The St. Gallen AI Consensus – Should AI have a vote?

A. Oseledchyk[1], P. Weber[2], B. Kasenda[1]

[1] Medical Oncology, University Hospital Basel, Basel, , Switzerland; [2] Breast Cancer Center, University Hospital Basel, Basel, Switzerland.

PD11-09
San Antonio Breast Cancer Symposium
December 9-12, 2025

Background

- Breast cancer research generates vast, heterogeneous trial data leading to therapeutic uncertainty
- St. Gallen consensus conferences provide expert guidance where evidence is ambiguous [1]
- Expert panels remain subject to cognitive biases and personal clinical experiences
- Large language models (LLMs) may offer potential to provide comprehensive evidence synthesis

Methods

Study design

- Cross-sectional study of four frontier LLMs (Claude, GPT-5 Thinking, Gemini, DeepSeek) on 81 selected St. Gallen 2025 questions. Use of identical standardized prompts with evidence restricted to publications through February 28, 2025, one pass per question.
- Outcomes: first-choice agreement with expert top answer choice; top-2 agreement (only in question with ≥3 options).
- Expert opinion quantified as normalized consensus strength to adjust for varying number of answer options (0-100 scale). Mixed-effects logistic regression to identify predictors of LLM-expert agreement
- Qualitative review of "no AI consensus" questions (≤2/4 LLMs matching expert first choice) and categorization of LLM performance as negative, neutral, or positive divergence

Objective

- To investigate concordance between LLM recommendations and expert consensus at the St. Gallen International Breast Cancer Conference 2025 [1]

Results

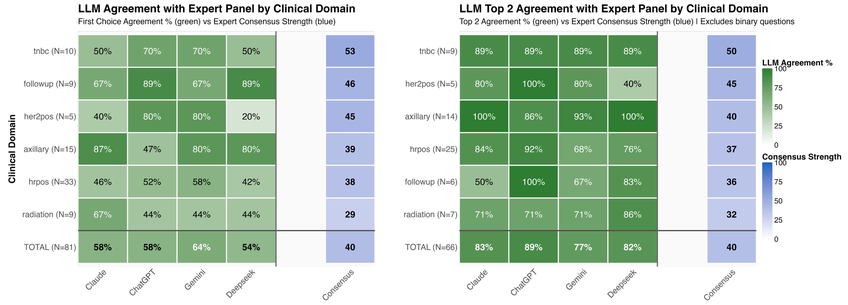


Figure 1. Concordance between LLMs and the St. Gallen 2025 expert panel, in total and by clinical domain. First-choice agreement (left) and Top-2 agreement (right, Top-2 excludes binary questions).

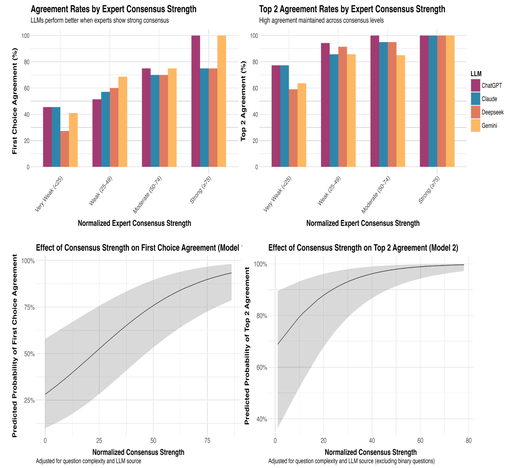


Figure 2. Association between normalized consensus strength and first-choice agreement (left) and top-2 agreement (right)

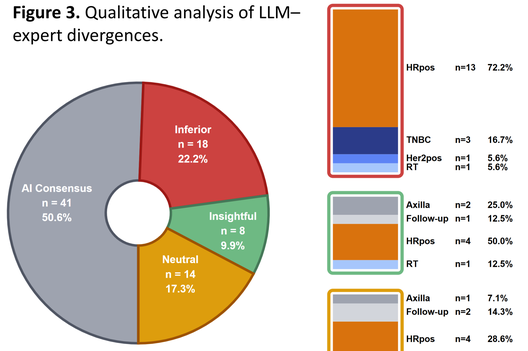


Figure 3. Qualitative analysis of LLM–expert divergences.

QR-Code - Link to Interactive Table of Panel and LLM voting

- First-choice agreement ranged 20-89% by clinical domain with an average of 58 – 64% across all domains; top-2 agreement reached 40-100% by domain and 77% - 89 % of total agreement rate **(Figure 1)**
- Expert consensus strength emerged as the only predictor of LLM concordance (OR 1.04 per 1-point increase, 95% CI: 1.02-1.06, p<0.001), demonstrating LLMs amplify prevailing evidence signals **(Figure 2)**
- Among 40 "no AI consensus" questions: 22% showed negative divergence (knowledge gaps), 17% neutral divergence (defensible alternatives in ambiguous scenarios), and 10% positive divergence (evidence-consistent alternatives) **(Figure 3)**
- Positive divergence examples included: appropriate de-escalation of axillary radiotherapy, stricter anthracycline indications based on trial evidence and use of neoadjuvant endocrine therapy
- Detailed table of all 81 questions, expert votes and LLM choices and rationale are summarized in an interactive table **(QR code;** link**:** https://sites.google.com/view/st-gallen-ai-consensus- 1/1 **)**

Conclusions

- LLM agreement is increasing when expert panels' consensus is strong
- LLMs function as "evidence amplifiers" rather than creative reasoners
- AI-supported discussions need to be studied prospectively to establish the role of AI-augmentation in expert panels

References:
[1] Ditsch, N., et al. Breast Care, 2025. 20(4): p. 288-297.

Funding: This work was supported with grants by Geistlich-Stucki Stiftung.

Poster presented at St. Antonio Breast Cancer Conference (SABCS) 2025
Corresponding author: Anton Oseledchyk, anton.oseledchyk@usb.ch

This presentation is the intellectual property of the author/presenter. Contact them at anton.oseledchyk@usb.ch for permission to reprint and/or distribute.