LLM tumor board summaries: what a four-institution expert study found

26 November 2026 · 4 min read · blogs.iictp.org/llm-tumor-board-case-summaries-recommendations-study

Short answer: large language models produced tumor board case summaries that clinicians rated moderately to highly, but their treatment recommendations scored clearly lower, in a study published online in npj Digital Medicine on 24 November 2026. Using data from four institutions, the authors report mean summary ratings of 3.57 to 4.59 out of 5 and recommendation ratings of 2.0 to 3.6. It matters because it separates a drafting aid from decision support.

TL;DR The short version

- A four-institution study in npj Digital Medicine on 24 November 2026 had experts rate LLM tumor board case summaries and treatment recommendations.
- Summaries scored 3.57 to 4.59 out of 5, but recommendations scored clearly lower, at 2.0 to 3.6.
- The results support LLMs as a drafting aid for summaries under clinician review, not as decision support for treatment.

Read online, listen to it, or change the text size: https://blogs.iictp.org/llm-tumor-board-case-summaries-recommendations-study

What did the study test?

A tumor board is a recurring meeting where specialists such as oncologists and pathologists discuss patient care. Wen-wai Yim and colleagues defined three AI tasks: summarising a case, generating options and predicting the meeting outcome. They evaluated models from the DeepSeek, GPT, Gemini and Qwen families.

How the evaluation worked

- Datasets came from four medical institutions.
- Medical oncologists wrote about 9,600 reference texts and made about 16,000 human judgments.
- Experts graded outputs directly and on detailed criteria.
- The authors compared those grades with automatic metrics, including using a model as judge and their own metric family called TBFact.

What is new compared with earlier work

The authors describe it as, to their knowledge, the largest expert-grading study of tumor board tasks with both granular and overall assessments. It also tests whether automatic scores agree with experts, which many earlier comparisons assumed.

Results reported by the authors

- Case summaries: average clinician ratings of 3.57 to 4.59 out of 5 across institutions.
- Recommendations: average ratings of 2.0 to 3.6.
- Agreement between model-as-judge metrics and clinicians was modest for summary subtasks (correlations around 0.2) and higher for recommendations (around 0.6).
- TBFact metrics were competitive with model-as-judge approaches for longer free text.
- Experts often rated answers highly even when they differed substantially from the reference text, and criterion scores related to overall judgment in a non-linear way.

Limitations and open questions

- This article relies on the abstract; read the full paper for data selection, cancer types, model versions and rating definitions.
- A rating is not a patient outcome. The abstract reports no prospective use in live meetings.
- Ratings varied between institutions, so one hospital’s result may not carry to another.
- The author list includes Microsoft Health AI staff and academic centres; one author declares external financial relationships described as unrelated to the study. Sources were read on 26 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to pilot summaries responsibly

1. Start with summaries only, drafted before the meeting and checked by the presenting clinician.
2. Agree the required facts: diagnosis, stage, pathology, imaging, prior treatment and open question.
3. Score drafts for missing and wrong facts, not only readability.
4. Keep model recommendations hidden from the board until a separate evaluation supports showing them.
5. Record corrections and review them monthly.

What it means for oncologists and hospital informatics

The gap between summary and recommendation scores is the useful signal. Assembling the case is where a draft may save time; choosing treatment still needs the board. Be cautious about vendor dashboards that report automatic quality scores for summaries, since the weak agreement reported here suggests they may not reflect clinical judgment.

Frequently asked questions

- **Can LLMs write tumor board summaries?** In this study clinicians rated them 3.57 to 4.59 out of 5 on average, so they still need checking.
- **Can LLMs recommend cancer treatment?** Recommendation ratings were lower, 2.0 to 3.6; the study does not support using them without expert decision.
- **Are automatic metrics reliable here?** Agreement with experts was modest for summaries.
- **Where was it published?** npj Digital Medicine, online on 24 November 2026.

Related reading

- Ambient AI scribe note quality
- WARD: longitudinal clinical agents
- How to evaluate an LLM application
- AI Practitioner certification

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Yim et al. Challenges in AI based tumor board case summarization and recommendations. npj Digital Medicine. Published online 24 November 2026. nature.com/articles/s41746-026-03442-3

Published by IICTP, an independent technical credential body, on 26 November 2026. The current version of this post is at https://blogs.iictp.org/llm-tumor-board-case-summaries-recommendations-study. More guides: blogs.iictp.org.