gutflix
GI learning, personalised

Summary

REASONING CAPABILITIES OUTPERFORM LITERATURE EXPOSURE: EVALUATING LARGE LANGUAGE MODELS ON GASTROENTEROLOGY BOARD QUESTIONS

Cem Simsek

UEG Week Berlin 2025 Poster

October 7, 2025

Reasoning-optimized large language models outperform non-reasoning models on European Board of Gastroenterology and Hepatology examination questions, with performance gains attributable to explicit reasoning capabilities rather than larger literature exposure.

KEY POINTS

- Reasoning-enhanced ChatGPT-o1 answered 171 of 203 EBGH questions correctly compared to 164 for ChatGPT-4o, rescuing 16 items that the non-reasoning model missed.
- Non-reasoning model accuracy showed statistically significant correlations with publication metrics (total publications r=0.837, p=0.026; full-text count r=0.823, p=0.033), while reasoning models showed weaker correlations (total publications r=0.312, p=0.566).
- Among previously failed questions retested in January 2025, ChatGPT-o1 Pro solved 15 of 17 remaining items, with 6 uniquely answered by this model and only 2 stumping all tested models.
- Reasoning models consistently detected tacit clinical cues and offered broader diagnostic or therapeutic reasoning in qualitative review.
- The authors conclude that explicit reasoning capabilities, rather than deep-search over larger corpora, drive performance on challenging gastroenterology board questions and may support domain-agnostic clinical decision-making.

VIEW THIS CONTENT ON GUTFLIX

https://gutflix.eu/search/d/8dfaecb4-a813-11f0-a617-0242ac140006

AI-generated summary

This summary was generated by an AI large language model based on the content transcript. It is for informational purposes only and should not be considered a substitute for clinical judgment. Always rely on your professional expertise and the full clinical context when making clinical decisions.