Title: Comparing English and Japanese ChatGPT Responses to Anesthesia-Related Medical Inquiries

Presenting author name: Kazuo Ando

Affiliation details of Presenting author Stanford University School of Medicine

Abstract:

Objective: The advancement of artificial intelligence (AI) integrated within large language models (LLMs) has the potential to enhance healthcare delivery efficiency. Despite the growing usage of LLMs, disparities in their effectiveness, especially across diverse languages, have not been thoroughly investigated. This research analyzes the performance of ChatGPT responses in English and Japanese concerning anesthesiology-related inquiries.

Method: Anesthesiologists proficient in both English and Japanese were enlisted as experts for this study. Ten frequently asked questions (FAQs) in the field of anesthesia were chosen and translated for evaluation. Expert evaluators assessed the responses from ChatGPT based on content quality indicators (such as accuracy, comprehensiveness, and safety) and communication quality parameters (including understanding, empathy/tone, and ethics).

Results: A total of eight anesthesiologists evaluated responses from English and Japanese LLMs. Overall, the quality of responses in English surpassed that of Japanese across all questions. Both content and communication quality were notably higher in English LLM responses compared to Japanese counterparts (both $p<0.001$). Measures of comprehensiveness and safety (pertaining to content quality; $p=0.0001$ and <0.001, respectively) and understanding (a communication metric; $p<0.001$) were superior in English LLM responses. Five out of the eight evaluators perceived English responses to be of higher quality than Japanese responses.

Conclusion: When evaluated by bilingual anesthesia experts, responses from English LLMs outperformed Japanese responses in addressing anesthesia-related FAQs. This study sheds light on potential language-related discrepancies in healthcare information dissemination and underscores the necessity for enhancing the quality of AI responses in underrepresented languages. Further research is warranted to investigate these differences in other commonly spoken languages and to compare LLM performance across various models.

Biography:

Anesthesiologist, Clinical Scientist