The triage and diagnostic accuracy of frontier large language models: an updated comparison to physician performance

Michael Joseph Sorich, Arduino Aleksander Mangoni, Stephen Bacchi, Bradley Douglas Menz, Ashley Mark Hopkins

Submitted to: Journal of Medical Internet Research
on: October 10, 2024

Table of Contents

The triage and diagnostic accuracy of frontier large language models: an updated comparison to physician performance

Michael Joseph Sorich[1] BPharm(Hons), GradDipMedStat, PhD; Arduino Aleksander Mangoni[1, 2] MD, PhD; Stephen Bacchi[3] MBBS, PhD; Bradley Douglas Menz[1] BPharm(Hons); Ashley Mark Hopkins[1] BPharm(Hons), PhD

[1]College of Medicine and Public Health Flinders University Adelaide AU
[2]Department of Clinical Pharmacology Southern Adelaide Local Health Network Adelaide AU
[3]Department of Neurology and the Center for Genomic Medicine Massachusetts General Hospital and Harvard Medical School Boston US

Corresponding Author:
Michael Joseph Sorich BPharm(Hons), GradDipMedStat, PhD
College of Medicine and Public Health
Flinders University
GPO Box 2100
Adelaide
AU

Abstract

Frontier large language models (LLMs) demonstrate triage and diagnostic accuracy comparable to physicians and significantly improve over earlier models and lay individuals, with collaborative LLM approaches enhancing diagnostic performance.

(JMIR Preprints 10/10/2024:67409)
DOI: https://doi.org/10.2196/preprints.67409

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 u**
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

Paper type: Research letter

Title: The triage and diagnostic accuracy of frontier large language models: an updated comparison to physician performance

Authors: Michael J. Sorich, PhD[1]; Arduino A. Mangoni, PhD[1,2]; Stephen Bacchi, PhD[3], Bradley D. Menz, B. Pharm (Hons)[1]; Ashley M. Hopkins, PhD[1]

[1]College of Medicine and Public Health, Flinders University, Adelaide, Australia

[2]Department of Clinical Pharmacology, Southern Adelaide Local Health Network, Adelaide, Australia.

[3]Department of Neurology and the Center for Genomic Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02138, USA

Corresponding author: Professor Michael Sorich, PhD, College of Medicine and Public Health, Flinders University, Adelaide, SA 5042, Australia (michael.sorich@flinders.edu.au)

Keywords: generative artificial intelligence; large language models; triage; diagnosis

Introduction:

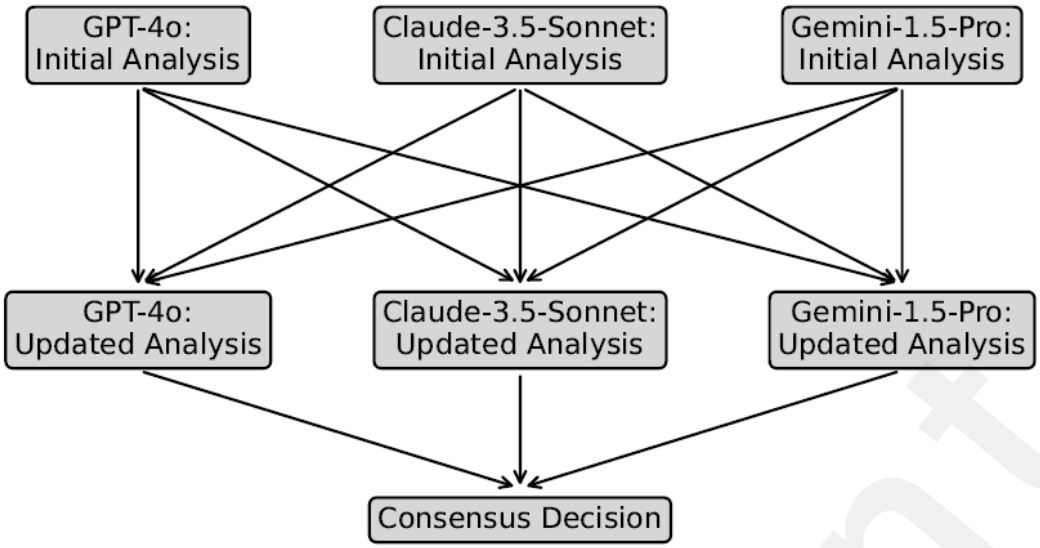
The medical capabilities of large language models (LLMs) are progressing rapidly [1-3]. Benchmarking LLMs against human performance with clinically relevant tasks enables tracking current capabilities and progress. The triage (level/urgency of care to seek) and diagnostic accuracy of the GPT-3 model were recently compared with 5000 lay individuals using the Internet and 21 practising primary care physicians [4]. The triage ability of GPT-3 was significantly inferior to that of physicians – having similar accuracy to lay individuals. The diagnostic ability was close to but below that of physicians [4]. Whether more recent frontier LLMs are still inferior to physicians on this benchmark is uncertain.

Methods:

The 48 case vignettes - including both common and severe conditions - validated by Levine and colleagues [4] were evaluated utilising GPT-4o-2024-05-13 (OpenAI), Claude-3.5-Sonnet (Anthropic), and Gemini-1.5-Pro-001 (Google) via a Python API. The LLMs were instructed to identify potential diagnoses and provide step-by-step reasoning. Subsequently, they reflected on the reasoning and selected the top three most likely diagnoses in order of likelihood. For triage prediction, the LLM was supplied with the vignette and the three diagnoses it predicted. It was instructed to identify the urgency of the required medical care, including its step-by-step reasoning.

A multi-agent workflow involving collaboration between the three distinct LLMs was also evaluated (Figure 1). Each LLM was provided with its initial analysis (decision plus reasoning) and the analyses of the two other LLMs. Each LLM was instructed to reflect on all analyses and update its proposed diagnoses/triage as appropriate. The consensus decision was then made based on a majority vote. Llama-3.1-405B (Meta) was used to identify the consensus diagnosis.

Figure 1: Large Language Model (LLM) Collaboration – a triage/diagnosis workflow involving initial analysis (the LLM's initial decision and step-by-step reasoning), updated analysis (reflect on all LLM initial analyses and update decision if appropriate), and consensus decision (majority vote of the individual LLM's updated decisions).

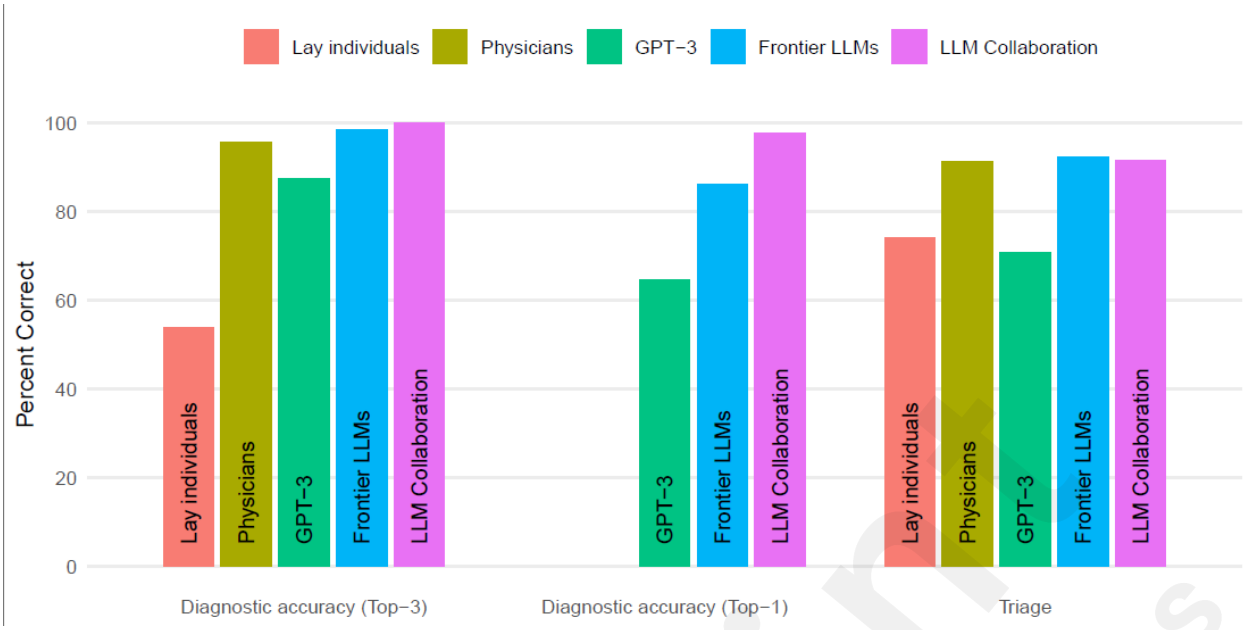


Diagnostic accuracy was evaluated by whether the correct diagnosis was one of the three proposed by the LLM (Top-3).[4] Additionally, the accuracy of the first-ranked diagnosis (Top-1) was assessed. Triage was assessed as urgent (emergency department or seeing a doctor within a day) vs non-urgent (seeing a doctor within a week or self-care) [4]. The prompts and LLM settings are provided in the supplemental methods.

Results:

The correct diagnosis was among the top three proposed diagnoses for 98.6% (142/144; Frontier LLMs) and 100% (48/48; LLM Collaboration) of cases (Figure 2). This compares to 87.5% for GPT-3, 54% for lay individuals, and 95.6% for physicians, as reported by Levine and colleagues [4]. The single most likely diagnosis prediction was correct for 86.1% (124/144; Frontier LLMs) and 97.8% (47/48; LLM Collaboration) of cases, compared to 64.6% for GPT-3 [4].

Figure 2: Diagnostic (Top-3 and Top-1) and triage accuracy of lay individuals with internet access[4], primary care attending physicians[4], GPT-3[4], Frontier LLMs (aggregate of GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro), and the LLM Collaboration (of GPT-4o, Claude 3.5 Sonnet and Gemini 1.5 Pro). Results for lay individuals, physicians and GPT-3 are reproduced from Levine and colleagues.[4] The Top-3 diagnostic accuracy assesses whether any of the three proposed diagnoses was correct, and the Top-1 diagnostic accuracy assesses whether the proposed first-ranked diagnosis was correct. Triage accuracy is based on identifying urgent or non-urgent situations. Note: Levine and colleagues did not report top-1 diagnostic accuracy for lay individuals and physicians.[4]



Triage was correct for 92.4% (133/144; Frontier LLMs) and 91.7% (44/48; LLM Collaboration). This compares to 70.8% for GPT-3, 74.1% for lay individuals, and 91.3% for physicians [4]. The most common error was overestimating the urgency.

Discussion:

For triage with these clinical vignettes, frontier LLMs now perform substantially better than lay individuals who could use the internet (before the availability of LLMs), and similarly to primary care physicians. This capability is consistent with recent evaluations of modern LLMs for emergency department triage [5, 6].

These results suggest that, with the help of contemporary LLMs, lay individuals may now better assess the level and urgency of medical care than was possible from internet searching in 2019. Formally evaluating this will be an important direction for future research.

The rapid progress in LLM capabilities poses challenges for tracking their current capability for health-related tasks. This includes challenges for traditional peer-reviewed publications, which can become outdated by the time of publication.

Additionally, we show that newer techniques involving collaboration between multiple distinct LLMs may improve diagnostic performance. Other methods, such as fine-tuning and in-context learning (e.g., integrating search functionality and demonstrations of how to work through complex cases), offer opportunities to improve performance [1, 2].

Acknowledgements: Dr Sorich is supported by a Beat Cancer Research Fellowship from the Cancer Council South Australia. Dr Hopkins holds an Emerging Leader Investigator Fellowship from the National Health and Medical Research Council, Australia (APP2008119). The PhD scholarship of Mr Menz is supported by the National Health and Medical Research Council, Australia (APP2030913). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflict of Interest Disclosures: Dr Sorich reported receiving grants from Pfizer, AstraZeneca Boehringer Ingelheim, and the National Health and Medical Research Council of Australia outside the submitted work. No other disclosures were reported.

Data availability: The case vignettes utilized are publicly available (doi: 10.1016/S2589-7500(24)00097-9). Prompts utilized are available via the supplemental methods.

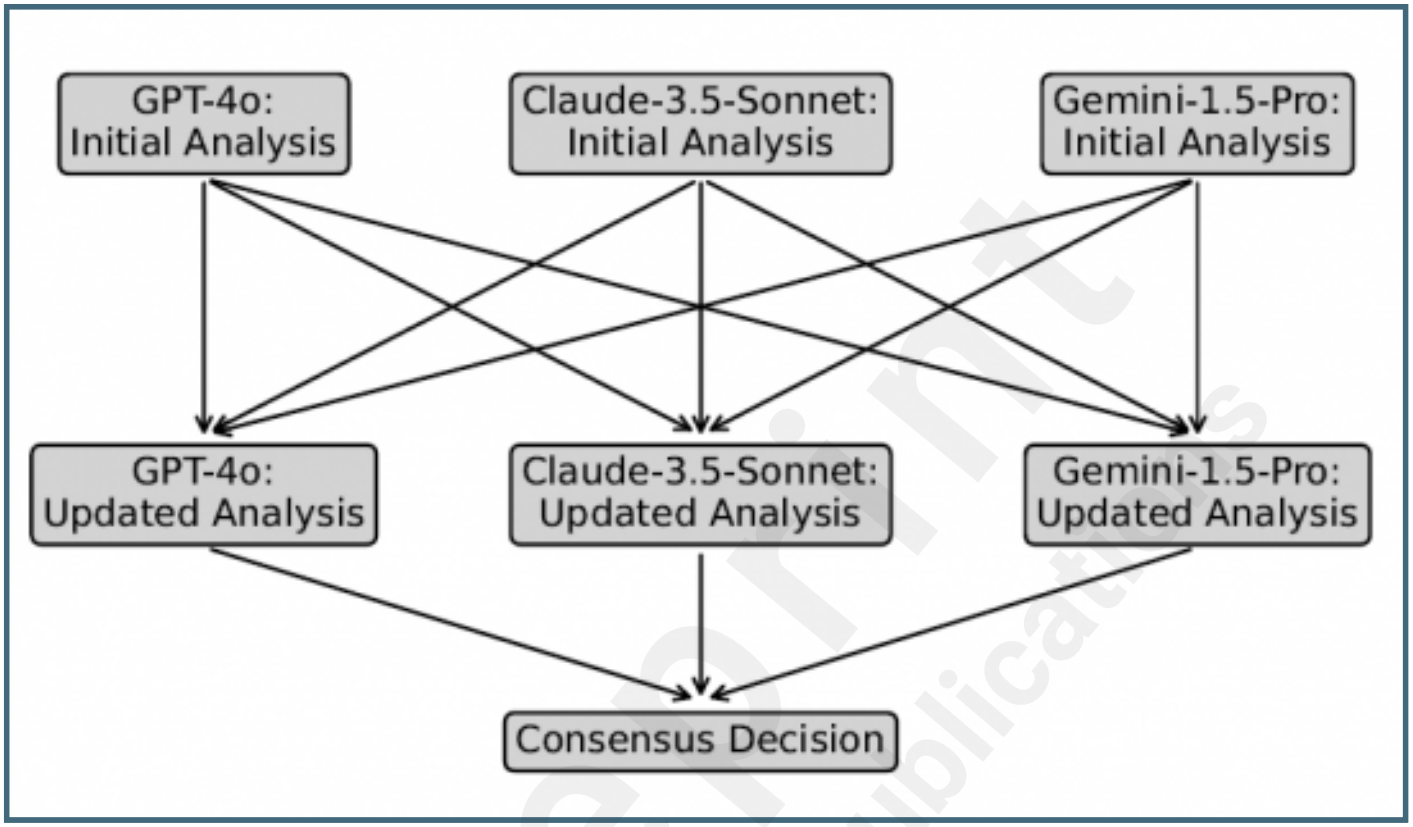
References:

1. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine2023 November 01, 2023:[arXiv:2311.16452 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2023arXiv231116452N.
2. Saab K, Tu T, Weng W-H, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of Gemini Models in Medicine2024 April 01, 2024:[arXiv:2404.18416 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv240418416S.
3. Sorich MJ, Menz BD, Hopkins AM. Quality and safety of artificial intelligence generated health information. BMJ. 2024;384:q596. doi: 10.1136/bmj.q596.
4. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. The Lancet Digital Health. 2024;6(8):e555-e61. doi: 10.1016/S2589-7500(24)00097-9.
5. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department. JAMA Network Open. 2024;7(5):e248895-e. doi: 10.1001/jamanetworkopen.2024.8895.
6. Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. J Med Internet Res. 2024;26:e53297. PMID: 38875696. doi: 10.2196/53297.

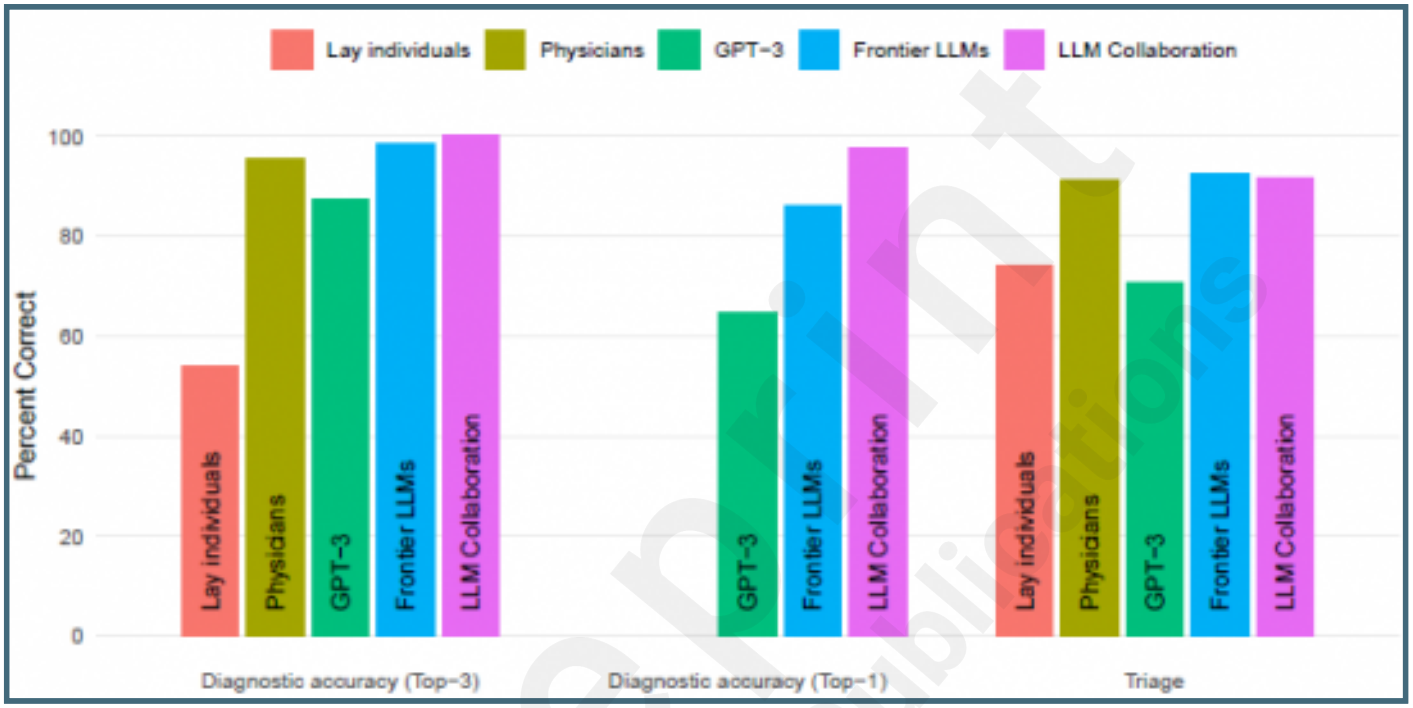
Supplementary Files

Figures

Large Language Model (LLM) Collaboration – a triage/diagnosis workflow involving initial analysis (the LLM's initial decision and step-by-step reasoning), updated analysis (reflect on all LLM initial analyses and update decision if appropriate), and consensus decision (majority vote of the individual LLM's updated decisions).



Diagnostic (Top-3 and Top-1) and triage accuracy of lay individuals with internet access [4], primary care attending physicians [4], GPT-3[4], Frontier LLMs (aggregate of GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro), and the LLM Collaboration (of GPT-4o, Claude 3.5 Sonnet and Gemini 1.5 Pro). Results for lay individuals, physicians and GPT-3 are reproduced from Levine and colleagues [4]. The Top-3 diagnostic accuracy assesses whether any of the three proposed diagnoses was correct, and the Top-1 diagnostic accuracy assesses whether the proposed first-ranked diagnosis was correct. Triage accuracy is based on identifying urgent or non-urgent situations. Note: Levine and colleagues did not report top-1 diagnostic accuracy for lay individuals and physicians [4].



Multimedia Appendixes

Supplemental methods.
URL: http://asset.jmir.pub/assets/887168107c0eca959d250428ba670180.docx