# Medical AI summaries often overstate findings despite precision in original research



Medical research is known for its rigorous and precise communication standards, emphasising the careful presentation of findings to match the scope and limitations of collected data. This discipline is grounded in a strict maxim: never to claim more than the evidence allows. However, recent studies reveal that this precision is often lost, not only in human summarisation but increasingly through the growing use of artificial intelligence (AI).

A comprehensive analysis published in The Conversation highlights how medical researchers typically hedge their claims. For example, a clinical trial on multiple myeloma treatment might report, “In a randomized trial of 498 European patients with relapsed or refractory multiple myeloma, the treatment increased median progression-free survival by 4.6 months, with grade three to four adverse events in 60 per cent of patients and modest improvements in quality-of-life scores, though the findings may not generalize to older or less fit populations.” Such detailed statements are precise but often complex and difficult for broader audiences to digest.

This complexity leads to simpler, more assertive summaries such as “The treatment improves survival and quality of life,” or “The drug has acceptable toxicity,” which lack critical qualifiers regarding patient populations, comparative measures, or conditions. Philosophers describe these as generics—statements that generalise without explicit detail or quantification. A systematic review of over 500 studies in top medical journals found that more than half included overgeneralizations beyond the specific populations studied, with over 80 per cent of these being generics, and fewer than 10 per cent providing any justification for such broader claims.

The tendency towards overgeneralisation reflects a deeper cognitive bias where both researchers and readers prefer simpler narratives amid complex data. This trend is now being amplified by large language models (LLMs) like ChatGPT, DeepSeek, LLaMA, and Claude, increasingly used by researchers, clinicians, and students to summarise scientific literature.

In a recent study testing 10 popular LLMs on their ability to summarise abstracts and articles from leading medical journals, nearly 5,000 AI-generated summaries were analysed. Results showed that some models produced over-generalizations in up to 73 per cent of summaries. A common error was transforming cautious claims such as “the treatment was effective in this study” into unqualified statements like “the treatment is effective,” which misrepresent the evidence's scope.

Human expert summaries were far less likely to include such broad generalizations, being nearly five times more precise compared to AI chatbots. Surprisingly, the latest models such as ChatGPT-4o and DeepSeek were prone to greater overgeneralisation, potentially due to training on already overgeneralized texts and reinforcement learning favouring confident, concise answers preferred by users.

These findings raise concerns about the risks of miscommunication, particularly in medical research where nuances relating to patient population, treatment effect size, and uncertainty critically influence clinical decisions. A global survey involving nearly 5,000 researchers found that almost half had integrated AI into their research workflows, and 58 per cent believed AI currently outperforms humans in literature summarisation. However, this study highlights the limitations of that optimism.

Addressing these challenges calls for multiple responses. Journal editorial policies can enforce clearer guidelines on reporting and interpreting data to reduce overgeneralized claims. Researchers should carefully select LLMs for summarisation, favouring models demonstrated to maintain contextual accuracy, such as Claude in this study. AI developers can incorporate prompts within LLMs that encourage cautious, precise language when summarising research findings. The methodology developed in this research offers a benchmark framework to assess the overgeneralisation propensity of AI tools before their broad application.

As medical research communication increasingly involves AI, maintaining precision in language—including nuanced expressions of data scope and uncertainty—remains essential. Both human researchers and AI systems share a tendency to overstate findings when summarising complex data. Elevating standards for both groups is necessary to ensure that medical evidence is accurately conveyed, supporting appropriate clinical use and patient care based on the true applicability of research outcomes.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.frontiersin.org/articles/10.3389/fmed.2024.1477898/full> - This study examines how large language models (LLMs) interpret complex medical information, highlighting instances where LLMs produce overgeneralized summaries, which can lead to misinterpretations of medical data.
2. <https://www.jmir.org/2024/1/e56614/> - This research explores the effectiveness of LLMs in extracting relevant information from medical records, revealing challenges in maintaining precision and accuracy, which can result in overgeneralized claims.
3. <https://journals.sagepub.com/doi/full/10.1177/20552076241293932> - This pilot study assesses the performance of LLMs in summarizing palliative care consultations, uncovering issues related to overgeneralization and the loss of critical qualifiers in AI-generated summaries.
4. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10723541/> - This systematic review discusses the challenges of maintaining factual accuracy in medical AI systems, emphasizing the tendency of LLMs to produce overgeneralized information that may not align with the evidence.
5. <https://arxiv.org/abs/2305.06299> - This paper evaluates the ability of GPT-3 to summarize and synthesize biomedical articles, highlighting instances where the model produces overgeneralized summaries that may not accurately reflect the original research findings.
6. <https://www.frontiersin.org/articles/10.3389/frai.2024.1477535/full> - This study compares the summarization capabilities of LLMs and medical students, revealing that AI-generated summaries often lack the depth and nuance found in human-generated content, leading to overgeneralizations.
7. <https://news.google.com/rss/articles/CBMiqgFBVV95cUxObXZFdXhWWGNteDBWeUw3djBDOXY1WFZzcXVIX1B0SjVBdFh5SXJzTkxZeFpfYnFIdTdMN3V0SGZKT3pjdHIyX29acG5tbG81M19PdmJ2UkhhOERwWUZUZ1FFU3N3UXdUalczMkhxVC1RRW9fdjNhVUlkZHA3R0JPWEhfRnlHLVRTQlVLVzBwcTZhS3BHeWhlUXhJTjRkc3ZaQmFMQ2JtNENqUQ?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data