nature medicine

Article

https://doi.org/10.1038/s41591-024-02855-5

Adapted large language models can outperform medical experts in clinical text summarization

Received: 23 October 2023

Accepted: 2 February 2024

Published online: 27 February 2024

Check for updates

Dave Van Veen [1,2] ✉, Cara Van Uden [2,3], Louis Blankemeier [1,2], Jean-Benoit Delbrouck [2], Asad Aali [4], Christian Bluethgen [2,5], Anuj Pareek [2,6], Malgorzata Polacin [5], Eduardo Pontes Reis [2,7], Anna Seehofnerová [8,9], Nidhi Rohatgi [8,10], Poonam Hosamani [8], William Collins [8], Neera Ahuja [8], Curtis P. Langlotz [2,8,9,11], Jason Hom [8], Sergios Gatidis [2,9], John Pauly [1] & Akshay S. Chaudhari [2,9,11,12]

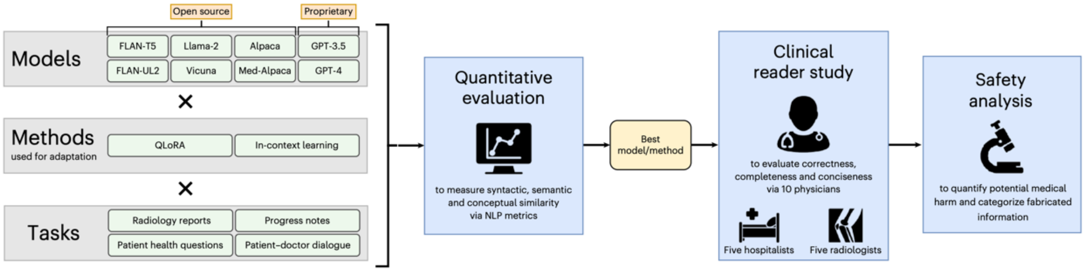
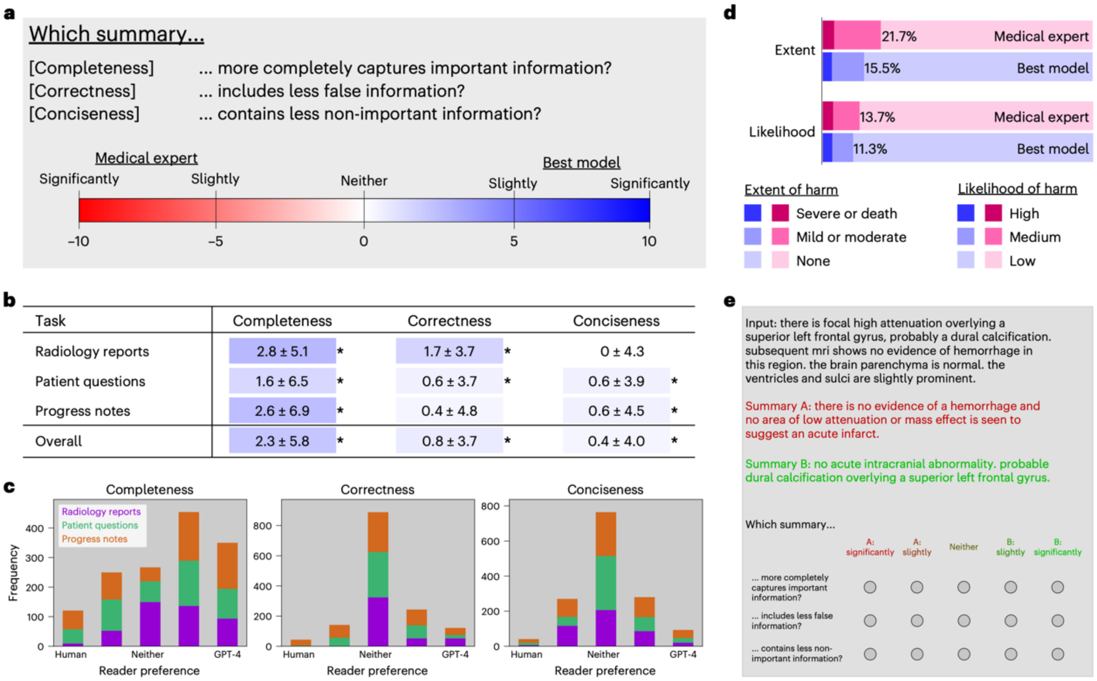


Fig. 1 | Framework overview. First, we quantitatively evaluated each valid combination (×) of LLM and adaptation method across four distinct summarization tasks comprising six datasets. We then conducted a clinical reader study in which 10 physicians compared summaries of the best model/method against those of a medical expert. Lastly, we performed a safety analysis to categorize different types of fabricated information and to identify potential medical harm that may result from choosing either the model or the medical expert summary.



Task	Completeness	Correctness	Conciseness
Radiology reports	2.8 ± 5.1 *	1.7 ± 3.7 *	0 ± 4.3
Patient questions	1.6 ± 6.5 *	0.6 ± 3.7 *	0.6 ± 3.9 *
Progress notes	2.6 ± 6.9 *	0.4 ± 4.8	0.6 ± 4.5 *
Overall	2.3 ± 5.8 *	0.8 ± 3.7 *	0.4 ± 4.0 *

Input: there is focal high attenuation overlying a superior left frontal gyrus, probably a dural calcification. subsequent mri shows no evidence of hemorrhage in this region. the brain parenchyma is normal. the ventricles and sulci are slightly prominent.

Summary A: there is no evidence of a hemorrhage and no area of low attenuation or mass effect is seen to suggest an acute infarct.

Summary B: no acute intracranial abnormality. probable dural calcification overlying a superior left frontal gyrus.

Which summary...

	A: significantly	A: slightly	Neither	B: slightly	B: significantly
... more completely captures important information?	○	○	○	○	○
... includes less false information?	○	○	○	○	○
... contains less non-important information?	○	○	○	○	○

Fig. 4 | Clinical reader study. a, Study design comparing summaries from the best model versus that of medical experts on three attributes: completeness, correctness and conciseness. **b**, Results. Highlight colors correspond to a value's location on the color spectrum. Asterisks (*) denote statistical significance by a one-sided Wilcoxon signed-rank test, $P < 0.001$. **c**, Distribution of reader scores for each summarization task across attributes. Horizontal axes denote reader preference as measured by a five-point Likert scale. Vertical axes denote frequency count, with 1,500 total cases for each plot. **d**, Extent and likelihood of possible harm caused by choosing summaries from the medical expert (pink) or best model (purple) over the other. **e**, Reader study user interface.