.

“Rapid improvement in ability of AI to reason using clinical guidelines” Çalışma Değerlendirmesi

Dr. Halil Siner

“Rapid improvement in ability of AI to reason using clinical guidelines” Çalışma Değerlendirmesi

Hazırlayan: Dr. Halil Siner
Araştırma Görevlisi Doktor, Afyonkarahisar Sağlık Bilimleri Üniversitesi Tıp Fakültesi, Kardiyoloji Anabilim Dalı

Çalışmanın adı: Rapid improvement in ability of AI to reason using clinical guidelines

Çalışmanın amacı ve metodolojisi: Büyük dil modelleri (LLM’ler), 2022’deki lansmanından sonraki iki ay içinde ChatGPT ile 100 milyondan fazla kullanıcının etkileşime girmesiyle hızlı bir şekilde yaygınlaşmıştır. Potansiyellerine rağmen, LLM’ler klinik pratikte akıl yürütme gerektiren görevlerde zorlanmıştır. 2024 yılındaki son gelişmeler, pekiştirmeli öğrenme yoluyla akıl yürütme yeteneği için optimize edilmiş OpenAI o1-preview gibi daha güçlü modelleri tanıtmıştır. Bu çalışmada, en yeni LLM’lerin güncel klinik kılavuzları kullanarak klinik akıl yürütme becerilerini anlamlı ölçüde geliştirip geliştirmediğini incelenmektedir.

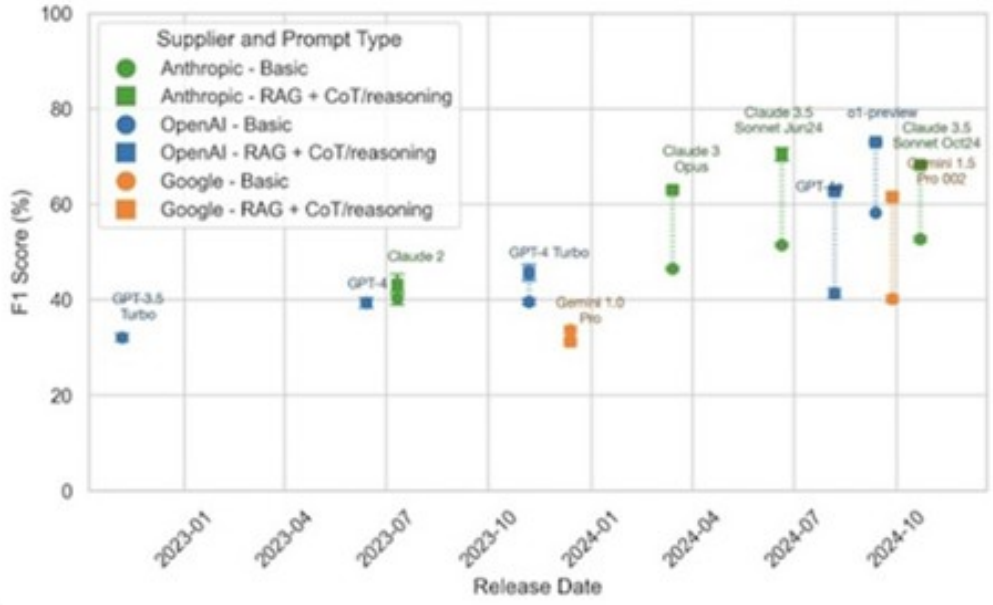
Çalışmanın sonuçları ve tartışma: Thomas ve ark.'nın çalışması, LLM’lerin klinik kılavuzlara uygun tedavi önerisi sunma performansının yalnızca bir yıl içinde %45,7'den %73,0'a yükselmesiyle dramatik bir sıçrama kaydettiğini ortaya koymaktadır.1 Bu gelişim, izole vakalar söz konusu olduğunda modellerin genel pratisyen hekimlerle kıyaslanabilir bir doğruluk düzeyine ulaştığını düşündürmektedir. Benzer biçimde, Roechl ve ark.'nın 2025 yılında JMIRx Med'te yayımlanan retrospektif çalışması, GPT-4 tabanlı modellerin gerçek dünya verilerinde kılavuz uyumlu karar alma süreçlerini büyük ölçüde otomatize edebildiğini, ancak bunun henüz mükemmeliyetten uzak olduğunu vurgulamaktadır.2 Öte yandan, RAG (Retrieval-Augmented Generation) destekli GPT-4 sistemlerinin tümör kurulu kararlarıyla %84 oranında tam uyum sağladığını bildiren çalışmalar da bu ivmenin farklı tıbbi alanlara yayıldığını teyit etmektedir.3

Ancak bu umut verici tablonun önemli sınırlılıkları bulunmaktadır. Thomas ve ark.'nın da dikkat çektiği üzere, çoklu komorbiditeleri olan kompleks vakalarda performans belirgin şekilde düşmektedir; bu durum yalnızca kardiyoloji alanına özgü değildir. Scientific Reports'ta yayımlanan bir çalışmada mevcut LLM'lerin esnek olmayan akıl yürütme nedeniyle zincir düşünce (chain-of-thought) gerektiren senaryolarda ciddi hatalar ürettiği ve sistematik bir şekilde erken hipotez kilitleme eğilimi sergilediği raporlanmıştır.4 NEJM AI'da yayımlanan bir kıyaslama çalışması ise o1, Gemini, Claude ve DeepSeek gibi en gelişmiş modellerin bile klinisyenlerin gerisinde kaldığını; özellikle sezgisel istatistiksel örüntü tanıma ve istisna yönetimi gerektiren durumlarda belirgin biçimde yetersiz kaldığını ortaya koymuştur.5

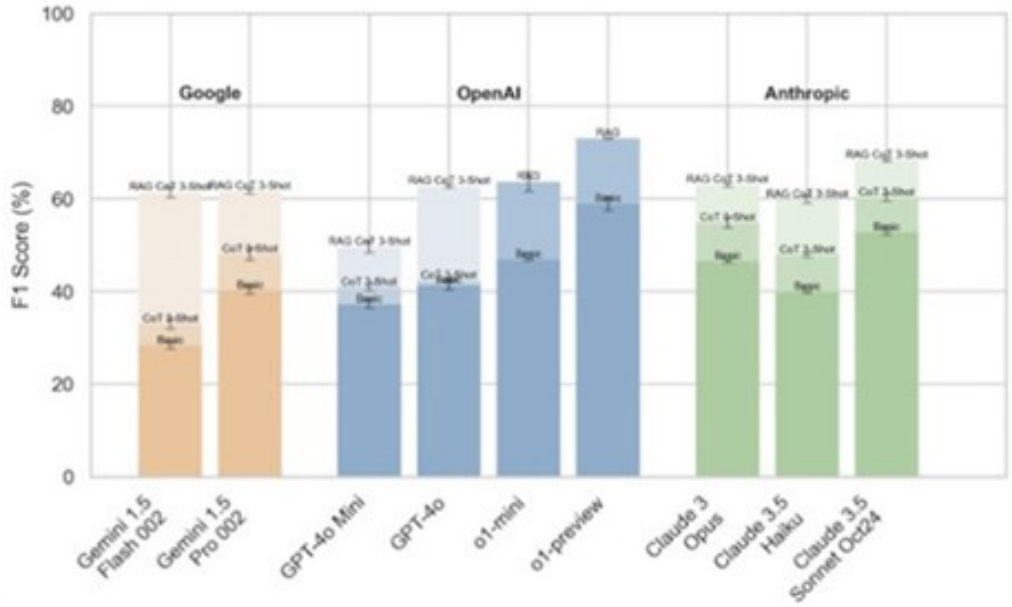
Erişim linki: https://academic.oup.com/ehjdh/article/7/Supplement_1/ztaf143.056/8422997

Kaynakça:

1. Khattak, S., et al., *Rapid improvement in ability of AI to reason using clinical guidelines.*European Heart Journal - Digital Health, 2026. 7(Supplement_1).
2. Roeschl, T., et al., *Assessing the Limitations of Large Language Models in Clinical Practice Guideline-Concordant Treatment Decision-Making on Real-World Data: Retrospective Study.* JMIRx Med, 2025. 6: p. e74899.
3. Abdullayev, N., et al., *European guideline informed RAG-based GPT-4 decision support tool in tumor board meetings for breast cancer treatment.* Eur J Surg Oncol, 2025. 51(11): p. 110384.
4. Kim, J., et al., *Limitations of large language models in clinical problem-solving arising from inflexible reasoning.* Sci Rep, 2025. 15(1): p. 39426.
5. McCoy, L., et al., *Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study.* NEJM AI, 2025.

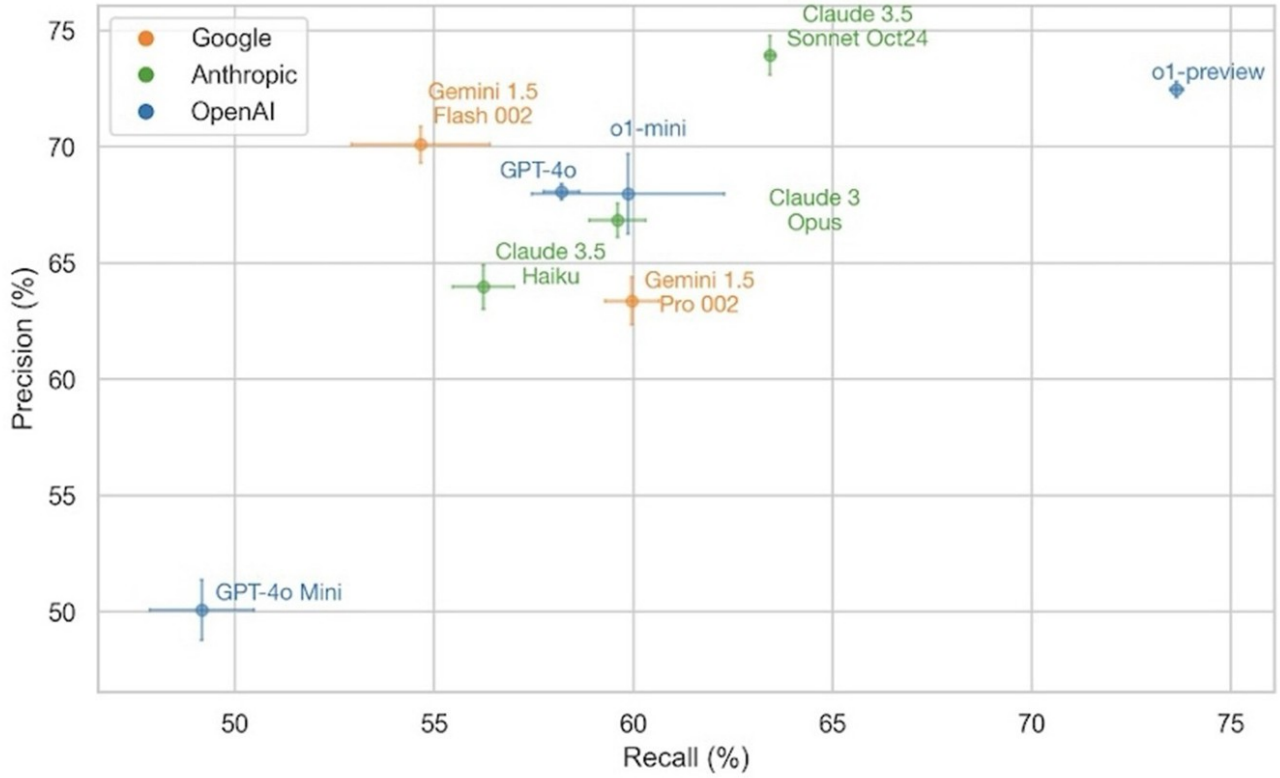


A



B

Improvement over time of LLMs for recommending medications based on clinical guidelines using retrieval augmented generation (RAG). F1 score of the most powerful LLMs released at different timepoints from Anthropic, OpenAI and Google (1A). Dotted line indicates the improvement in F1 score when the model used an advanced prompt (RAG CoT 3-shot, squares) that used RAG to include the latest guidelines compared to a basic prompt that did not contain the latest guidelines using RAG (basic prompt, circles). F1 score of the most recent LLMs from Anthropic, OpenAI and Google, showing improvement in performance with prompts that utilise CoT 3-shot and RAG in the majority of models (1B). Of note, we did not use CoT 3-shot prompts with the OpenAI o1 LLMs as OpenAI recommend against this as it can interfere with the advanced reasoning process. Instead these models used a basic prompt with or without RAG. Overall F1 score calculated for all 201 clinical scenarios for each prompt type for each model in triplicate. Error bars represent 95% confidence intervals.



Comparison of the latest LLMs from OpenAI, Anthropic and Google for recommending medications based on clinical guidelines. Precision and recall using the most advanced prompts available (RAG CoT 3-shot) for all models apart from o-1 preview and o-1 mini which used RAG without CoT or multi-shot learning as advised by OpenAI. Overall recall and precision calculated for all 201 clinical scenarios in triplicate. Error bars represent 95% confidence intervals.