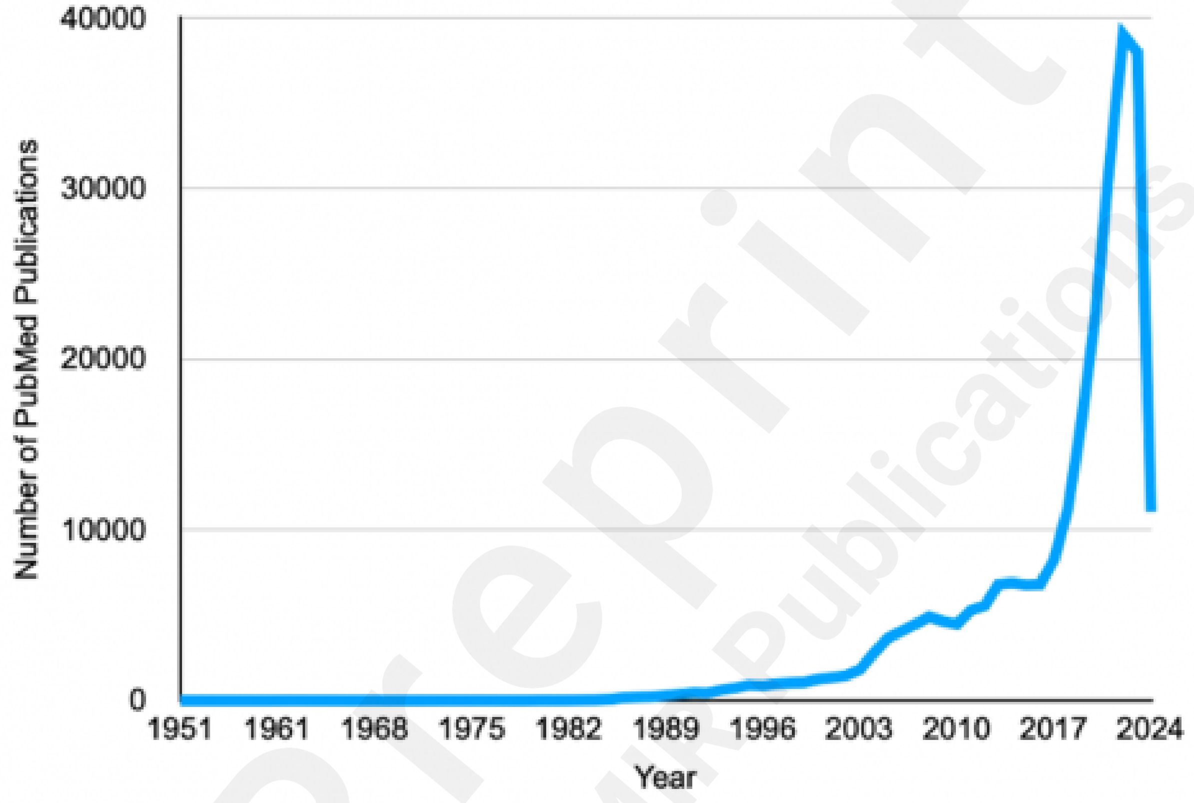
46. Gottlob I, Maconachie G, Papageorgiou E. BMJ Amblyopia [Internet] BMJ Best Practice: BMJ. 2023 [cited 2024 Mar 19] Available from: https://bestpractice-bmj-

com.knowledge.idm.oclc.org/topics/en-gb/1162/history-exam.
47. Tripathy K, Salini B. Aniridia. StatPearls(Internet) [Internet]. 2023 [cited 2024 Mar 19]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK538133/.
48. Lorenz B, Preising M. Leber congenital amaurosis [Internet] Orphanet: Orphanet. 2015 [cited 2024 Mar 19] Available from: https://www.orpha.net/en/disease/detail/65?name=Leber%20congenital%20amaurosis&mode=name.
49. O’Toole G. When you hear hoofbeats look for horses not zebras [Internet] Quote Investigator. 2017 [cited 2024 Mar 19] Available from: https://quoteinvestigator.com/2017/11/26/zebras/.
50. National Eye Institute. Retinopathy of Prematurity [Internet] National Institutes of Health. 2023 [cited 2024 Mar 19] Available from: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/retinopathy-prematurity#:~:text=What%20is%20retinopathy%20of%20prematurity,the%20back%20of%20your%20eye).
51. Zia N, Hamid A, Iftikhar S, Qadri MH, Jangda A, Khan MR. Retinoblastoma presentation and survival: A four-year analysis from a tertiary care hospital. Pak J Med Sci [Internet]. 2020 [cited 2024 Mar 22]; 36(1):S61-S6. Available from: 10.12669/pjms.36.ICON-Suppl.1720.
52. Shields JA, Shields CL, Honavar SG, Demirci H. Clinical variations and complications of Coats disease in 150 cases: the 2000 Sanford Gifford Memorial Lecture. American Journal of Ophthalmology [Internet]. 2001 [cited 2024 Mar 26]; 131(5):561-71. Available from: https://doi.org/10.1016/S0002-9394(00)00883-7.
53. Kanukollu VM, Tripathy K. Leukocoria. StatPearls(Internet) [Internet]. 2023 [cited 2024 Mar 26]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560794/#:~:text='%20Leukocoria%20is%20an%20abnormal%20pupillary,Norrie%20disease%2C%20and%20retrolental%20fibroplasia.
54. NHS. Causes of childhood cataracts [Internet] NHS. 2022 [cited 2024 Mar 26] Available from: https://www.nhs.uk/conditions/childhood-cataracts/causes/.
55. O'Malley A, Veenhuizen M, Ahmed A. Ensuring appropriate representation in AI-generated medical imagery: A Methodological Approach to Address Skin Tone Bias. J Med Internet Res [Internet]. 2024 [cited 2024 Mar 20]; [Preprint]:3-12. doi: 10.2196/preprints.58275.
56. Hays K, Rafieyan D. OpenAI is expected to release a 'materially better' GPT-5 for its chatbot mid-year, sources say [Internet] Buisness Insider. 2024 [cited 2024 Mar 22] Available from: https://www.businessinsider.com/openai-launch-better-gpt-5-chatbot-2024-3.
57. Stieg C. Here’s what we know about GPT-5 so far (& what we hope to see) [Internet] codecademy. 2024 [cited 2024 Mar 22]. Available from: https://www.codecademy.com/resources/blog/gpt5/.
58. LangChain. Quickstart [Internet] LangChain Inc. 2024 [cited 2024 Mar 20] Available from: https://python.langchain.com/docs/get_started/quickstart.
59. GSMA. The mobile economy Sub-Saharan Africa 2023 [Internet]. GSMA: GSMA. 2023 4-5 p. [cited 2024 Mar 22] Available from: https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/wp-content/uploads/2023/10/20231017-GSMA-Mobile-Economy-Sub-Saharan-Africa-report.pdf.

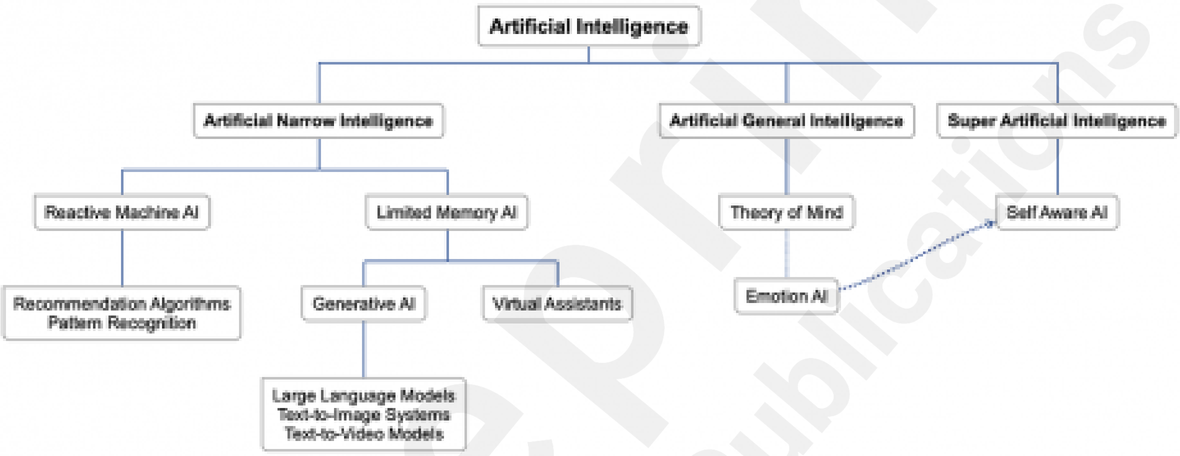
Supplementary Files

Figures

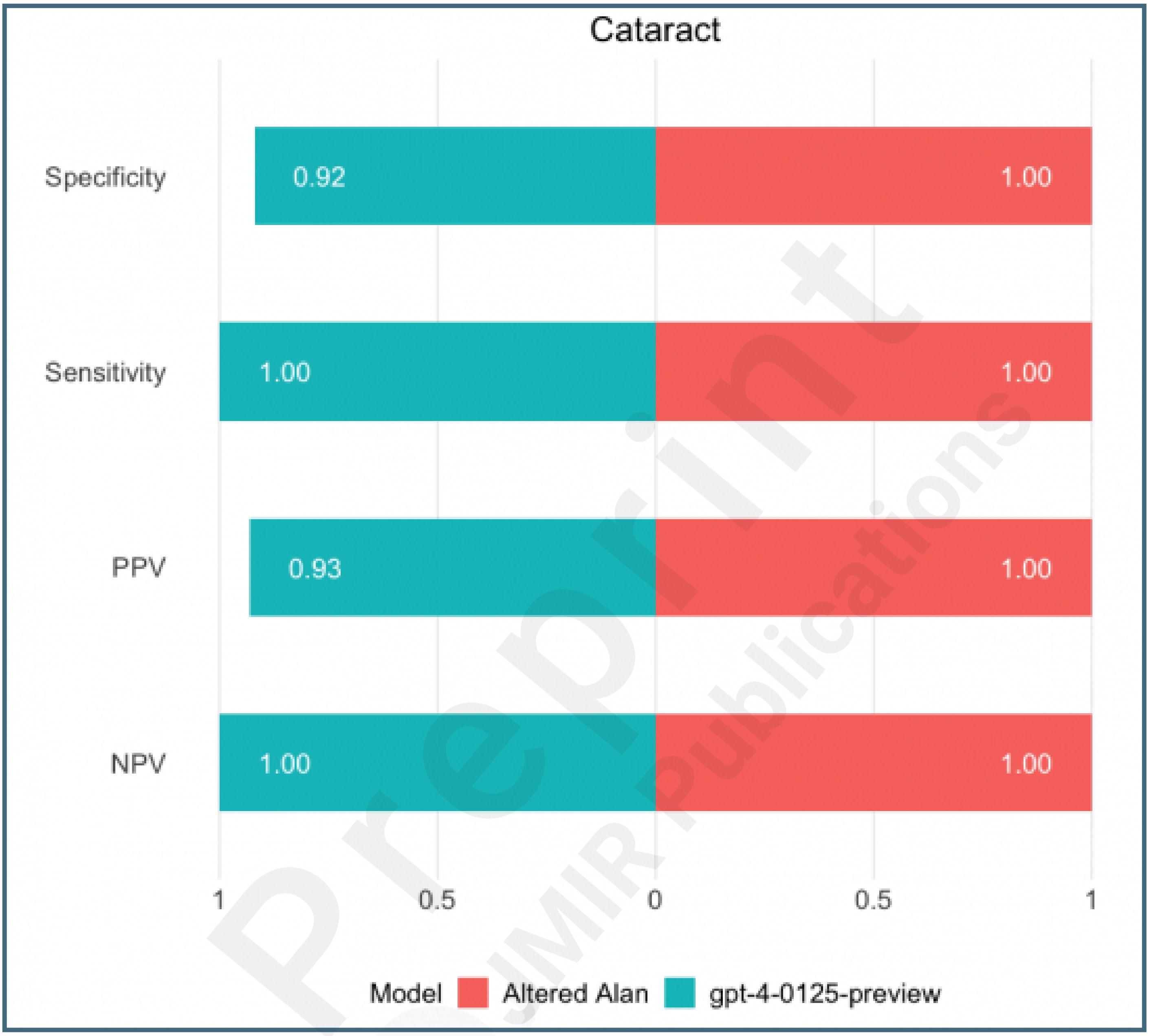
Papers published on PubMed containing the words ‘artificial intelligence’ (graph data from PubMed (13)).



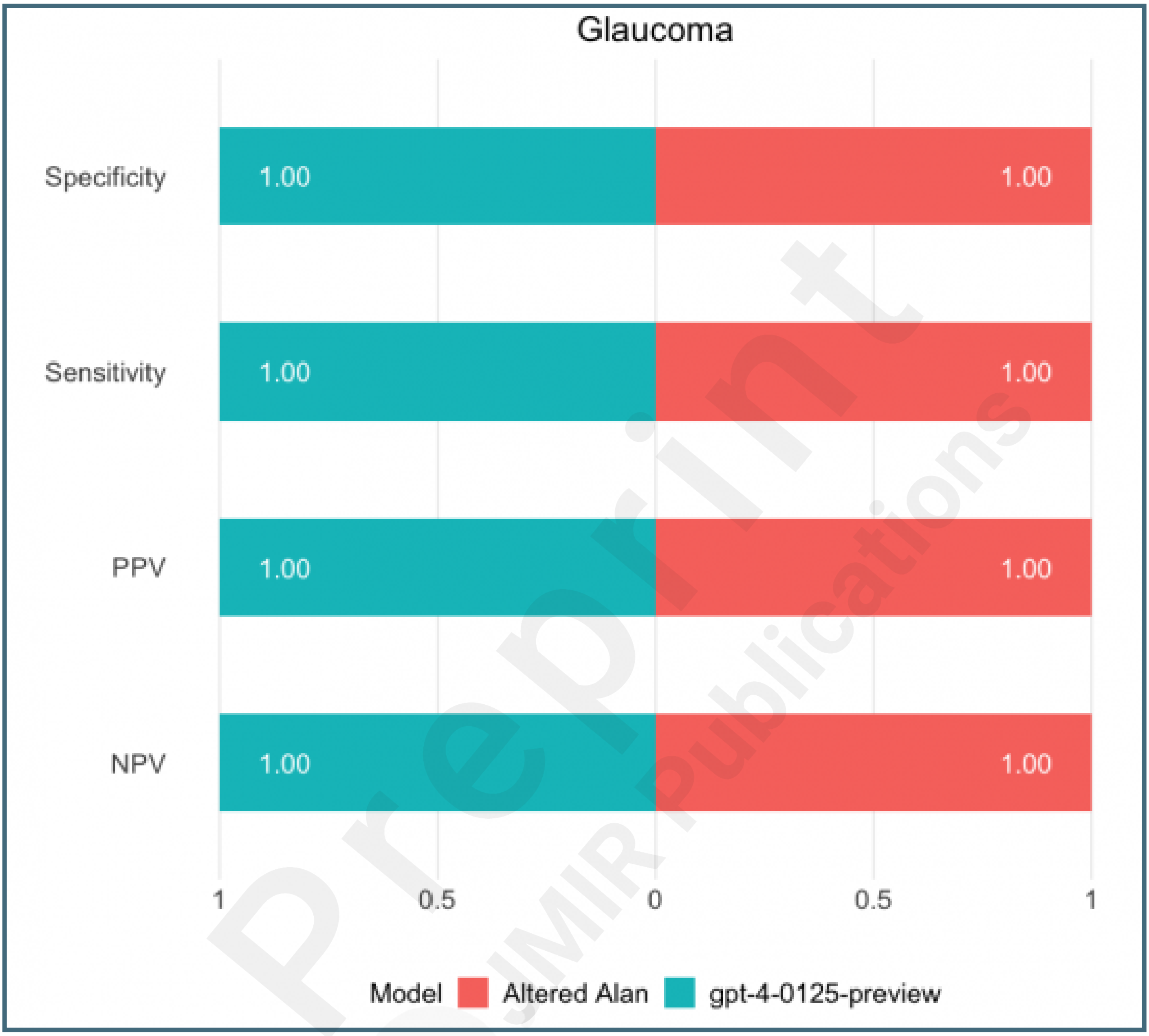
A mind map detailing the types of artificial intelligence.



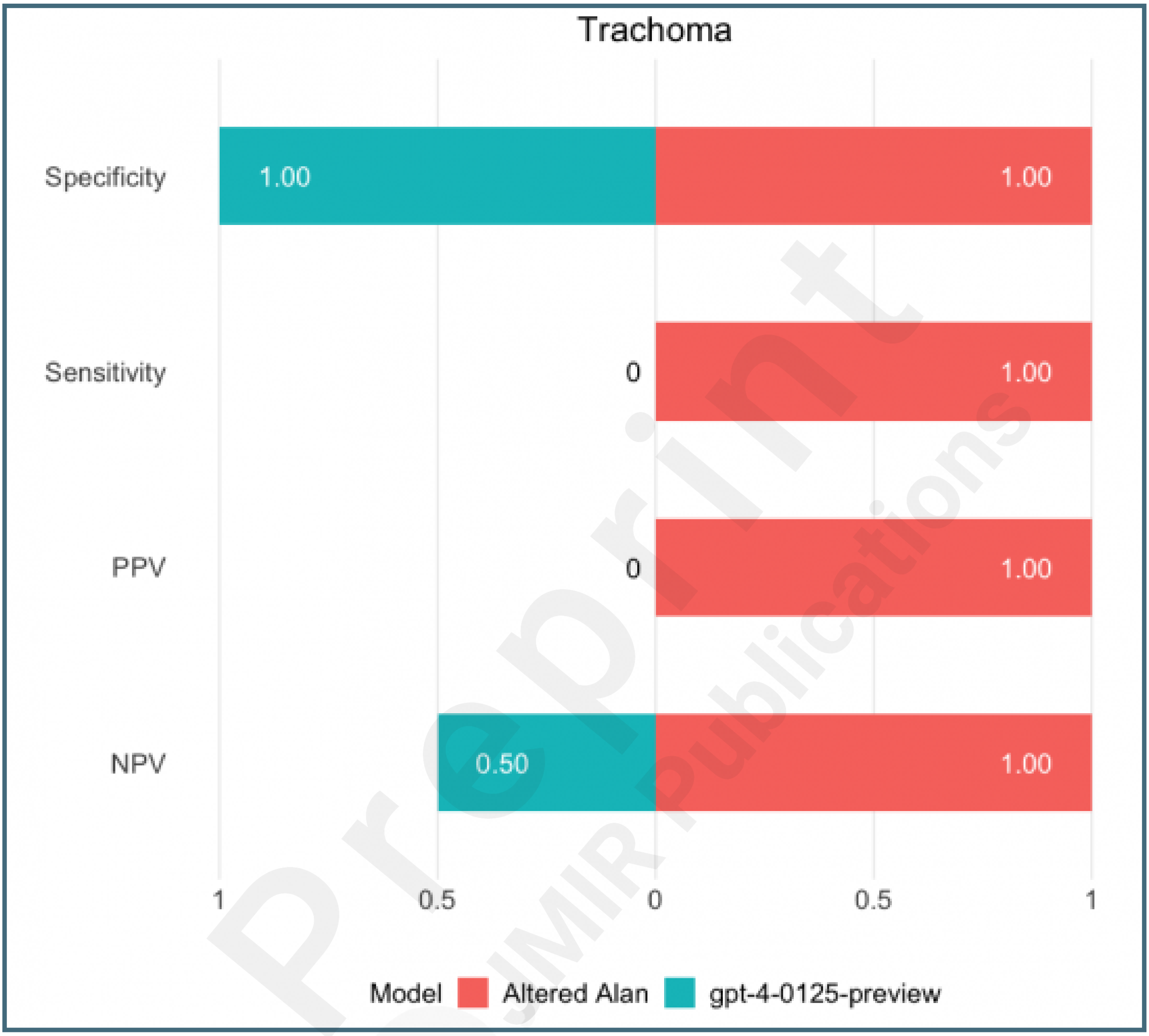
A comparison of AI model performance for cataract.



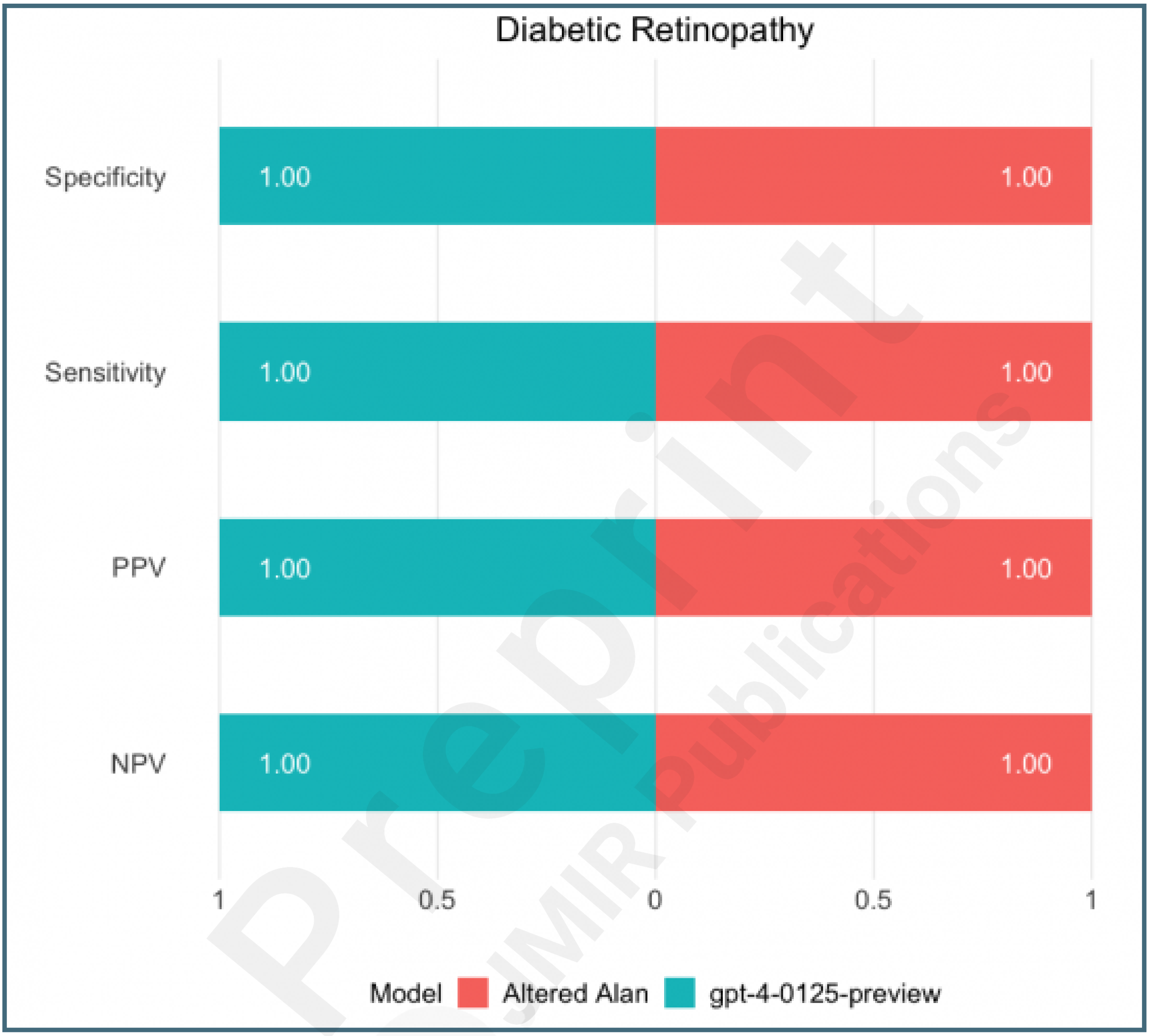
A comparison of AI model performance for glaucoma.



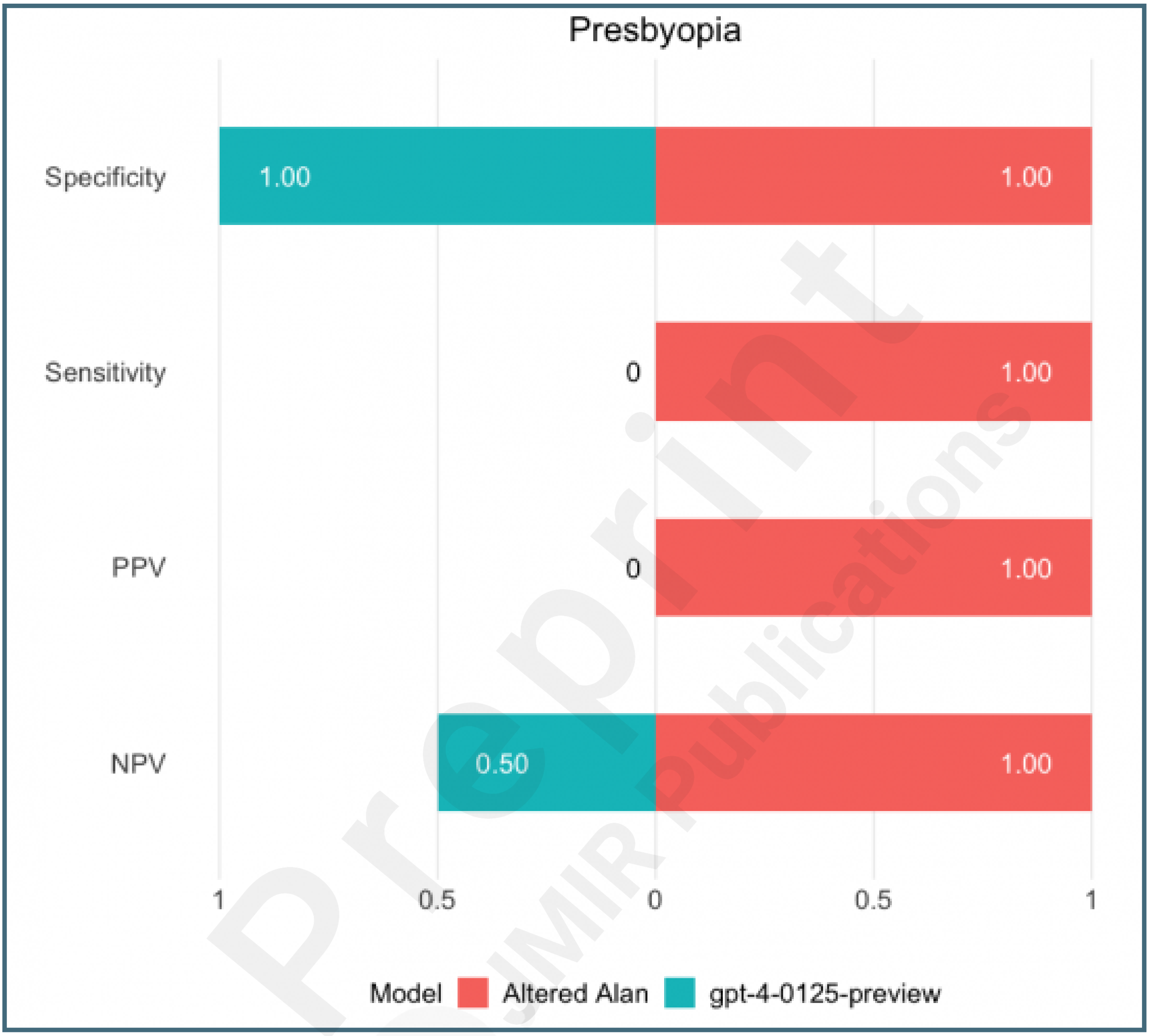
A comparison of AI model performance for trachoma.



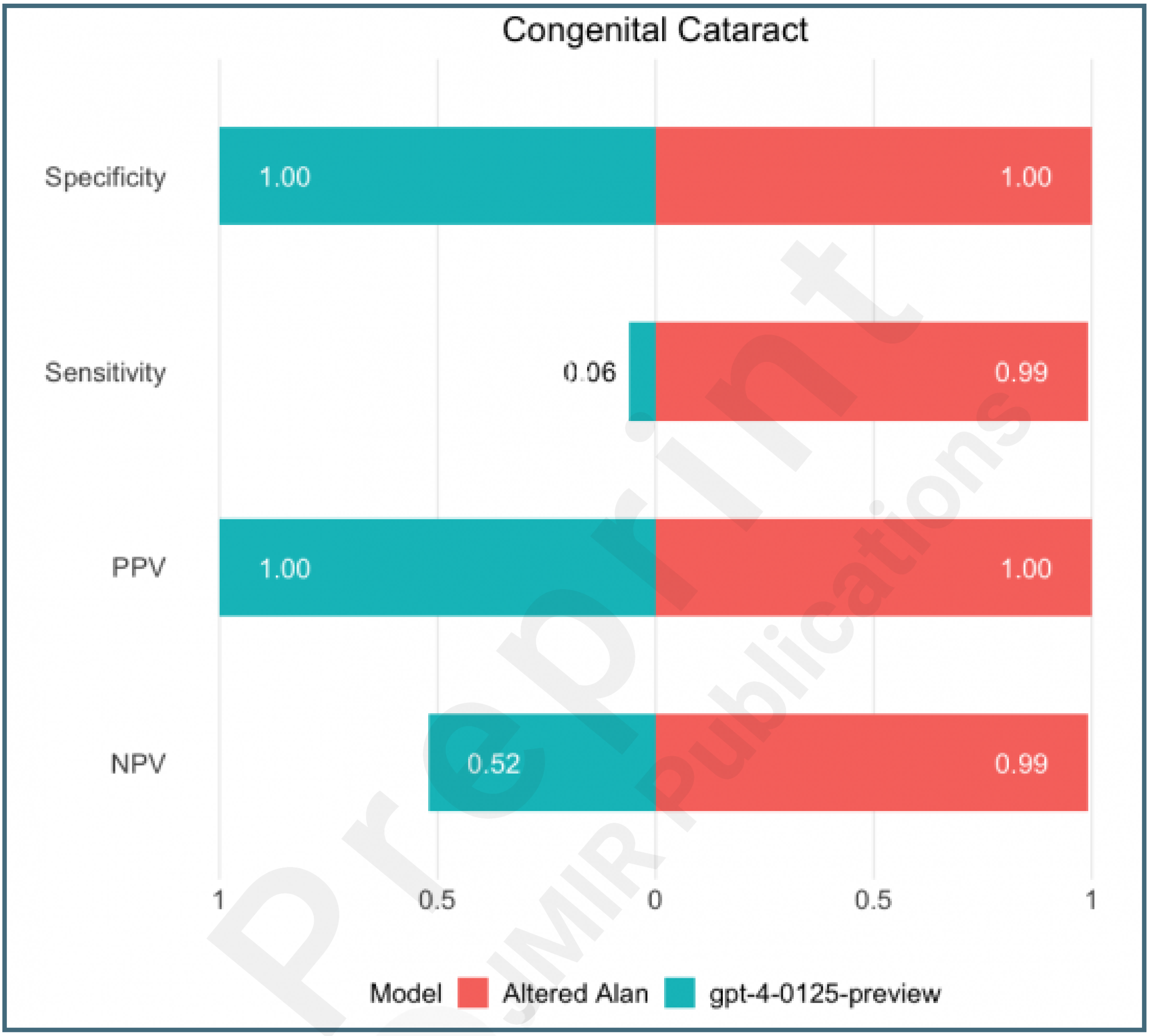
A comparison of AI model performance for diabetic retinopathy.



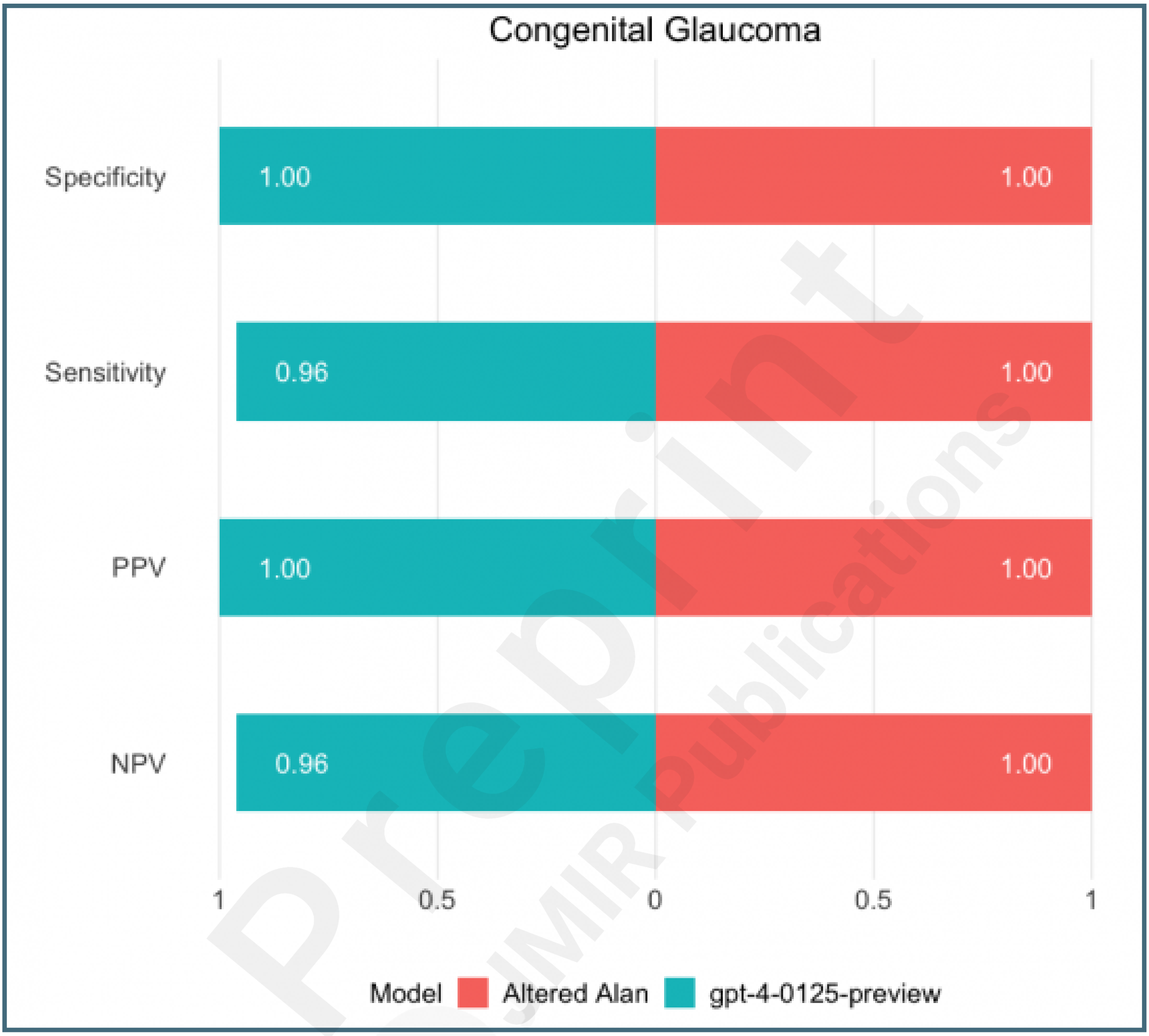
A comparison of AI model performance for presbyopia.



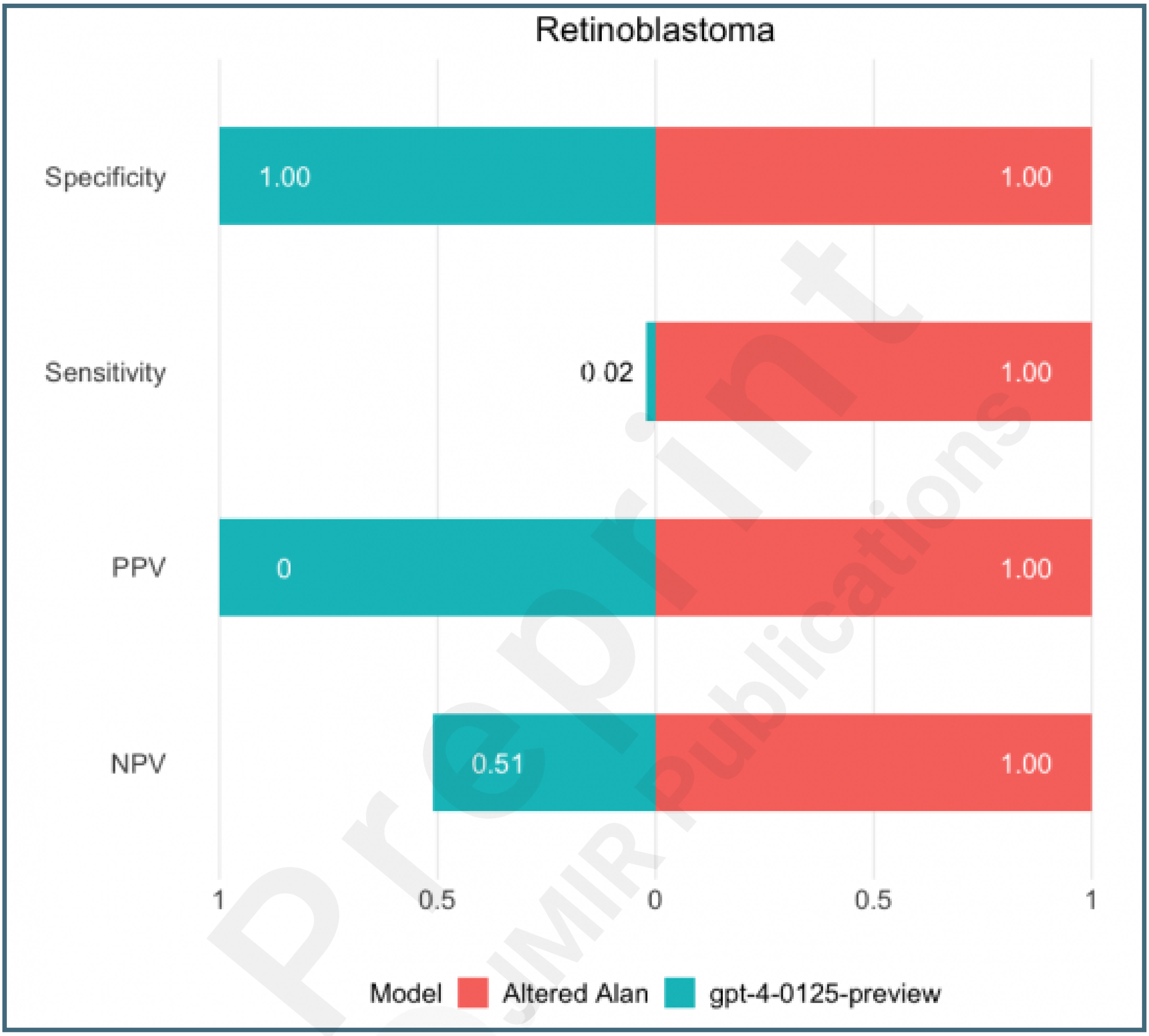
A comparison of AI model performance for congenital cataract.



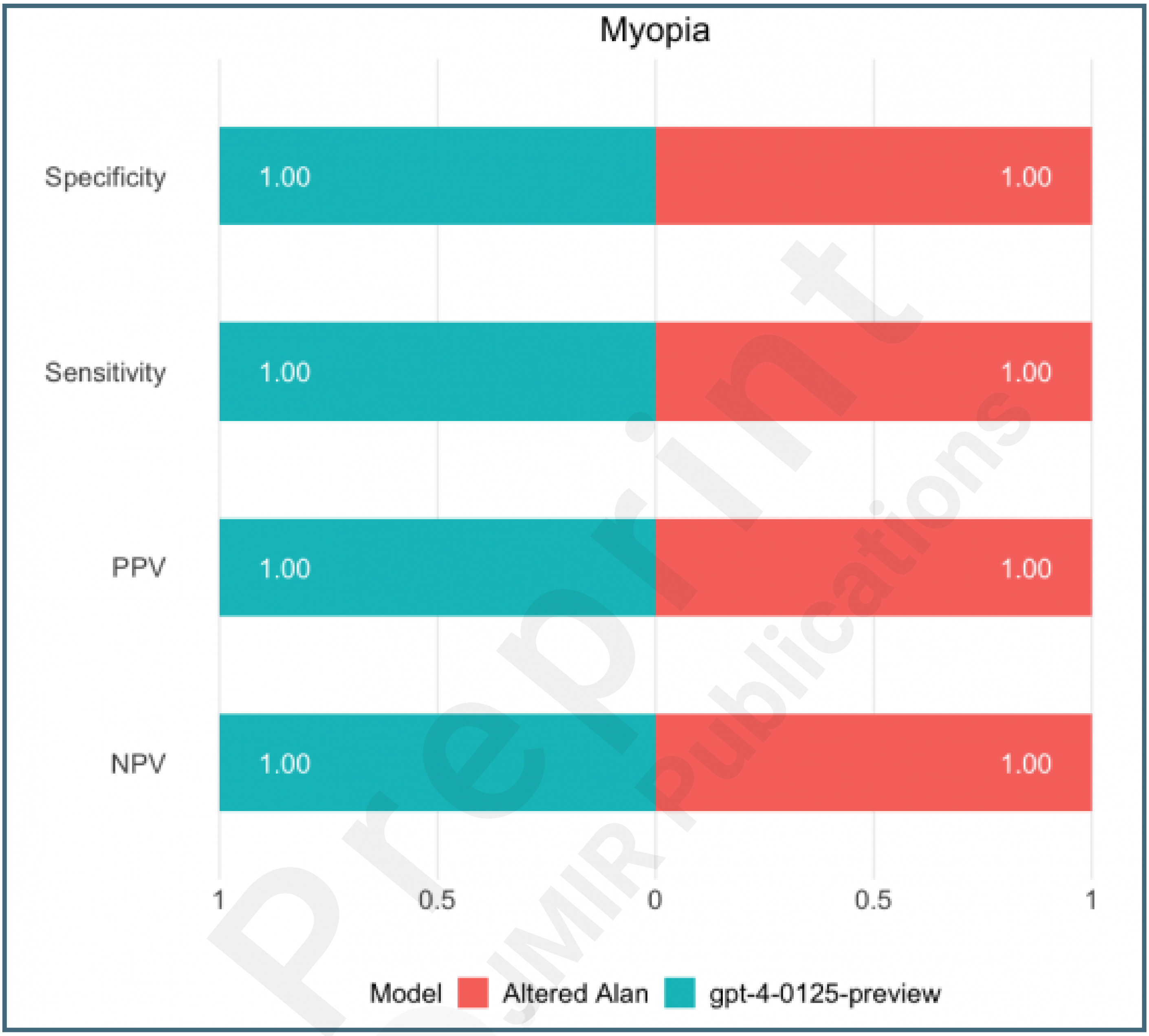
A comparison of AI model performance for congenital glaucoma.



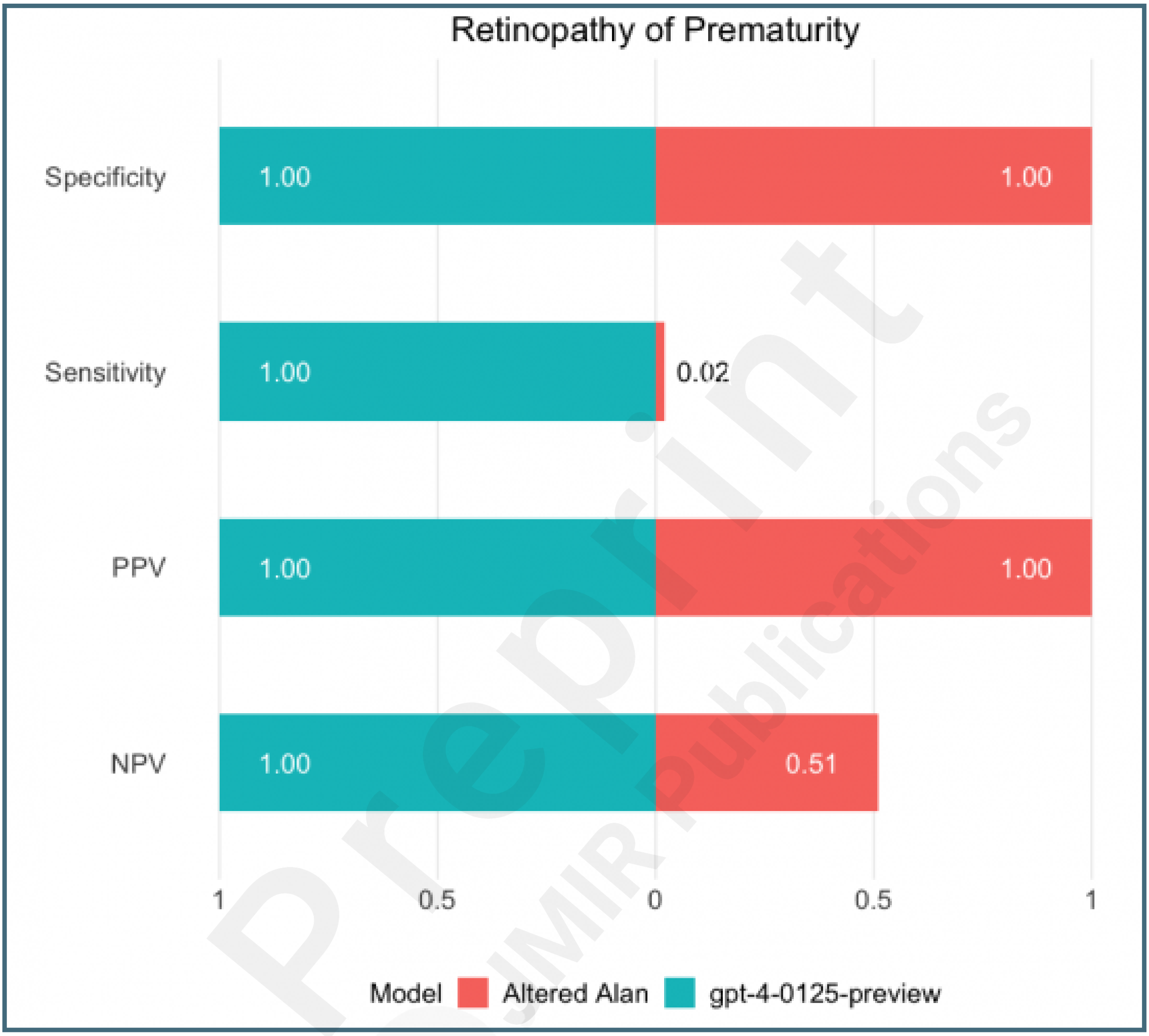
A comparison of AI model performance for retinoblastoma.



A comparison of AI model performance for myopia.



A comparison of AI model performance for retinopathy of prematurity.


Retinopathy of Prematurity
Specificity
Sensitivity
PPV
NPV
Model
🟥 Altered Alan
🟦 gpt-4-0125-preview