

Combating Provider Burnout: ChatGPT 4.0 is Equivalent to a Plastic Surgery Provider in Answering Image-Based Postoperative Patient Questions

Boris Joutovsky, DO¹, Amanda Beneat, MD¹, Nisha Sondhi, MD¹, Anna Jacobs MD¹, Katie E. Weichman, MD^{1,2}, and Lawrence Draper, MD^{1,2}
¹New York University - Long Island School of Medicine ² Hansjörg Wyss Department of Plastic Surgery, New York University School of Medicine

Department of Surgery
NYU - Langone Long Island

INTRODUCTION

Dependence on EMR for patient communication has risen since COVID-19, contributing to increased provider burnout.¹ Patient difficulty in describing lesions has led to unnecessary clinic or ED visits. Utilizing large language models (LLMs) such as ChatGPT to respond to patient questions has become popular since its creation.²

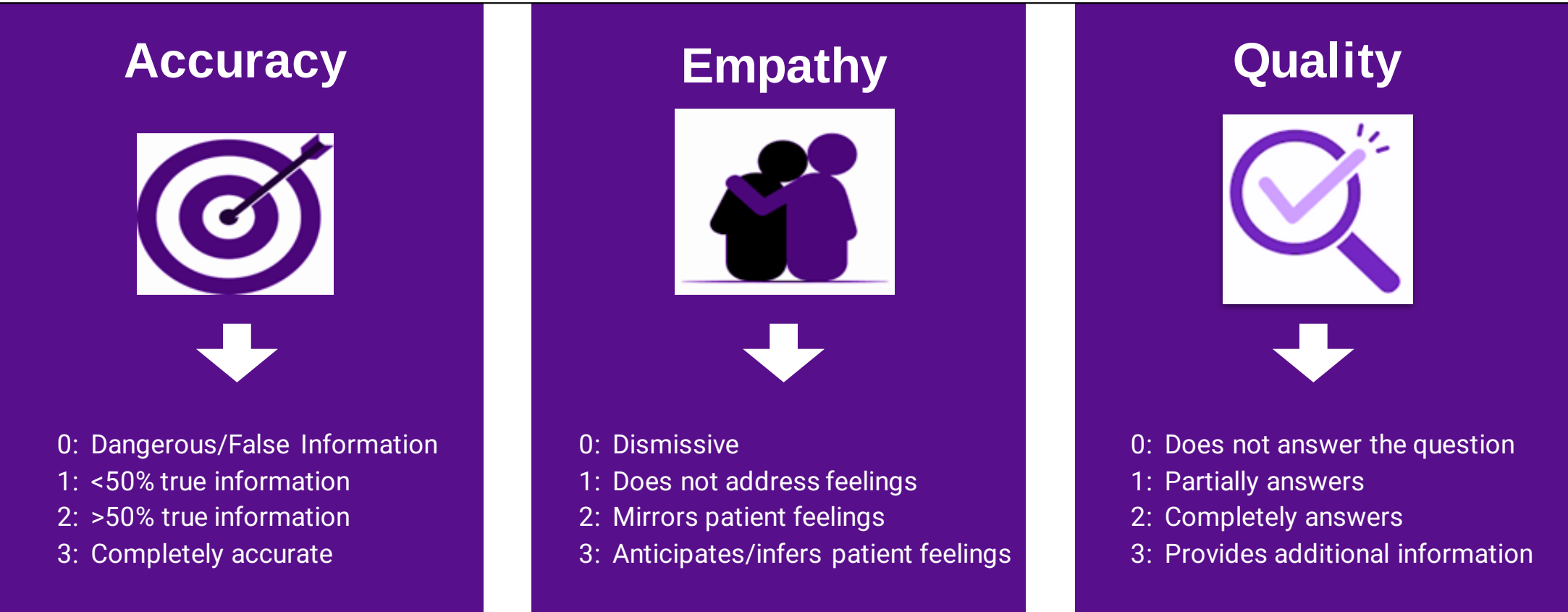
Studies have proven reliability of LLMs to address questions accurately with **low rates of false information**.³ To date, there are no studies in utilizing LLMs to response to **image-based** patient questions in breast surgery. Adding this functionality could further alleviate provider burden.

METHODS

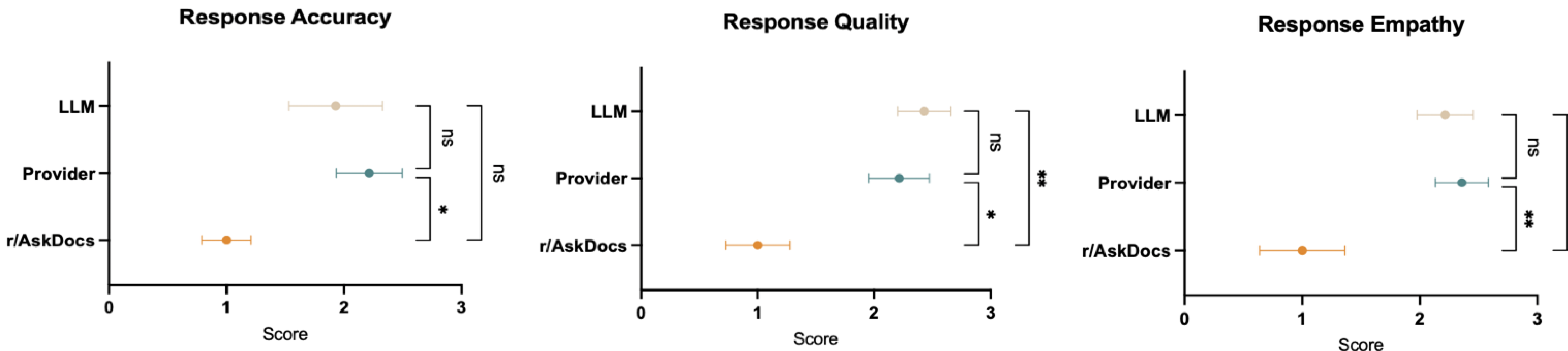
Question Selection:^{3,4}
Patient questions were sourced from the Reddit forum r/AskDocs using terms like “plastic surgery,” “cosmetic surgery”, and “plastics”. Only questions with images and responses from verified physicians, nurse practitioners (NPs), or physician assistants (PAs) were used.

AI and Provider Response Generation:
De-identified questions were entered into a new ChatGPT 4.0 session for individual responses. The same questions were given to a clinic PA who handles patient inquiries, simulating real patient interactions.

Analysis:
Responses from AI, verified providers, and clinic PA were blindly evaluated by NPs and PAs in plastic (n=6) and breast surgery (n=4) clinics for breast related questions on:



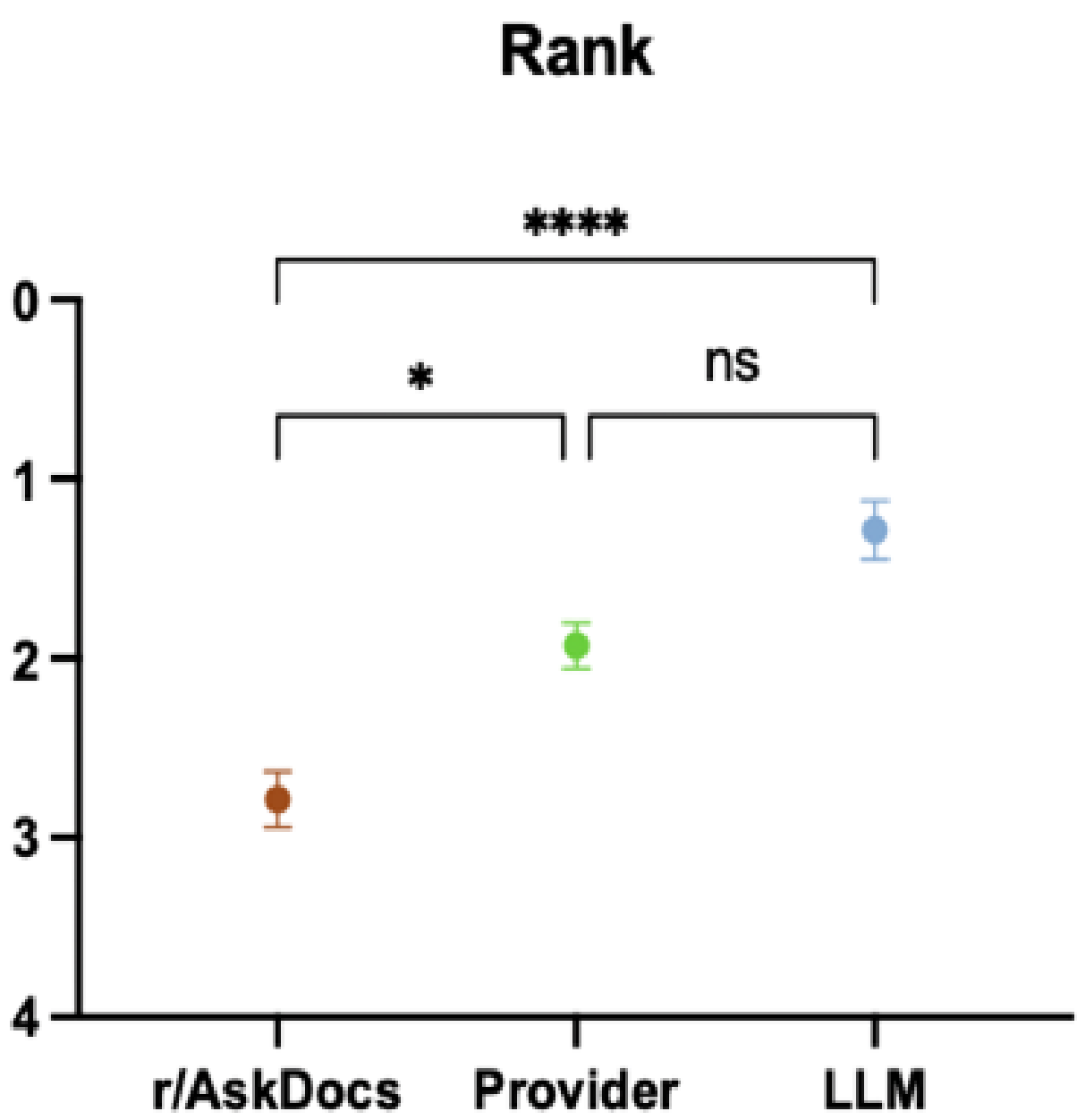
RESULTS



a. LLMs consistently outperformed online responses and were equivalent real-world provider responses.

Providers were provided the original question and all three responses in a blinded format and asked to grade each response on a scale of 0-3 with 0 being the worst score and 3 being the best possible score. ChatGPT 4.0 scored well in all three domains tested and was superior to online provider responses to each question. Responses from a dedicated plastic surgery provider were equivalently scored.

p-value * < 0.05; **< 0.001

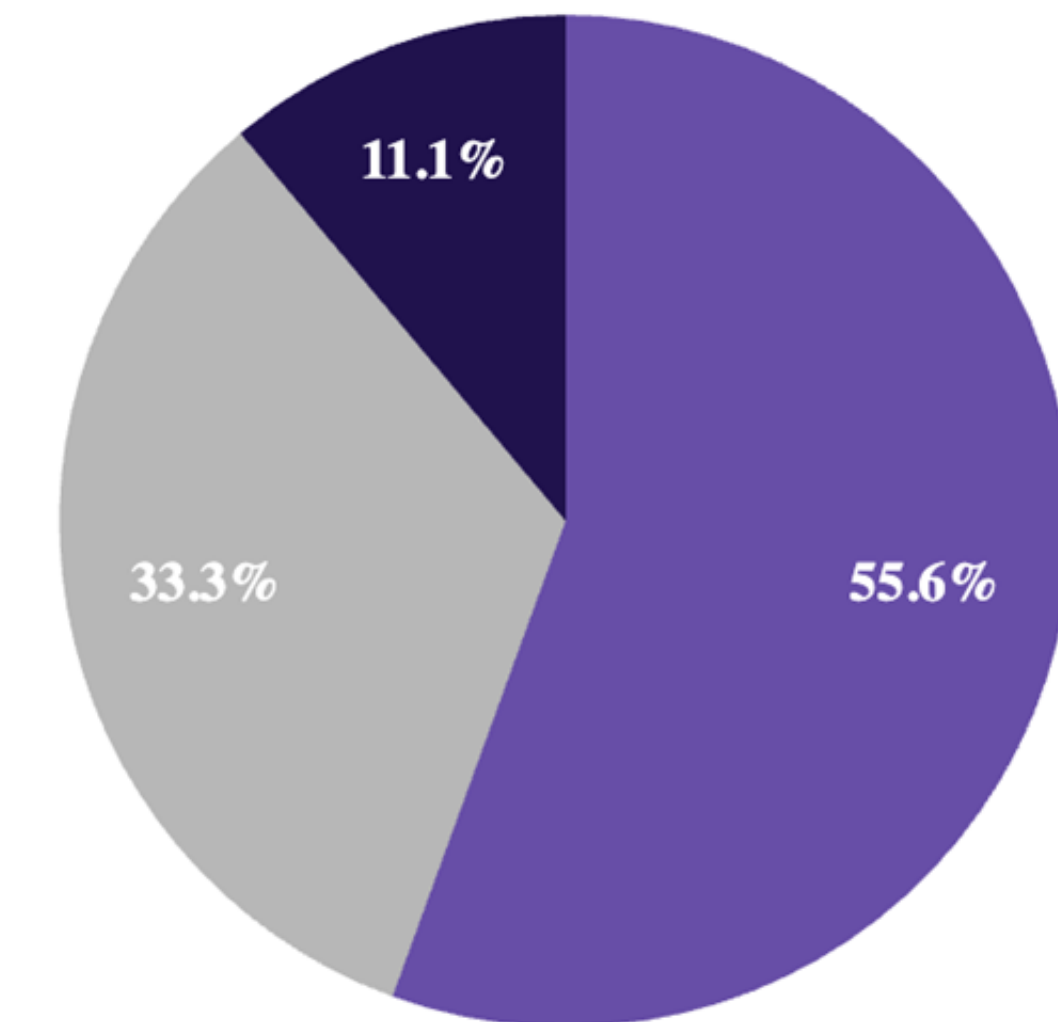


b. LLMs were ranked highest among all three response types

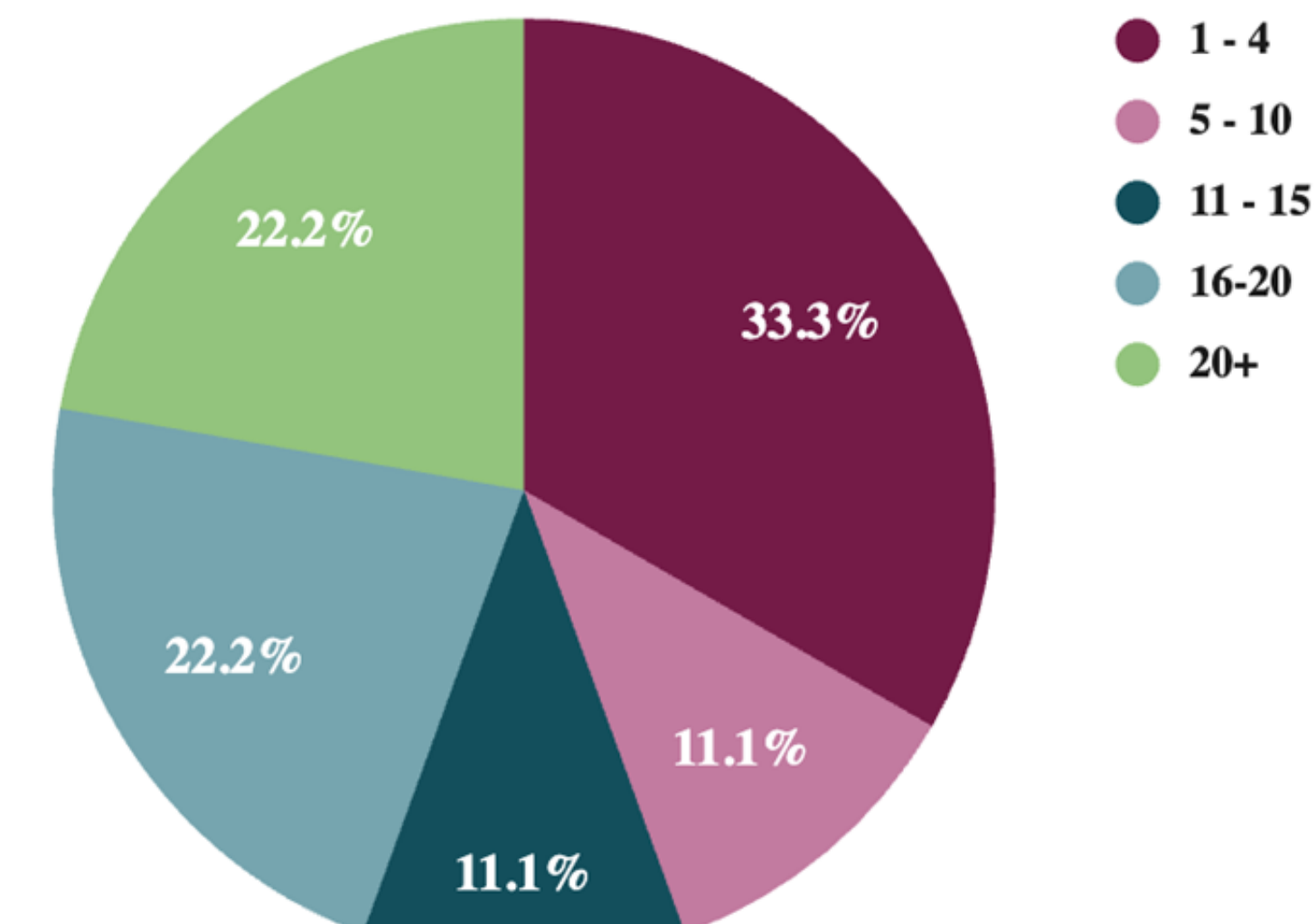
Providers asked to rank each response from best (1) to worst (3). ChatGPT 4.0 and the plastic surgery provider both out-ranked online provider responses. Additionally, ChatGPT responses outranked the plastic surgery provider responses, although this was not significant.

p-value * < 0.05; **** < 0.00001

Provider Age



Years in Practice



c. Selected Demographics for Survey Responders

FUTURE DIRECTIONS

- Integration of LLMs as “chat bots” into clinical practice
- Pre- and post-operative instructions, **personalized** to each surgeon
- Responding to common patient questions, **including image-based questions**, with responses that align with each provider’s preference and **match advice that they would typically provide**.



Want to learn more? Scan the QR code!

CONCLUSIONS

LLMs are able to provide **accurate, empathetic, and high quality** responses to patient questions, even with a **visual component**.

In terms of empathy, LLMs performed similarly to providers when addressing patient questions. This demonstrates that LLMs have evolved beyond an unfeeling, robotic interaction. Furthermore, using an LLM in practice will likely not negatively impact **levels of satisfaction**.

LLMs could be integrated into clinical practice to aid in **alleviating provider workload** and thereby **reduce burnout**

REFERENCES

- Almulhem JA, Aldekhyyel RN, Binkheder S, Temsah MH, Jamal A. Stress and Burnout Related to Electronic Health Record Use among Healthcare Providers during the COVID-19 Pandemic in Saudi Arabia: A Preliminary National Randomized Survey. *Healthcare (Basel)*. 2021 Oct 14;9(10):1367. doi: 10.3390/healthcare9101367. PMID: 34683047; PMCID: PMC8544434.
- Busch, F., Hoffmann, L., Rueger, C. *et al*. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med* 5, 26 (2025). <https://doi.org/10.1038/s43856-024-00717-2>
- Raghunathan R, Jacobs AR, Sant VR, King LJ, Rothberger G, Prescott J, Allendorf J, Seib CD, Patel KN, Suh I. Can large language models address unmet patient information needs and reduce provider burnout in the management of thyroid disease? *Surgery*. 2025 Jan;177:108859. doi: 10.1016/j.surg.2024.06.075. Epub 2024 Oct 18. PMID: 39424485.
- Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023 Jun 1;183(6):589-596. doi: 10.1001/jamainternmed.2023.1838. PMID: 37115527; PMCID: PMC10148230.