Generative AI: Applications to Systematic Literature Reviews, Evidence Synthesis and RWE

Rachael L. Fleurence, PhD

Senior Advisor, National Institutes of Health

November 19, 2024

ISPOR Europe, Barcelona Spain

Applications of Generative AI to SLRs: Outline

1. Applications in SLR, ES and RWE
2. Overview of limitations
3. NICE position statement
4. LLM evaluation framework for HEOR needed


Generative AI for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations

Rachael L. Fleurence, PhD, MSc[1], Jiang Bian, PhD[2 3 4], Xiaoyan Wang, PhD[5 6], Hua Xu, PhD[7], Dalia Dawoud, PhD[8 9], Mitch Higashi, PhD[10], Jagpreet Chhatwal, PhD[11 12]

Fleurence et al. https://arxiv.org/abs/2407.11054

Before we get into the details…

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

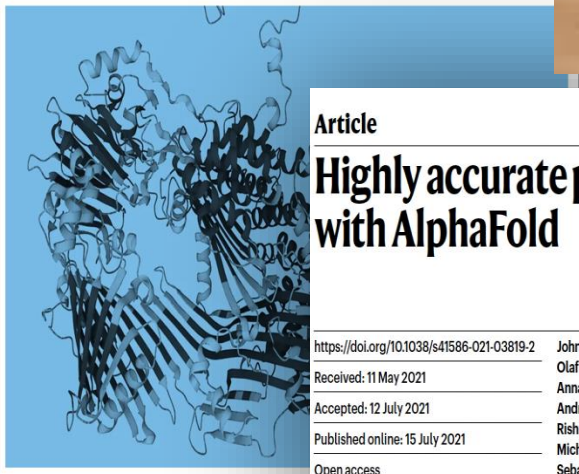


IMAGE: DEEPMIND

Aug 2nd 2022

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]

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

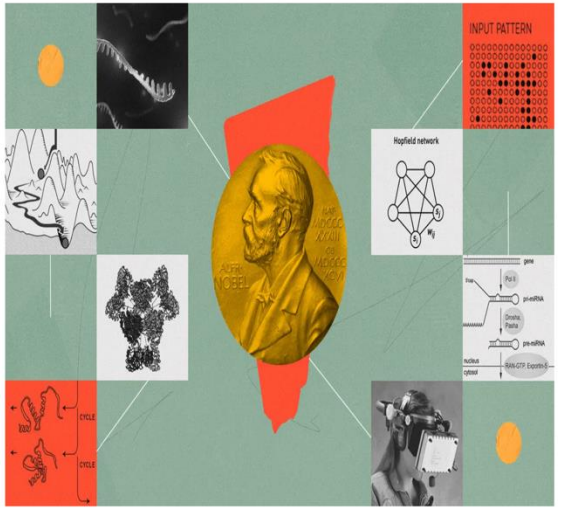


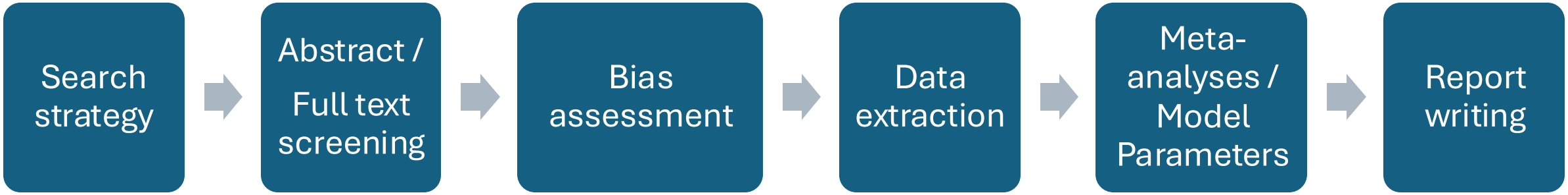
ILLUSTRATION: JAVIER PALMA

Oct 10th 2024

Share

Applications of Generative AI in SLRs

- SLRs are **time-consuming** and **labor-intensive** (6-18 months, FTEs)



Enhancing Search Strategies

Search strategy

- **Capabilities:**
  - Can propose MeSH terms and keywords for biomedical search engines (e.g., PubMed).
- **Challenge:**
  - **"Hallucinations"**: Risk of fabricated citations, requiring manual verification or advanced techniques (e.g. RAG)

J Am Soc Nephrol. 2023 May 31;34(8):1302–1304. doi: 10.1681/ASN.0000000000000166

Retrieve, Summarize, and Verify: How Will ChatGPT Affect Information Seeking from the Medical Literature?

Qiao Jin [1], Robert Leaman [1], Zhiyong Lu [1,✉]

Author information Article notes Copyright and License information

PMCID: PMC10400098 PMID: 37254254

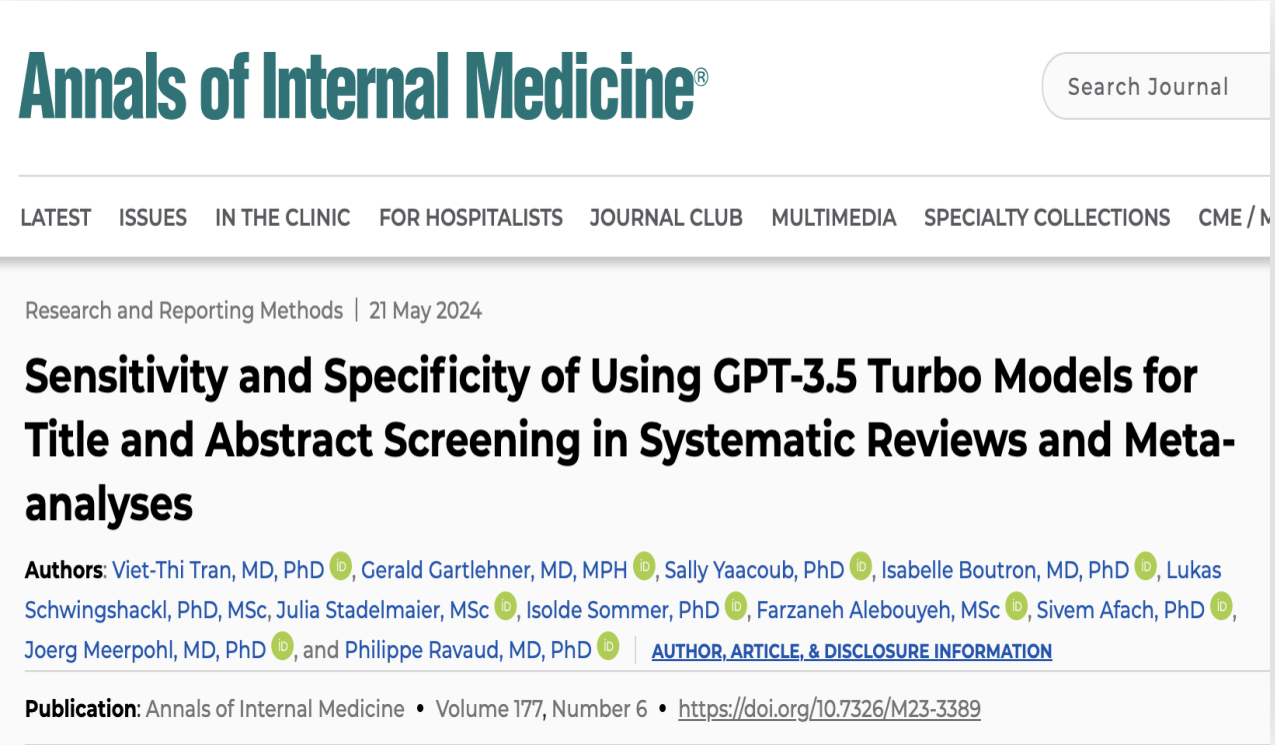
Hallucination = An incorrect output produced by a generative AI model that is not based on the input data or reality. This content is factually incorrect, misleading, or fabricated.

Jin Q, Leaman R, Lu Z. Retrieve, Summarize, and Verify: How Will ChatGPT Affect Information Seeking from the Medical Literature? *J Am Soc Nephrol*. Aug 1 2023;34(8):1302-1304. doi:10.1681/ASN.0000000000000166

Abstract screening

Automating abstract screening

- **Aim:** Study investigated the sensitivity and specificity of GPT-3.5 Turbo as a single reviewer, for title and abstract screening in systematic reviews.
- **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

Bias Assessment

Bias Assessment

- Study assessed **2 LLMs (ChatGPT and Claude)** and had **3 experts** assessing 30 RCTs, using a structured prompt to assess Risk Of Bias Assessment
- **High accuracy** rates for both LLMs **(>84.5%),** compared with human reviewers, across 10 specific domains.
- Findings suggest LLMs have **substantial accuracy** in assessing ROB in RCTs.


JAMA Network Open

Original Investigation | Statistics and Research Methods

Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models

Honghao Lai, MM; Long Ge, MD; Mingyao Sun, MSN; Bei Pan, MD; Jiajie Huang, MSN; Liangying Hou, MD; Qiuyu Yang, MD; Jiayi Liu, MM; Jianing Liu, MSN; Ziying Ye, MM; Danni Xia, MM; Weilong Zhao, MM; Xiaoman Wang, MD; Ming Liu, MD; Jhalok Ronjan Talukdar, PhD; Jinhui Tian, MD; Kehu Yang, MD; Janne Estill, PhD

Lai H, Ge L, Sun M, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. *JAMA Netw Open*. May 1 2024;7(5):e2412687.

Using LLMs Data Extraction

Data extraction

- **High Accuracy**: Can be effective in replicating data extraction tasks.
- **Case Studies**:
  - **Gartlehner et al.:** LLM reached **96.3% accuracy** in data extraction compared to human reviewers.
  - **Reason et al.**: Achieved over **99% accuracy** in replicating data extraction from 4 network meta-analysis.
- **Challenges**:
  - Difficulties handling tables and graphs.
  - Issues with accurately reporting data, e.g. may include data from introduction or conclusion sections as results.
- **Practical Application**:
  - LLMs can provide a "first draft" tool for data extraction but for now need human validation.

Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Res Synth Methods*. Mar 3 2024;doi:10.1002/jrsm.1710

Reason T et al. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. *Pharmacoecon Open*. Mar 2024;8(2):205-220. doi:10.1007/s41669-024-00476-9

Meta-analysis and Code Generation

Meta-analysis

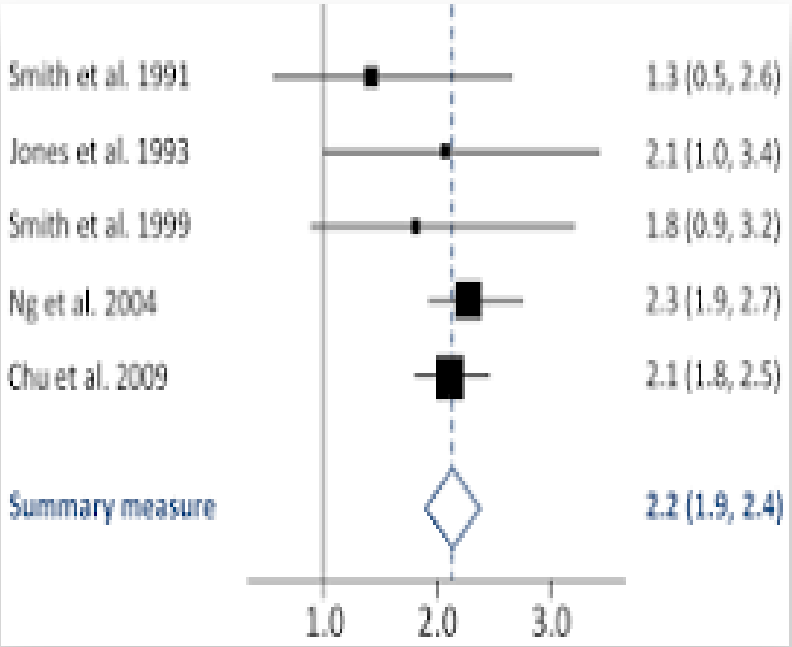
- **Capabilities**:
  - LLMs can **generate code** for conducting meta-analyses (e.g., in R and Python).
  - LLMs can **debug** code and help fix coding errors

- **Findings**:
  - **High Accuracy**: Reported by some studies, such as Reason et al.
  - **Limitations**: Earlier studies have shown a propensity for errors but these may be due to user inexperience and/or LLM capabilities.

Evidence Synthesis: Meta-analysis and Model Parameters

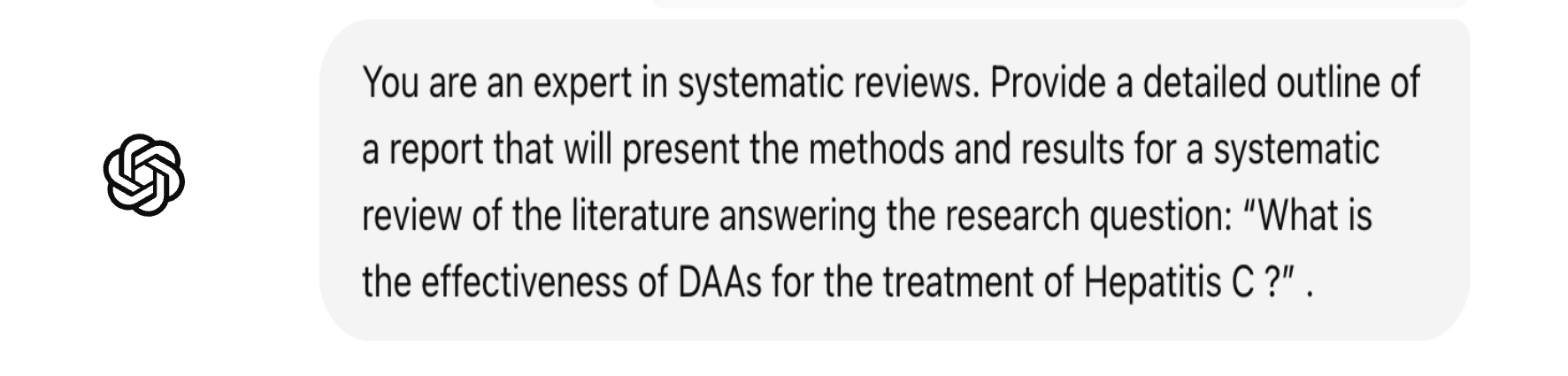
Meta-analysis

- ChatGPT can generate **Python** and **R code** to perform a meta-analysis.
- The code can be implemented in the appropriate interface (e.g. Google Colab).
- ChatGPT is excellent at **debugging code** and problem solving as errors arise.
- However, **expert knowledge** is still essential to determine the appropriate type of analysis (e.g., fixed-effects or random-effects) and to execute the code correctly in Python or R.
- If all has been validated, these results can be used as **inputs** in **decision models** similar to traditional meta-analysis outputs.



Drafting Reports with LLMs

Report writing



- **LLMs capabilities:** excel at summarizing and writing (with the right prompts).
- **Capabilities**: Foundation models can generate initial drafts of systematic literature review (SLR) reports.
- **Potential**: Can produce drafts of reasonable quality, but human review and validation is essential to ensure accuracy and reliability.

Generative AI for Real-World Evidence Generation

- Use of LLMs for extracting insights from **electronic health records** (EHRs) and other unstructured data.
- Benefits: Improved **accuracy** and **efficiency.**
- Limitations: Data **privacy**, potential **inaccuracies** in coding.

NEJM AI

NEJM AI 2024; 1 (5)
DOI: 10.1056/AIdbp2300040

DATASETS, BENCHMARKS, AND PROTOCOLS

Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying

Ali Soroush, M.D., M.S.,[1,2,3] Benjamin S. Glicksberg, Ph.D.,[1,2] Eyal Zimlichman, M.D., M.Sc.,[4,5] Yiftach Barash, M.D., M.Sc.,[5,6] Robert Freeman, R.N., M.S.N., N.E.-B.C.,[7] Alexander W. Charney, M.D., Ph.D.,[1,2] Girish N Nadkarni, M.D., M.P.H.,[1,2] and Eyal Klang, M.D.[1,2,5,8]

Received: July 16, 2023; Revised: February 4, 2024; Accepted: March 1, 2024; Published: April 19, 2024

Some Limitations of Foundation Models and LLMs

- **Accuracy Concerns**: LLMs can produce errors in tasks such as abstract classification and data extraction. There's also the risk of hallucinations (e.g. non-existent citations).

- **Human Oversight is Essential**: While some studies suggest that LLMs can achieve accuracy levels comparable to human efforts, this isn't always consistent. Continuous human oversight and validation are crucial to ensure quality and reliability.

- **Reproducibility Issues**: Different LLMs (and even different prompts) may yield varying results, complicating efforts to replicate studies and findings.

- **Potential for Bias**: Models trained on datasets with inherent biases, can inadvertently skew results.

- **Data Privacy Risks**: Using patient-level data (e.g. in meta-analyses) raises significant privacy and security concerns, necessitating stringent safeguards.

- **Explainability** refers to how well the internal mechanics of a system can be described in human terms. Generative AI models are often seen as "black boxes" due to their complex structures and large data sets, making explainability and interpretability difficult to represent.

NICE Position Statement: Generative AI for SLRs and Evidence Synthesis

- AI can automate **key stages** of systematic reviews and meta-analyses improving **efficiency**, though **validation** is ongoing.
- Ensuring **transparency and explainability** in AI-driven processes is critical to maintain **trust** and **accountability**.
- Methodological rigor must be upheld by applying **established frameworks** (e.g., Cochrane, PALISADE) to **minimize bias** and **validate** AI outputs in evidence synthesis.

NICE. Use of AI in evidence generation: NICE position statement. 2024. Accessed 20 September, 2024.

Towards an HEOR Evaluation Framework for Trustworthy AI ?

LLM Characteristics Description

- **Model Identification and Versioning**
- **Training Data Sources and Scope**
- **Training Methodology and Resources**

LLM Output Evaluation

- Accuracy
- Completeness
- Factuality
- Fairness, Bias, Toxicity
- Deployment Metrics
- Calibration and Uncertainty

Conclusions

Early applications of Generative AI in HEOR show **promise,** but human involvement remains essential

Future outlook: as **user expertise** and **model performance** improve, LLMs are likely to augment SLRs.

Evaluation frameworks for trustworthy AI in HEOR are needed: There are **no shortcuts** to high quality science.