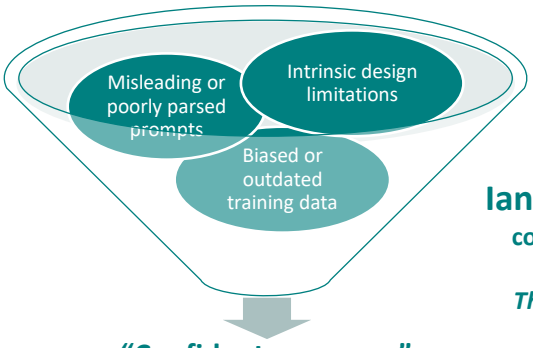
From GPT to GP? Testing ChatGPT's medical competence

Andrea Bertuzzi (Researcher)

The release of ChatGPT, Open AI's latest generative large language model (LLM), has garnered media attention recently.[1] ChatGPT is a 175-billion parameter model using deep learning algorithms trained on huge amounts of data; the breakthrough feature that made it a public phenomenon is its interface, designed to generate human-like responses to users' prompts.

Many users have already tested ChatGPT's ability to answer queries on technical subjects, **BUT** the tool is far from infallible!



Ian Bogost – contributing writer, *The Atlantic*

"[ChatGPT] doesn't make accurate arguments or express creativity, but instead **produces textual material** in a form corresponding with **the requester's explicit or implicit intent**, which might also contain truth under certain circumstances."[1]

LLMs could have applications in various areas of healthcare, from disease surveillance to medical education.[2] However, previous LLM iterations have shown serious limitations when tested on clinical knowledge through generative question-answering tasks. Thanks to its dialogic design, ChatGPT could provide better and novel use cases, but can it perform (at least) comparably to a medically trained human?

How does ChatGPT perform on the United States Medical Licensing Examination? The implications of Large Language Models for medical education and knowledge assessment[3]

Objective

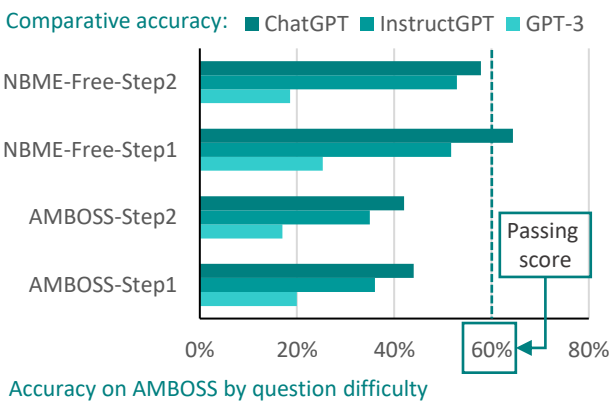
- To assess the performance of ChatGPT on questions within the scope of the **United States Medical Licensing Examination** (USMLE) **Step 1 and Step 2 exams**.

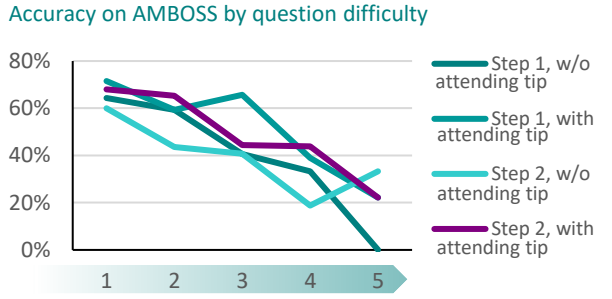
Methods

- Medical Education data sets from the **AMBOSS** question bank (two 100-question sets) and the **National Board of Medical Examiners (NBME)** (two 120-questions sets).
- Questions containing figures or tables were excluded.
- Answers assessed for **logical reasoning** and use of **information** both **internal** and **external** to the question.
- Incorrect answers assessed for **logical**, **informational** and **statistical errors**.

Results

- ✓ ChatGPT achieved **accuracy >40% on all four data sets** and qualified for a **"pass" on NBME-Step1** (64.4%).
- ✓ ChatGPT **outperformed previous LLMs** InstructGPT and GPT-3.
- ✓ ChatGPT **performance decreased with increasing question difficulty** on the AMBOSS data sets.
- ✓ **All answers** to NBME sets, both correct and incorrect, provided a **logical justification**.





Our thoughts:

- ChatGPT's output is known to be **conditioned by the prompt structure**; also, it is impossible to rule out **potential biases** present in Open AI's training dataset (which remains undisclosed) or introduced by the model's structure.
- In light of this, the authors' claim that ChatGPT **"performs at a level expected of a third-year medical student"** must be strongly put into context – **"on one data set out of four and under curated conditions"**.
- However, the authors propose a convincing use case for future iterations of ChatGPT as an **adjunct for peer group education**.
- Overall, the study shows that ChatGPT is a **definite improvement** on previous LLMs in terms of medical application potential, but excessive enthusiasm (or concern) is misplaced for now.

1. Bogost I. ChatGPT Is Dumber Than You Think [Internet]. The Atlantic. 2022. Available from: https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/
2. Marr B. Revolutionizing Healthcare: The Top 14 Uses Of ChatGPT In Medicine And Wellness [Internet]. Forbes. [cited 2023 Apr 3]. Available from: https://www.forbes.com/sites/bernardmarr/2023/03/02/revolutionizing-healthcare-the-top-14-uses-of-chatgpt-in-medicine-and-wellness/?sh=68a50ddc6e54
3. Gilson A et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 8;9:e45312. doi: 10.2196/45312. PMID: 36753318; PMCID: PMC9947764.