A new meaning for NLP – the trials and tribulations of natural language processing with GPT-3 in ophthalmology

Journal:	*British Journal of Ophthalmology*
Manuscript ID	bjophthalmol-2022-321141.R1
Article Type:	Review
Date Submitted by the Author:	n/a
Complete List of Authors:	Nath, Siddharth; McGill University, Ophthalmology and Visual Sciences Marie, Abdullah; Queen's University Belfast, School of Medicine and Dentistry Ellershaw, Simon; University College London, UKRI Centre for Doctoral Training in AI-enabled Healthcare Korot, Edward; Stanford University, Byers Eye Institute Keane, Pearse; Moorfields Eye Hospital, National Institute for Health Research, Biomedical Research Centre for Ophthalmology, UCL Institute of Ophthalmology
Keywords:	Diagnostic tests/Investigation, Telemedicine

A new meaning for NLP – the trials and tribulations of natural language processing with GPT-3 in ophthalmology

Siddharth Nath, **MD, PhD**[1,4]; *Abdullah Marie*[2], *Simon Ellershaw*, **MSc**[3], *Edward Korot,* **MD**[4,5], *Pearse A. Keane*, **MD, FRCOphth**[4]

[1]Department of Ophthalmology and Visual Sciences, Faculty of Medicine and Health Sciences, McGill University, Montréal, Québec, Canada

[2]School of Medicine, Dentistry, and Biomedical Sciences, Queen's University Belfast, Belfast, Northern Ireland

[3]UKRI Centre for Doctoral Training in AI-enabled Healthcare, University College London, London, United Kingdom

[4]National Institute for Health Research, Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom

[5]Byers Eye Institute, Stanford University, Palo Alto, California, USA

***Correspondence to:**

Dr. Pearse A. Keane
National Institute for Health Research
Biomedical Research Centre for Ophthalmology
Moorfields Eye Hospital NHS Foundation Trust
UCL Institute of Ophthalmology
London EC1V 2PD
United Kingdom
pearse.keane1@nhs.net

ABSTRACT

Natural language processing (NLP) is a subfield of machine intelligence focused on the interaction of human language with computer systems. NLP has recently been discussed in the mainstream media and the literature with the advent of Generative Pre-trained Transformer 3 (GPT-3), a language model capable of producing human-like text. The release of GPT-3 has also sparked renewed interest on the applicability of NLP to contemporary health care problems. This article provides an overview of NLP models, with a focus on GPT-3, as well as discussion of applications specific to ophthalmology. We also outline the limitations of GPT-3 and the challenges with its integration into routine ophthalmic care.

INTRODUCTION

Ophthalmologists have conventionally been trained to fear the term ‘NLP’, for it indicates ‘no light perception’ – a level of vision where the patient is effectively blind. In a more modern context, and outside of ophthalmology, however, NLP has grown to most commonly mean ‘natural language processing’ – a subset of machine intelligence focused on the interaction of human language with computer systems.(1)

For as long as computers have existed, NLP has been an area of interest, with Alan Turing’s proposal of what is now called the ‘Turing Test’ – an experiment for determining whether the language generated by a computer is distinguishable from that produced by a human – having been in place since 1950.(2) Modern language models have come remarkably close to passing a Turing Test, with OpenAI’s Generative Pre-trained Transformer 3 (GPT-3) being able to write entire pieces of prose nearly indistinguishable from human authors.(3–5)

GPT-3 is the latest in a series of autoregressive language models that uses deep learning to generate text in a human-like language format. Trained using the Common Crawl, WebText, Books1, and Books2 datasets, and all of English-language Wikipedia, in a ‘few-shot’ approach, GPT-3 is considered to have learned a snapshot of the entire internet.(6,7) Like other autoregressive language models, GPT-3 predicts the next text element in a sequence of words, based upon a query, task, or instruction, given in natural human language. The emphasis with NLP models such as GPT-3, is the production of a result in human-like text, allowing for broad use cases and easy, user-friendly interaction. The reason this is a useful evaluation task pertains to the sentiment of a model correctly predicting the next word in a sequence suggesting that it has a general understanding of the preceding sequence’s meaning and context.

These language models collectively form a group known as ‘transformers’, which are capable of predicting text through a complex mechanism termed ‘attention’, responsible for assigning a contextual value to singular inputs in order to determine the appropriate output. Attention builds upon mechanisms used by earlier language models (called recurrent neural networks or RNNs), which were sequential in nature and would consider all prior inputs in order to determine an output.(8) Transformer models utilize an encoder (responsible for translating the input sequence into a vector which can be interpreted by the model) and a decoder (which converts the model’s processing into human text) in their architecture to complete their queries.

Within the realm of transformers, GPT-3 is especially unique in that it was trained using

175 billion parameters, over 10-times more than prior models with similar architecture, and that it is able to learn using a ‘few-shot’ approach. Effectively, where other language models require substantial inputs to contextualize a response, GPT-3 is able to glean appropriate context from only a few short phrases, or even, single words, and is able to output text that is relevant without additional training. Moreover, GPT-3 differs from other transformers in that it was developed primarily with a decoder architecture. Instead of coding user inputs as a vector for the model to assess, GPT-3 generates responses by processing its own outputs, in the context of prior queries, through the model, effectively ‘learning from itself’. GPT-3 has gained considerable attention in the mainstream media for its ability to process natural language inputs across an array of disciplines and for the breadth of data on which it is trained. For instance, it has been used to write code (and even complete software programs) through natural language instruction, develop dialogue for virtual reality experiences, translate between languages, and even analyze customer experiences.(9–11)

Despite its powerful processing capabilities, GPT-3, and other autoregressive language models like it, have considerable, and even potentially harmful, limitations. As it was trained on a large portion of internet content, GPT-3 incorporates the biases found across the web in its own processing. These include: gender, racial, and geopolitical biases. In the assessment of gender biases, for example, the developers of GPT-3 focused on the associations between gender and occupation. They noted that occupations requiring higher levels of education, such as finance or academic professor, were more likely to be followed by male identifiers, while those such as nurse, midwife, receptionist, and housekeeper, were more likely to be followed by female identifiers.(6,12) GPT-3 also suffers from difficulty with ‘common sense physics’, and lacks context about the world (and consequently, the tasks requested of it) as it is not grounded in domains outside of language, such as video or real-world interaction. Like all autoregressive language models, GPT-3 is also unable to correct itself once it begins to make mistakes. In writing prose, for instance, it is unable to go back and edit, and often one mistake will lead to many more, as it uses preceding words to predict its next output.

Nonetheless, given its impressive capabilities and the potential for implementation without the need for ground-up programming, GPT-3 has garnered significant attention from the health care community, and offers compelling solutions to contemporary clinical problems.(13) This article will provide an overview of some of the existing health care applications of GPT-3, and discuss potential uses and challenges in applying it to ophthalmology.

EXISTING HEALTH CARE APPLICATIONS OF GPT-3

NLP models have found applications across various health care disciplines, most prominently in the assessment of electronic health record (EHR) data. Language models are particularly well-suited to this task, as they can process large volumes of text at a scale unachievable by human assessors and require little-to-no visual context for extracting requested data. Researchers have developed NLP models for a variety of EHR use-cases, from extracting physician-reported pain data from oncologic consultation notes, to identification of potential clinical trial participants by way of extracting inclusion and exclusion criteria.(14,15)

Given its initial release and limited availability through a restricted-access program requiring formal application and approval, health care applications of GPT-3 remain sparse. Most recently, Logé and colleagues at Stanford University developed a dataset for assessing bias in medical question-answering systems and tested it using both GPT-3 and its predecessor, GPT-2.(16) This work represents a particularly challenging use of autoregressive language models as it introduces a highly complex and multifactorial clinical problem. Pain-management in the clinic is complicated by the individual lived experience of pain across patients, variable manifestations of pain sequelae, inherent subjectivity in pain reporting, and clinician biases, both implicit and explicit.(17–19) They noted that GPT-3 advocated for treating every patient with pain management, however, there was significant variability across individual medical contexts and simulated patient gender and race. GPT-3 was 3.6% more likely to refuse pain treatment to black patients than white patients, and, equally more likely to refuse pain treatment to women.(16) This work demonstrates, that, although highly trained autoregressive language models like GPT-3 have the potential to address complex clinical problems, clinicians should be wary that inherent societal biases may be incorporated in their training data and resulting implementations.

POTENTIAL APPLICATIONS OF GPT-3 IN OPHTHALMOLOGY

Within ophthalmology, NLP has been trialed in EHR-driven use-cases as well.(20) For instance, NLP pipelines have been developed to identify patients with an array of ocular diseases, including glaucoma, herpes zoster ophthalmicus, cataracts, and pseudoexfoliation syndrome.(21–

24) NLP has also been developed to triage ophthalmic referrals, and has found uses in predicting the rate of antibiotic use and intraoperative complications from operative notes, as well as predicting patient quality of life in association with vision loss in the setting of chronic ocular disease.(25–28) All of the above applications required *de novo* development of pipelines along with requisite planning for large training and validation datasets. Although impressive in their capacity, these pipelines also remain siloed in their own use cases, unable to pull data across fields to inform their tasks. For instance if a patient with cataracts presents and is triaged by a herpes zoster ophthalmicus model, the outputs of the model will not be meaningful.

GPT-3 offers ophthalmologists and researchers the opportunity to build from a ready-to-use application programming interface (API) and remains 'informed' about multiple disciplines as its training effectively encompasses the English Internet. GPT-3 is also noteworthy in that it can be trained using minimal input (called zero-shot, or few-shot learning), improving the ease with which solutions can be developed and deployed and also reducing upstart costs. Few-shot learning removes the need for costly expansive training datasets as GPT-3 can effectively 'learn' from a handful of inputs, and sometimes no inputs at all. For instance, GPT-3 could be used to form the backbone of an ophthalmic triage system that offers human-like responses to emergent questions, with minimal training on potential prompts. Such a system could be implemented in emergency departments or optometrist offices to ensure clinical resources are used in a timely, equitable, and efficient manner. Moreover, this system could also implement GPT-3's language translation capabilities to accept questions in an array of local languages and relay responses back in a patient-friendly dialect. Intelligent triage systems have already shown promise in the practice of ophthalmology during the COVID-19 pandemic, and an implementation with GPT-3 would fill an important, unmet need.(29) GPT-3 could also be used to develop a consultation note creation system that standardizes ophthalmology notes across subspecialties and clinicians. Although NLP, and more broadly, machine intelligence, have been used to develop ambient virtual scribe systems, such as AutoScribe, these implementations remain limited as they can only record dialogue and offer suggestions to clinicians, requiring review of raw speech data and substantial cognitive and time effort.(30) As GPT-3 has been shown to write prose from 'few-shot' learning, it could conceivably be trained to create consult notes for ophthalmology from a few keywords entered by clinicians. Coupling this capacity with dictation software could create a powerful EMR tool which would create standardized, mail-ready consult notes. Such a system could streamline already overloaded ophthalmology clinics, improve the timeliness of responses to referring physicians, and ease involvement of patients in research. GPT-3's intelligent language capabilities could also be used to create teaching tools for global ophthalmology outreach. For instance, charities like Orbis, could implement GPT-3 to improve translation of highly technical medical data and streamline patient education and local physician training on medical missions. Finally, and perhaps most compellingly, GPT-3's ability to write nascent code with little training could aid other ophthalmologists and vision science researchers in developing their own AI solutions. Our group has previously evaluated performance of clinicians without coding experience in the development of AI solutions and we have resolved that clinician-driven AI requires appropriate education of physicians in the nuances of the field.(31,32) A GPT-3-based coding platform would leap many of these educational requirements and open the field of AI solution development to end-user clinicians, paving the way for previously unexplored applications.

CHALLENGES IN IMPLEMENTATION OF GPT-3 IN OPHTHALMOLOGY

Despite these possibilities, use of GPT-3 in clinical practice carries inherent challenges. Because GPT-3 was trained using a large proportion of the internet, it carries within its processing capabilities congruent biases of gender, race, and politics, which within a health care setting, could be devastating in patient-facing implementations. Such biases could work to alienate already vulnerable patient populations, lead to inequitable and inadequate care, and further cement such biases in contemporary medical practice. To combat this, clinicians using GPT-3 should consider building safeguards; for instance, within ophthalmology, as gender and race data is not required for triaging many ocular emergencies, such data could be omitted at the collection stage.

In addition to inherent biases, implementation of GPT-3 within clinical practice is also limited by its inability to correct mistakes, and more worryingly, press on with an inaccurate processing stream. For instance, GPT-3 was widely criticized in the mainstream media for encouraging a simulated patient to commit suicide when trialed in a psychiatric question-answer implementation.(33) In order to prevent similar catastrophic results from an ophthalmic implementation of GPT-3, clinicians could ensure that key outputs are assessed for raw quality by a human evaluator. Although this increases required resources, it would still create a 'semi-autonomous' system, which, if implemented for instance as a triage tool, could nonetheless improve access to care. To minimize the need for clinician resources, a filtering system could be developed, whereby only significant interventions, if suggested by an intelligent system, need to be reviewed by a human evaluator. Interventions or recommendations which do not have potential for harm to patients may not require human approval, thereby reducing the required human cognitive resources. Such systems would be analogous to modern academic medical practice, where junior doctors often make simple clinical decisions independently, but review more complex interventions with senior supervisors or consultants.

Implementation of GPT-3 into ophthalmology is also limited by its lack of image processing capability. Ophthalmology is a highly visual specialty, with examination findings at the slit lamp or multimodal imaging guiding most treatment decisions. To address this, clinicians could turn to transformer models based on GPT-3 which combine text and imaging data, for their solutions. OpenAI, for instance, has recently announced the Contrastive Language-Image Pre-training (CLIP) model and the DALL-E 2 model, both of which combine text and image processing functionality. In fact, DALL-E 2 has gained considerable attention recently for its ability to create nascent images from text input, offering ophthalmologists a powerful augment to GPT-3's language processing capabilities.(34,35) Lastly, implementation of GPT-3 into routine ophthalmic care has been limited by access to the platform itself. GPT-3 was initially available only through a web-based API, managed by OpenAI, but licensed 'exclusively' to Microsoft. As of November 2021, OpenAI has expanded availability of the API to end users in a pre-defined list of countries, however, use of the API remains subject to adherence with OpenAI's guidelines.(36) While democratization of AI tools and arguments for open access versus privately developed algorithms are beyond the scope of this article, it remains important to consider that any potential health care applications require compliance with institutional and regulatory agency privacy laws as well as consistent reliable performance with limited downtime. Clinicians should consider real-world integration as an important aspect of any AI solution prior to piloting implementations.

CONCLUSIONS

In summary, autoregressive language models such as GPT-3, offer clinicians tremendous potential for addressing complex clinical problems. Within ophthalmology, GPT-3 is especially compelling as it has the capability to improve use of valuable ophthalmic clinical resources, improve clinic flow, and bring AI-solution development to end-user clinicians. Ophthalmologists and researchers should remain wary of GPT-3's inherent biases and consider their impact on patients within each specific use-case. Careful implementation of GPT-3 and combination text-image models, such as DALL-E 2, holds the potential for transforming patient care and revolutionizing contemporary ophthalmology.

COMPETING INTERESTS

Dr. Nath, Mr. Marie, and Mr. Ellershaw, have no financial disclosures. Dr. Korot has acted as a consultant for Google Health and Genentech and is an equity holder in Reti Health. Dr. Keane has acted as a consultant for DeepMind, Roche, Novartis, Apellis, and BitFount, and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan, and Bayer.

ACKNOWLEDGEMENTS

Dr. Keane is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1).

REFERENCES

1. Guida G, Mauri G. Evaluation of natural language processing systems: Issues and approaches. Proc IEEE. 1986 Jul;74(7):1026–35.

2. Turing A. Computing Machinery and Intelligence (1950) [Internet]. The Essential Turing. 2004. Available from: http://dx.doi.org/10.1093/oso/9780198250791.003.0017

3. Elkins K, Chun J. Can GPT-3 pass a writer's Turing test? J cult anal [Internet]. 2020 Sep 14; Available from: https://culturalanalytics.org/article/17212.pdf

4. Gpt-3. A robot wrote this entire article. Are you scared yet, human? Guardian. 2020;

5. Floridi L, Chiriatti M. GPT-3: Its Nature, Scope, Limits, and Consequences. Minds Mach. 2020 Dec 1;30(4):681–94.

6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv [cs.CL]. 2020. Available from: http://arxiv.org/abs/2005.14165

7. Mehmood MA, Shafiq HM, Waheed A. Understanding regional context of World Wide Web using common crawl corpus. In: 2017 IEEE 13th Malaysia International Conference on Communications (MICC). 2017. p. 164–9.

8. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [Internet]. arXiv [cs.CL]. 2014. Available from: http://arxiv.org/abs/1406.1078

9. Vincent J. OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws. The Verge. 2020;

10. Scout H. GPT-3 and AI in Customer Support [Internet]. 2021 [cited 2021 Aug 24]. Available from: https://www.helpscout.com/blog/ai-in-customer-support/

11. Vincent J. OpenAI's text-generating system GPT-3 is now spewing out 4.5 billion words a day [Internet]. The Verge. 2021 [cited 2021 Aug 25]. Available from: https://www.theverge.com/2021/3/29/22356180/openai-gpt-3-text-generation-words-day

12. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017 Apr 14;356(6334):183–6.

13. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med. 2021 Jun 3;4(1):93.

14. Naseri H, Kafi K, Skamene S, Tolba M, Faye MD, Ramia P, et al. Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. J Biomed Inform. 2021;120:103864.

15. Beck JT, Vinegra M, Dankwa-Mullan I, Torres A, Simmons CC, Holtzen H, et al. Cognitive technology addressing optimal cancer clinical trial matching and protocol feasibility in a community cancer practice. J Clin Orthod. 2017 May 20;35(15_suppl):6501–6501.

16. Logé C, Ross E, Dadey DYA, Jain S, Saporta A, Ng AY, et al. Q-Pain: A Question Answering Dataset to Measure Social Bias in Pain Management [Internet]. 2021 [cited 2021 Aug 24]. Available from: https://openreview.net/pdf?id=Ud1K-l71AI2

17. Mossey JM. Defining racial and ethnic disparities in pain management. Clin Orthop Relat Res. 2011 Jul;469(7):1859–70.

18. Coghill RC. Individual differences in the subjective experience of pain: new insights into mechanisms and models. Headache. 2010 Oct;50(9):1531–5.

19. Mularski RA, White-Chu F, Overbay D, Miller L, Asch SM, Ganzini L. Measuring pain as the 5th vital sign does not improve quality of pain management. J Gen Intern Med. 2006 Jun;21(6):607–12.

20. Yang LWY, Ng WY, Foo LL, Liu Y, Yan M, Lei X, et al. Deep learning-based natural language processing in ophthalmology: applications, challenges and future directions. Curr Opin Ophthalmol. 2021 Sep 1;32(5):397–405.

21. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc. 2012 Mar;19(2):225–34.

22. Barrows RC Jr, Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proc AMIA Symp. 2000;51–5.

23. Zheng C, Luo Y, Mercado C, Sy L, Jacobsen SJ, Ackerson B, et al. Using natural language processing for identification of herpes zoster ophthalmicus cases to support population-based study. Clin Experiment Ophthalmol. 2019 Jan;47(1):7–14.

24. Stein JD, Rahman M, Andrews C, Ehrlich JR, Kamat S, Shah M, et al. Evaluation of an Algorithm for Identifying Ocular Conditions in Electronic Health Record Data. JAMA Ophthalmol. 2019 May 1;137(5):491–7.

25. Tan Y, Bacchi S, Casson RJ, Selva D, Chan W. Triaging ophthalmology outpatient referrals with machine learning: A pilot study. Clin Experiment Ophthalmol. 2020 Mar;48(2):169–73.

26. Smith DH, Johnson ES, Russell A, Hazlehurst B, Muraki C, Nichols GA, et al. Lower visual acuity predicts worse utility values among patients with type 2 diabetes. Qual Life Res. 2008 Dec;17(10):1277–84.

27. Gaskin GL, Pershing S, Cole TS, Shah NH. Predictive Modeling of Risk Factors and Complications of Cataract Surgery. Eur J Ophthalmol. 2016 Jul 1;26(4):328–37.

28. Liu L, Shorstein NH, Amsden LB, Herrinton LJ. Natural language processing to ascertain two key variables from operative reports in ophthalmology. Pharmacoepidemiol Drug Saf. 2017 Apr;26(4):378–85.

29. Wu X, Chen J, Yun D, Yuan M, Liu Z, Yan P, et al. Effectiveness of an Ophthalmic Hospital-Based Virtual Service during the COVID-19 Pandemic. Ophthalmology. 2021 Jun;128(6):942–5.

30. Crampton NH. Ambient virtual scribes: Mutuo Health's AutoScribe as a case study of artificial intelligence-based technology. Healthc Manage Forum. 2020 Jan;33(1):34–8.

31. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health. 2019 Sep;1(5):e232–42.

32. Keane PA, Topol EJ. AI-facilitated health care requires education of clinicians. Lancet. 2021 Apr 3;397(10281):1254.

33. Daws R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. AI News. 2020;

34. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision [Internet]. arXiv [cs.CV]. 2021. Available from: http://arxiv.org/abs/2103.00020

35. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-Shot Text-to-Image Generation [Internet]. arXiv [cs.CV]. 2021. Available from: http://arxiv.org/abs/2102.12092

36. OpenAI. OpenAI's API now available with no waitlist [Internet]. OpenAI. 2021 [cited 2022 Apr 5]. Available from: https://openai.com/blog/api-no-waitlist/

NHS
National Institute for
Health Research

Biomedical Research Centre for Ophthalmology
2nd floor, Richard Desmond Children's Eye Centre
Moorfields Eye Hospital NHS Foundation Trust
City Road
London
EC1V 2PD

April 22, 2022

Professor Frank Larkin
Editor-in-Chief, *British Journal of Ophthalmology*

RE: bjophthalmol-2022-321141 - A new meaning for NLP - the trials and tribulations of natural language processing with GPT-3 in ophthalmology

Dear Frank,

We would like to extend our sincerest thanks to the editors and reviewers for your comments, suggestions, and constructive criticisms of our manuscript.We have carefully revised our work by modifying the text and responding to reviewers and the editorial office in a point-by-point format, with editorial comments presented in bold text and our responses directly below in non-bolded format. We have grouped our responses under the headings: Responses to Reviewer #1, Responses to Reviewer #2, and Responses to Editor-in-Chief. Thank you kindly for considering our paper and for the opportunity to provide a revised submission. We do truly hope that our revised manuscript will be of interest to the broad readership of the *British Journal of Ophthalmology*.

Responses to Reviewer #1

General comments

This paper provides a timely overview of this important intersection between Ophthalmology and NLP that has garnered growing interest as the availability of GPT3 has accelerated the translation of clinical AI text applications from bench to bedside.

Thank you for your feedback. We hope that our manuscript will serve as a resource for ophthalmologists and researchers embarking on development of digital health solutions with GPT-3.

Major comments

The manuscript is highly informative for readers. There may be some updates wrt page 4 (and similar repeat on page 7): "Given its recent release and limited availability through a restricted-access program requiring formal application and approval") e.g. open AI themselves has released an API for GPT3 access without the waitlist: https://openai.com/blog/api-no-waitlist/

We agree; since its initial launch via a semi-private API, access to GPT-3 has now expanded considerably. We have revised the sections on page 4 and page 7 of the manuscript to reflect the present accessibility of GPT-3, while maintaining the role its initial exclusivity may have had in limiting the rapid development of ophthalmological applications.

The manuscript provides an excellent discussion of potential applications for AI such as GPT3 in Ophthalmology, such as triage. Some existing literature may be relevant for this discussion, such as this study during COVID-19 whereby the use of AI-based chatbots for over 10,000 eye-related visits were reported: Wu X, Chen J, Yun D, et al. Effectiveness of an Ophthalmic Hospital-Based Virtual Service during the COVID-19 Pandemic. Ophthalmology 2021;128(6):942-945. doi:10.1016/j.ophtha.2020.10.012

Thank you for your suggestion. We have included the above study as a reference within the revised manuscript and commented upon its importance.

The authors have highlighted important limitations and potential safety issues from GPT3 that may warrant a semi-autonomous model. Some further expansion for the final point regarding the need for clinicians to consider real-world integrations may be beneficial for readers. Are there any key considerations or suitable clinical operational models for GPT3 that the authors may recommend?

While there are not any previously defined clinical operational models within the literature for use with GPT-3, one well-established model which implementations of GPT-3 could be made analogous to is that of the junior doctor and the consultant. Much like how a junior doctor is able to make independent decisions when the risk for harm to a patient is minimal, a GPT-3-driven implementation may not require supervision when interventions or care it recommends have little to no potential harm for patients. Akin to junior doctors reviewing complex cases with their consultant supervisors, in situations where an intelligent GPT-3-based system recommends a route which has the potential for significant harm to a patient, it could create an automatic ‘flag’ which prompts a human assessor to intervene and audit the proposed pathway. We have amended the revised manuscript to include the above as a potential clinical operational model, in line with your suggestions.

Minor comments
Minor awkward phrasing “most equated with...” (pg 3 line 8-9)

The phrasing of the above sentence has been revised to improve readability.

<u>**Responses to Reviewer #2**</u>

This paper provides a brief review of the language model GPT-3 and a perspective on how it can be used in ophthalmology. Please find my comments below.

Thank you for your time and effort taken to assess our manuscript and provide feedback.

First, considering the audience, the description on GPT-3 could be more precise. For instance, in the Introduction, it mentions the 'few-shot' approach without any explanations. Potential readers may not understand it.

Thank you for this suggestion. We have improved upon the description of GPT-3 within the Introduction. We have incorporated further explanation of the architecture of language models, and how GPT-3 works, including its ‘few-shot’ learning capabilities.

Second, a more thorough review and comparison of GPT-3 with other language models are needed. For instance, one important difference of GPT-3 compared with other language models is that it focuses on the decoder part (for text generation, summarization, and translation purpose) whereas many other similar language models focus on the encoder part (mostly for text representation). The intro only mentions their similarities but not the differences. This is important because it shows the unique part of GPT-3 for specific use cases in the Application section.

We appreciate this suggestion. In line with your comments, we have added a section to the Introduction where GPT-3 is compared against existing language models, with emphasis on its unique attributes. We agree this lends better to the discussion of use cases later on in the manuscript.

In addition, there are similar articles on GPT-3 in the medical domain (e.g., https://www.nature.com/articles/s41746-021-00464-x). The study should compare with the existing studies and discuss in more depth especially in the ophthalmology domain.

Thank you for raising this point. The above referenced study is included in our bibliography and is an excellent starting point for clinicians looking to become equated with GPT-3. Our study differs in that it focuses specifically on applications of GPT-3 in ophthalmology with potential real-world examples and key barriers to implementation. In line with your suggestion and feedback from the review process, we have further expanded upon the specific ophthalmic applications.

<u>**Responses to Editor-in-Chief**</u>

The journal encourages submissions on applications of AI and relevant new developments in this field. This manuscript is well written and the section headings well chosen. However as pointed out by reviewer #2 and as you will be aware, the challenge for authors is to introduce this and similar topics to non-expert readers with adequate explanation of concepts and significance. Please bear this in mind throughout your revision.

Thank you for your comments and positive feedback. We have revised our manuscript in line with the comments put forth by both reviewers in order to improve the readability of our paper for a non-expert audience.

We thank the reviewers and editors for their input and feedback and for the opportunity to provide an improved report. We do hope our revised manuscript will be of interest to the broad readership of *British Journal of Ophthalmology*.

Respectfully re-submitted on behalf of the group,

Pearse Keane

Pearse A Keane, MD MSc FRCOphth MRCSI
Honorary Consultant Ophthalmologist, Moorfields Eye Hospital
Professor of Artificial Medical Intelligence, University College London (UCL)
UK Research & Innovation, Future Leaders Fellow