Prompt Engineering: Harnessing Generative AI for HEOR

Dr Rachael L. Fleurence*, PhD, MSc
Head of Evidence and AI Solutions
Value Analytics Labs
May 14, 2025

*Formerly at the National Institutes of Health, Bethesda, MD, USA

What is Generative AI ?

Generative AI (Gen AI)

AI systems that generate text, images, or other content based on input data, often creating new and original outputs.

Large Language Models (LLMs)

A subset of foundation models trained on extensive text data to perform tasks like recognizing, summarizing, translating, predicting, and generating text based on large datasets.

How Large Language Models (LLMs) Use Context

- **Context Retention:** LLMs generate relevant and coherent responses by retaining information from earlier inputs.
- **Example:** If a conversation includes *clinical trial data*, the model continues to generate responses within a healthcare framework.
- **Analogy:** Similar to a thoughtful colleague who remembers previous questions and responds in a consistent, context-aware manner.

How Good is Gen AI ?

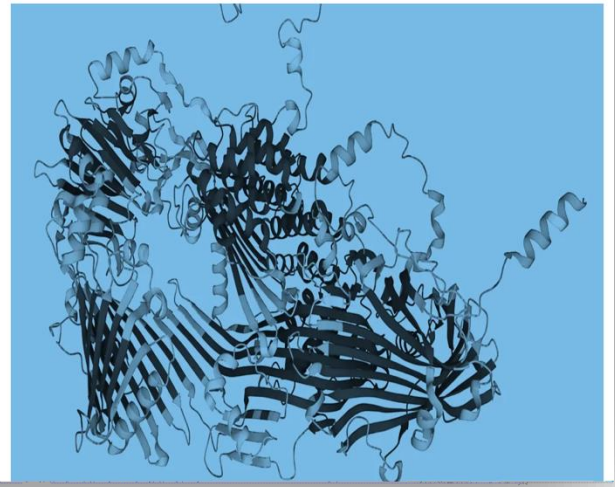
Very good !

The Economist ☰ Menu | Weekly edition | The world in brief | Search

Podcasts | Babbage

How artificial intelligence cracked biology's biggest problem

Our podcast on science and technology. This week, we examine how DeepMind's AI system predicted the structure of virtually every known protein—and what the breakthrough means for both science and machine learning



The Economist ☰ Menu | Weekly edition | The world in brief | Search

Science & technology | The 2024 Nobel prizes

AI wins big at the Nobels

Awards went to the discoverers of micro-RNA, pioneers of artificial-intelligence models and those using them for protein-structure prediction

Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]

And Improving Rapidly

Generate an image of a three-legged stool.

September 2024

May 2025

But Prompting Language is Important ...

Generate an image of full glass of red wine.

Can you fill it to the brim ?

Applications of Generative AI in HEOR

ISPOR REPORT · Volume 28, Issue 2, P175-183, February 2025

Download Full Issue

Generative Artificial Intelligence for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations: An ISPOR Working Group Report

Rachael L. Fleurence, PhD [1] · Jiang Bian, PhD [2,3,4] · Xiaoyan Wang, PhD [5,6] · Hua Xu, PhD [7] · Dalia Dawoud, PhD [8,9] · Mitchell Higashi, PhD [10] · Jagpreet Chhatwal, PhD [11] on behalf of the ISPOR Working Group on Generative AI

Key Applications of Generative AI in HEOR

- **Systematic Literature reviews**: Automation of search terms, abstract screening, data extraction, and code generation for meta-analyses.
- **Health economic modeling**: Conceptualization, validation, parameterization, country adaptations.
- **Real World Evidence**: Analysis of large unstructured data from clinical notes and imaging.
- **Dossier Development**: Outline, Drafting.

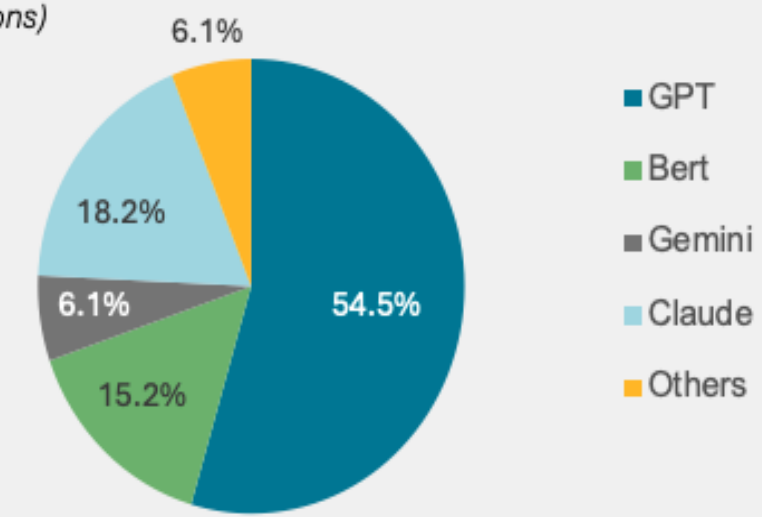
A Few Examples

A Review of the Uses of Gen AI for SLRs

LLMs Assessed

Among all LLMs, GPT was the most frequently evaluated LLM across studies (Figure 2).

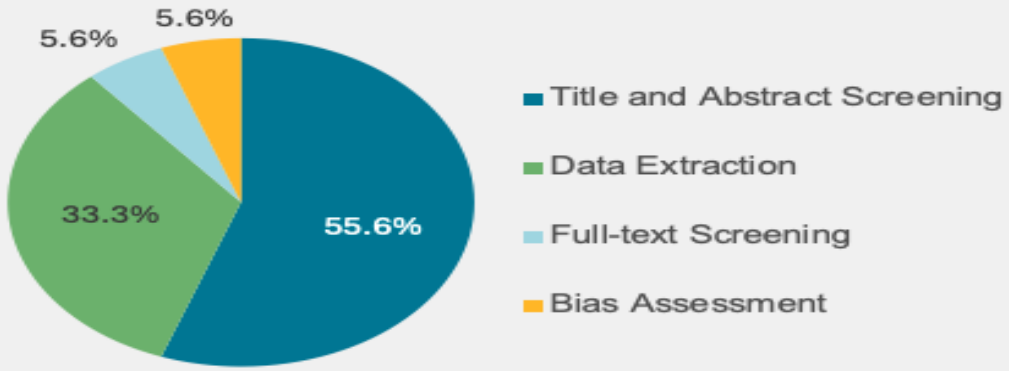
Figure 2: *Distribution of LLMs Evaluated (among 33 independent evaluations)*



Performance Evaluation

Among all SLR steps, title and abstract screening using LLMs was the most frequently evaluated one across studies (Figure 3).

Figure 3: *Distribution of SLR Tasks Evaluated (among 18 independent evaluations)*



LLMs demonstrated strong performance across various systematic review tasks (Table).

- **Title screening:** Sensitivity 94.3%–96.2%; specificity 85.5%–99.6%.
- **Abstract screening:** Accuracy 80%–97.5%; sensitivity 62%–95%; specificity 65%–98.7%.

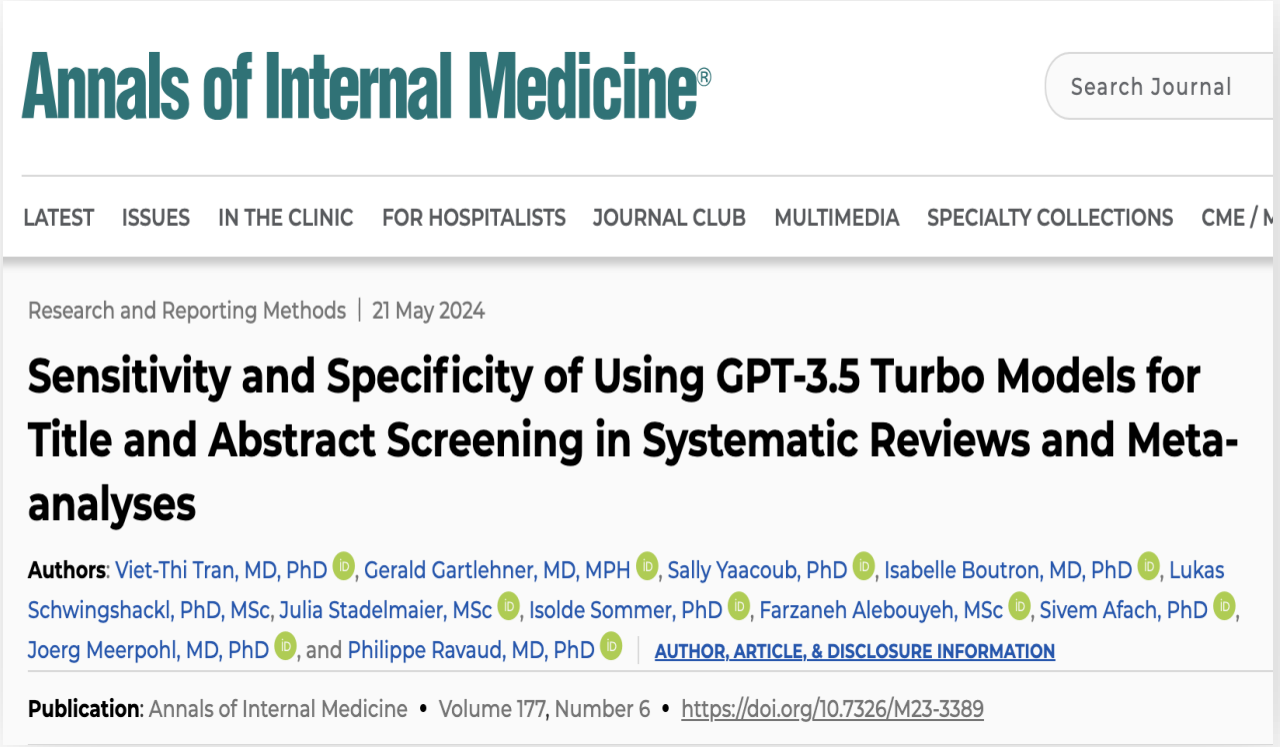
Reference: Chhatwal et al. Assessing the Effectiveness of Large Language Models in Automating Systematic Literature Reviews: Findings from Recent Studies. ISPOR Montreal 2025

Example: Automating title/abstract screening

SLR

- **Results:** Sensitivities ranged from **81.1% to 96.5%** and specificities ranged from **25.8% to 80.4%.**

- **Conclusion**: GPT-3.5 Turbo model may be used as a **second reviewer** for title and abstract screening



Annals of Internal Medicine®

Search Journal

LATEST ISSUES IN THE CLINIC FOR HOSPITALISTS JOURNAL CLUB MULTIMEDIA SPECIALTY COLLECTIONS CME / M

Research and Reporting Methods | 21 May 2024

Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses

Authors: Viet-Thi Tran, MD, PhD, Gerald Gartlehner, MD, MPH, Sally Yaacoub, PhD, Isabelle Boutron, MD, PhD, Lukas Schwingshackl, PhD, MSc, Julia Stadelmaier, MSc, Isolde Sommer, PhD, Farzaneh Alebouyeh, MSc, Sivem Afach, PhD, Joerg Meerpohl, MD, PhD, and Philippe Ravaud, MD, PhD | AUTHOR, ARTICLE, & DISCLOSURE INFORMATION

Publication: Annals of Internal Medicine • Volume 177, Number 6 • https://doi.org/10.7326/M23-3389

Tran VT et al. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. *Ann Intern Med*. Jun 2024;177(6):791-799. doi:10.7326/m23-3389

HEM

Developing Economic Models Using Gen AI

ABSTRACT ONLY · Volume 27, Issue 12, Supplement , S7, December 2024

P28 Development of De Novo Health Economic Models Using Generative AI

J Chhatwal [1] · IF Yildirim [2] · S Samur [3] · E Bayraktar [2] · T Ermis [2] · T Ayer [4]

Affiliations & Notes ⌄ Article Info ⌄

Some Limitations

- Accuracy
- Comprehensiveness
- Factuality
- Robustness
- Reproducibility
- Privacy and Security

Reference: Fleurence et al. ISPOR Working Group on Gen AI. ArXiv, 2024

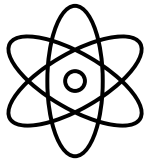
Final Thoughts

User Expertise vs. Model Performance

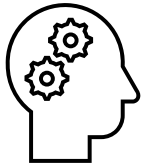
Generate an image of full glass of red wine.

Can you fill it to the brim ?

Conclusions



Early applications of Gen AI in HEOR show a lot of **promise**



User expertise and **model performance** are improving in tandem

A **learning community** is important on this journey – consider joining the **ISPOR AI Community of Interest** !

Join the Artificial Intelligence Community of Interest

How to Join:

1. Scan the code.
2. Login with your email and ISPOR password.
3. Select “Community of Interest - Artificial Intelligence.”
4. Click Save.

The AI COI is open to all ISPOR members.

*The Artificial Intelligence Community of Interest is currently seeking volunteers for **Community Engagers** for their Online Community – Sign Up Today!*