UCL Social Research Institute

Evidence Synthesis Group WG on automation and technology

UCL

Developing guidance and recommendations for the responsible use of AI in systematic reviews

James Thomas

Professor of Social Research & Policy

EPPI Centre
Evidence for Policy & Practice

Outline

- The robots are ~~coming~~ here
- We need to be ready to benefit
- We need to maintain standards
- Guidance development

Picture generated by JT using DALL-E)

YouTube GB

ai literature review

Home

Shorts

Subscriptions

our channel

tory

lists

videos

Later

deos

Tales Offi...

& Free

amx

anDanLv45

DM

nore

g

SciSpace AI Literature Review Workspace - Find and survey relevant papers in minutes

20K views • 8 months ago

SciSpace (formerly Typeset)

Do effortless **literature reviews** using SciSpace. Try it for free now: https://typ.st/45GYk3Z L

Subtitles

Intro | What is a literature review? | Enter SciSpace | The... 11 chapters

How to write literature review using AI. Free AI tool for literature search/review. Scispace.

7.1K views • 1 month ago

XploreBio

Learn how to use this free **AI** tool for **literature** search, summarize papers and drive conclus this tutorial. Query: **literature** ...

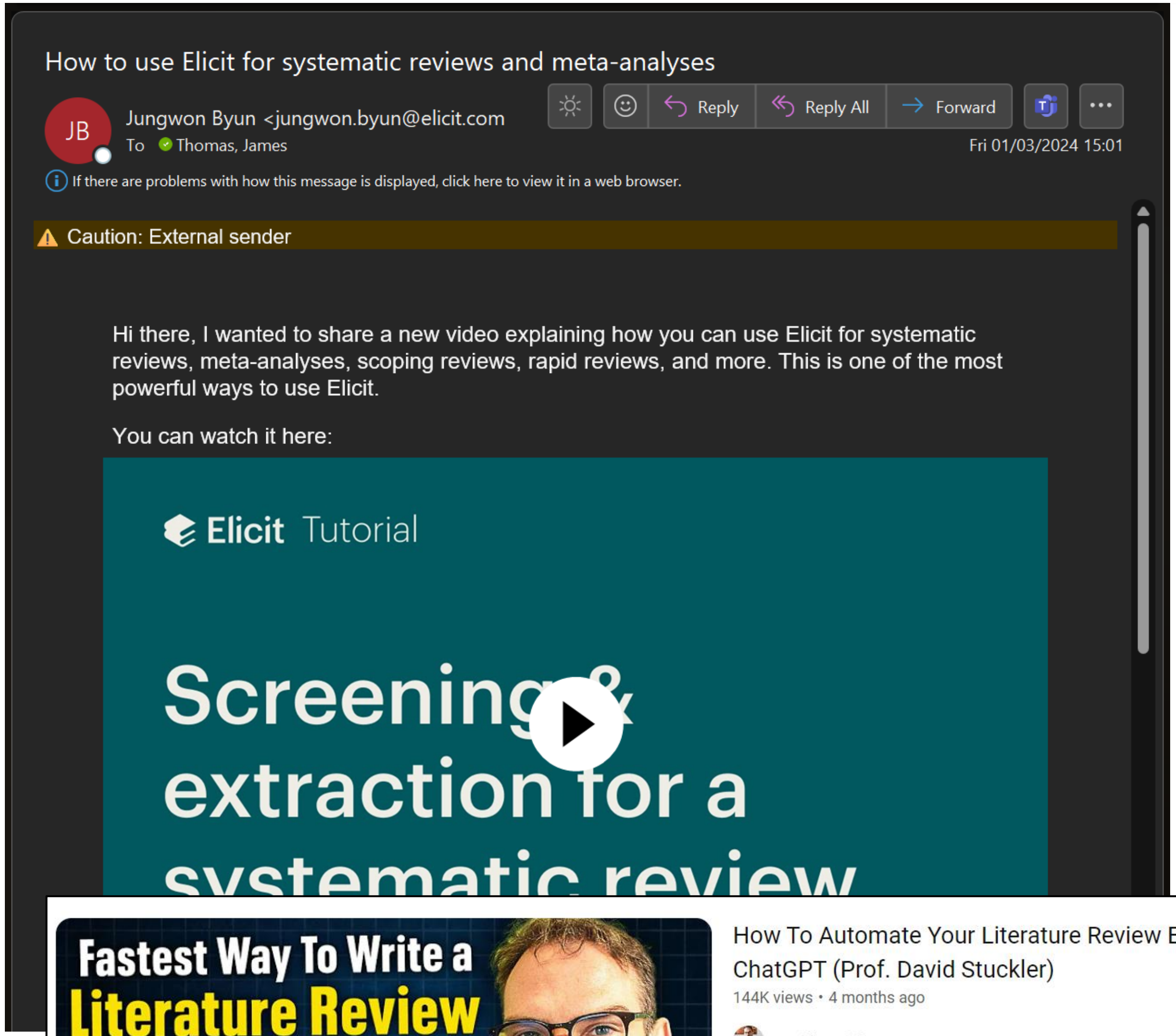
AI Written Literature Review - Generate Paper in Minutes - AI Literature Review Generator

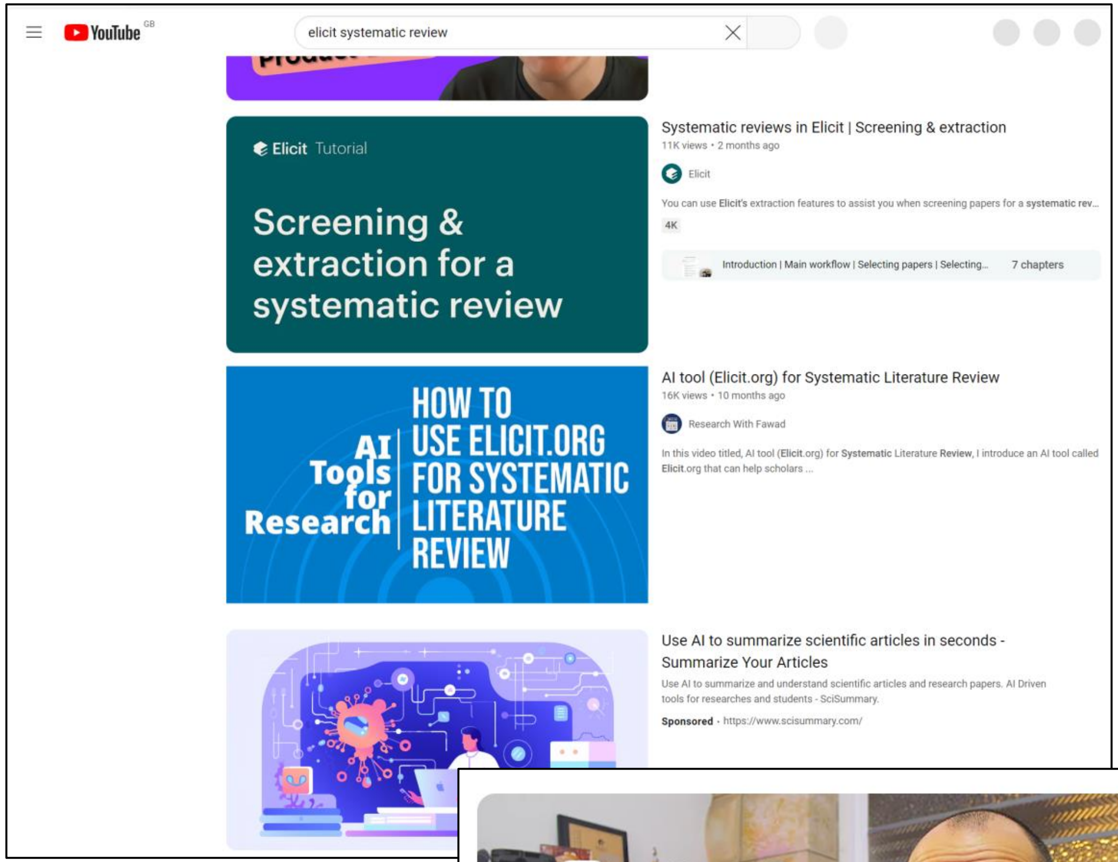
AI Literature Review with Reliable and Trustworthy Citations in APA, MLA, and More.

Sponsored • https://www.aithor.com/ai-thesis

Write A Masterpiece Systematic Literature Review With AI [Next Level Strategies]

21K views • 1 month ago


How to use Elicit for systematic reviews and meta-analyses
JB
Jungwon Byun <jungwon.byun@elicit.com
To Thomas, James
Reply
Reply All
Forward
Fri 01/03/2024 15:01
If there are problems with how this message is displayed, click here to view it in a web browser.
Caution: External sender
Hi there, I wanted to share a new video explaining how you can use Elicit for systematic reviews, meta-analyses, scoping reviews, rapid reviews, and more. This is one of the most powerful ways to use Elicit.
You can watch it here:
Elicit Tutorial
Screening & extraction for a systematic review

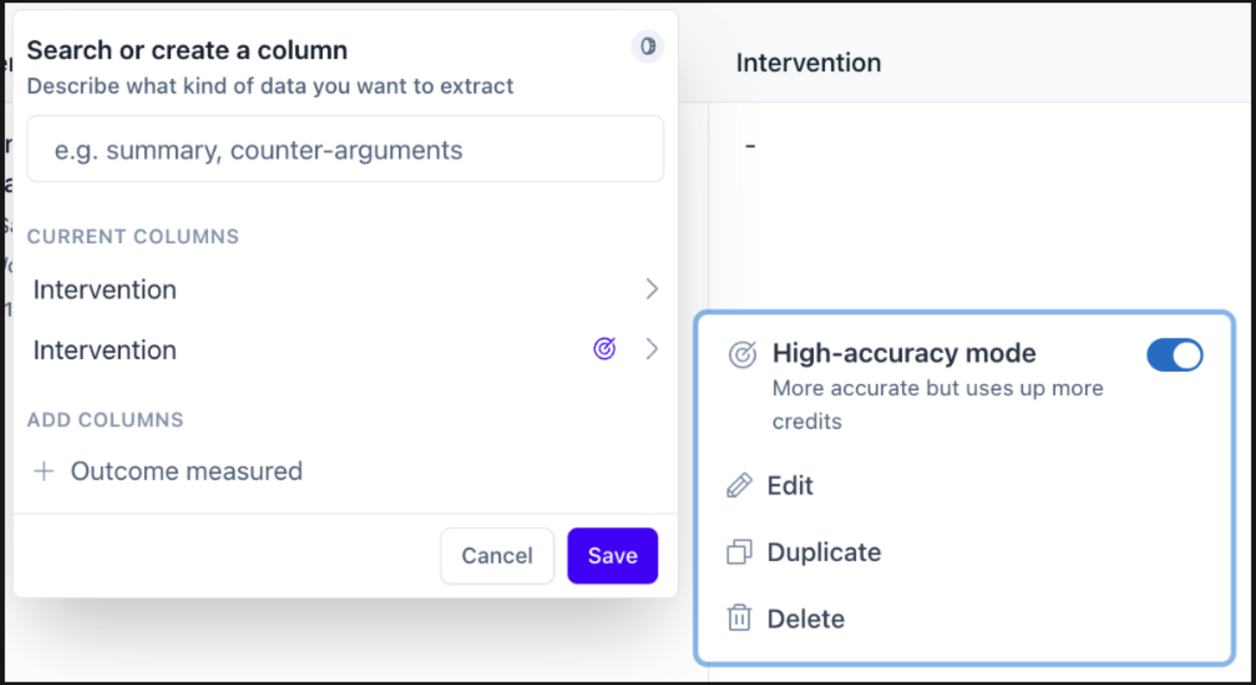

YouTube GB
elicit systematic review
Elicit Tutorial
Screening & extraction for a systematic review
Systematic reviews in Elicit | Screening & extraction
11K views • 2 months ago
Elicit
You can use Elicit's extraction features to assist you when screening papers for a systematic rev...
4K
Introduction | Main workflow | Selecting papers | Selecting...
7 chapters
AI Tools for Research
HOW TO USE ELICIT.ORG FOR SYSTEMATIC LITERATURE REVIEW
AI tool (Elicit.org) for Systematic Literature Review
16K views • 10 months ago
Research With Fawad
In this video titled, AI tool (Elicit.org) for Systematic Literature Review, I introduce an AI tool called Elicit.org that can help scholars ...
Use AI to summarize scientific articles in seconds - Summarize Your Articles
Use AI to summarize and understand scientific articles and research papers. AI Driven tools for researches and students - SciSummary.
Sponsored · https://www.scisummary.com/


Fastest Way To Write a Literature Review
Stage 1
Find Your Research Question
Stage 2
Stage 3
Stage 4
17:20
How To Automate Your Literature Review ETHICALLY Using ChatGPT (Prof. David Stuckler)
144K views • 4 months ago
David Stuckler
... Writing the literature review How To Read Research Papers Effectively: https://youtu.be/WVv2j...
Intro | Finding your research question | Developing an outline...
6 chapters


Fastest
Lit Review
EVER!
9:30

What is high-accuracy mode?

High-accuracy mode gives better results when adding columns and extracting data. In our testing, high-accuracy mode had about 1/2 the error rate of standard columns. High-accuracy mode is particularly useful for conducting systematic reviews and meta-analyses.



High-accuracy mode is only available to Elicit Plus subscribers, and costs about 250 credits per answer.

Learn more about high-accuracy mode here.

Improvements

As of today, we're using a new technique for high-accuracy mode. Our testing found that our new technique reduces the error rate by about 8% compared to our old technique.

- Elicit can be used in ‘high accuracy mode’ for systematic reviews and meta-analyses
- Apparently the error rate is reduced by 8% compared with… something else
- Published evaluations by developers of new tools are poor to non-existent

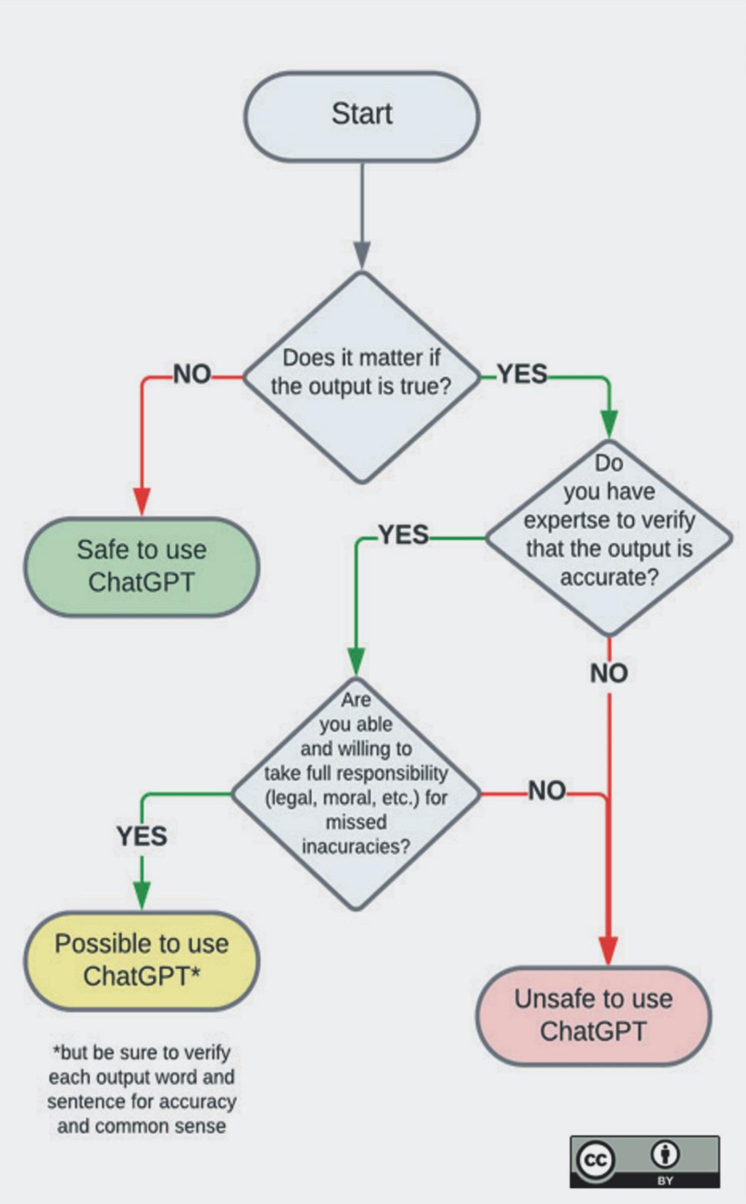
What does this mean for the systematic review field?

- Automation in systematic reviewing is coming fast
- It may be hugely disruptive – possibly akin to the impact of systematic reviews on EBM / evidence informed policy
- There is a danger that established standards for evidence synthesis are compromised / left behind
  - Either because we fail to adapt
  - Or because we allow good evidence synthesis to be displaced by less rigorous (but cheaper) approaches

Picture generated by JT using DALL-E)

When can we use this new technology?

Guidance and standards are emerging



Flowchart devised by Aleksandr Tiulkanov, AI and Data Policy Lawyer, January 2023

Process for developing guidance and recommendations for responsible use of AI in systematic reviews

- ICASR, Cochrane, Campbell, JBI + others involved
- If you want to get involved please register here
  - https://forms.office.com/e/Dg2vwD8agf

- Draft timeline
  - July – draft open for feedback and input
  - September – special session at Global Evidence Summit
  - September – December – further rounds of feedback & revision
  - December – version 1.0 available online
  - Mid-2025 – update (if necessary)

Research integrity

- Considering how accepted principles of research integrity apply can be helpful
  - Rigour
  - Honesty
  - Transparency and open communication
  - Care and respect
  - Accountability

Rigour

- The use of an AI tool in a systematic review must be clearly justified by good evidence
- Rigorous and valid evaluation is key
- Are findings replicable?
- Prevent contamination between training and testing datasets is vital
- We need to build a cumulative evidence base – hence, Studies Within a Review (SWAR)

Development pipeline to justify the use of the Cochrane RCT Classifier

Conventional machine learning model trained on 280,000 records from Cochrane Crowd

Model was calibrated to achieve 99% recall on a second (‘Hedges’) dataset (~50,000 records)

Model was validated on 92,000 studies included in Cochrane intervention reviews

Model was deployed for live use in Cochrane review workflows

Being rigorous in development and testing

Development and evaluation of a classification task using a language model

Prompt development with development dataset

Prompt testing with a *different* dataset

The language model can then apply the prompts to the remaining data

Being rigorous in development and testing

Development and evaluation of a classification task using a language model

Prompt development with development dataset

Prompt testing with a *different* dataset

The language model can then apply the prompts to the remaining data

Critical to avoid contamination between development and testing!

Rigour

- The use of an AI tool in a systematic review must be clearly justified by good evidence
- Rigorous and valid evaluation is key
- Are findings replicable?
  - Deterministic vs non-deterministic / probabilistic algorithms
- Avoiding contamination between training and testing datasets is vital
- We need to build a cumulative evidence base – hence, Studies Within a Review (SWAR)

Honesty

- Honesty about tool performance
- Honesty about making claims in advance of using a tool (e.g. when bidding for work)
- Honesty about evaluation – no sneaky tests of language model prompts outside a proper evaluation framework ☺ (or at least, no contamination of data)

Care and respect

- Language models are known to be biased
- Some development processes remove the most obvious and objectionable output (usually)
  - But biases remain
- We need to be very careful before trusting that it will not generate bias even in a systematic review context

Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study

Travis Zack, Eric Lehman*, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, Atul J Butte, Emily Alsentzer*

Summary

Background Large language models (LLMs) such as GPT-4 hold great promise as transformative tools in health care, ranging from automating administrative tasks to augmenting clinical decision making. However, these models also pose a danger of perpetuating biases and delivering incorrect medical diagnoses, which can have a direct, harmful impact on medical care. We aimed to assess whether GPT-4 encodes racial and gender biases that impact its use in health care.

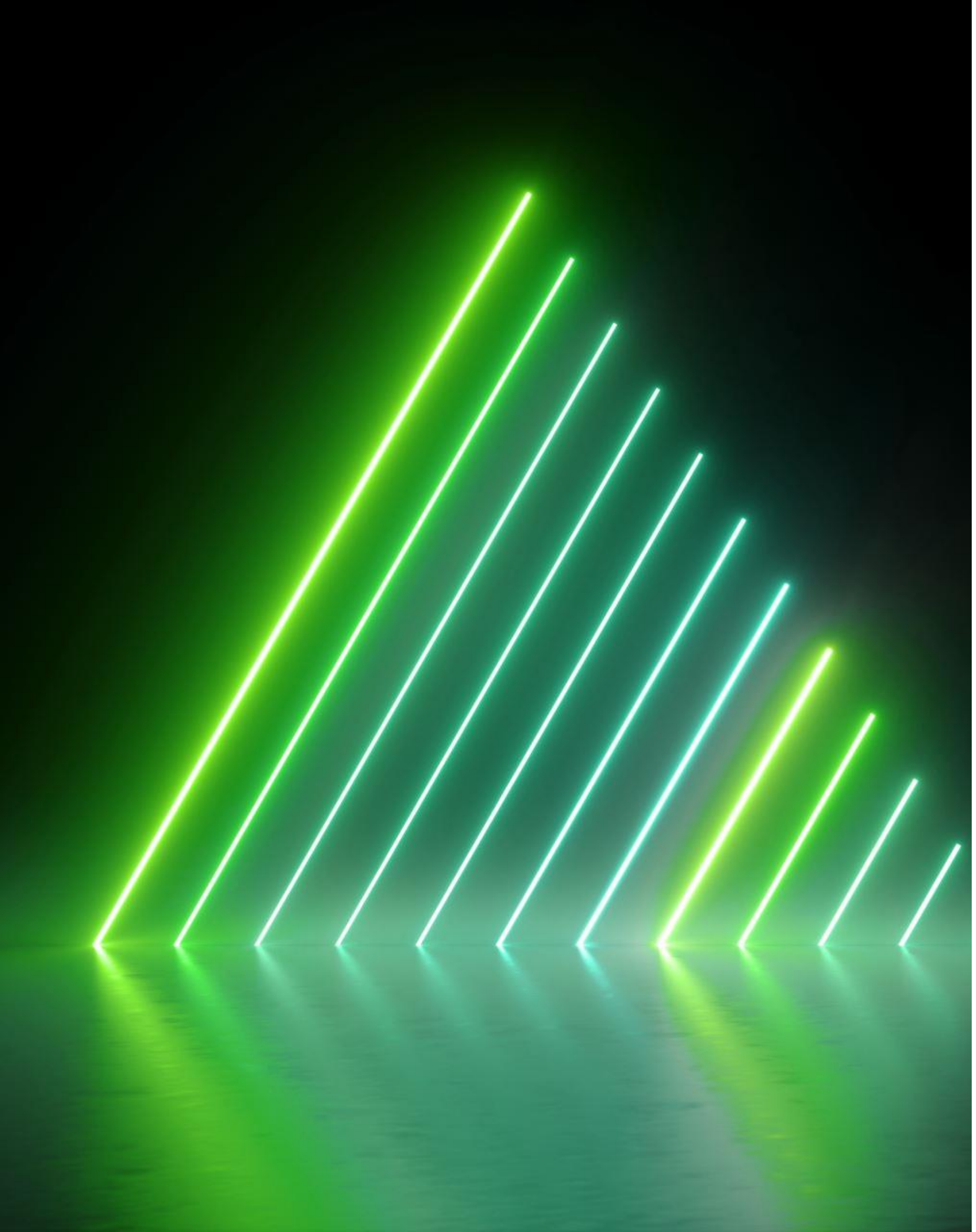
Methods Using the Azure OpenAI application interface, this model evaluation study tested whether GPT-4 encodes racial and gender biases and examined the impact of such biases on four potential applications of LLMs in the clinical domain—namely, medical education, diagnostic reasoning, clinical plan generation, and subjective patient assessment. We conducted experiments with prompts designed to resemble typical use of GPT-4 within clinical and medical education applications. We used clinical vignettes from NEJM Healer and from published research on implicit bias in health care. GPT-4 estimates of the demographic distribution of medical conditions were compared with true US prevalence estimates. Differential diagnosis and treatment planning were evaluated across demographic groups using standard statistical tests for significance between groups.

Findings We found that GPT-4 did not appropriately model the demographic diversity of medical conditions, consistently producing clinical vignettes that stereotype demographic presentations. The differential diagnoses created by GPT-4 for standardised clinical vignettes were more likely to include diagnoses that stereotype certain races, ethnicities, and genders. Assessment and plans created by the model showed significant association between demographic attributes and recommendations for more expensive procedures as well as differences in patient perception.

Interpretation Our findings highlight the urgent need for comprehensive and transparent bias assessments of LLM tools such as GPT-4 for intended use cases before they are integrated into clinical care. We discuss the potential sources of these biases and potential mitigation strategies before clinical implementation.

Transparency and open communication

- How does the tool work?
- How can I replicate / confirm your results?
- Honesty about conflicts of interest
- In evaluation methods

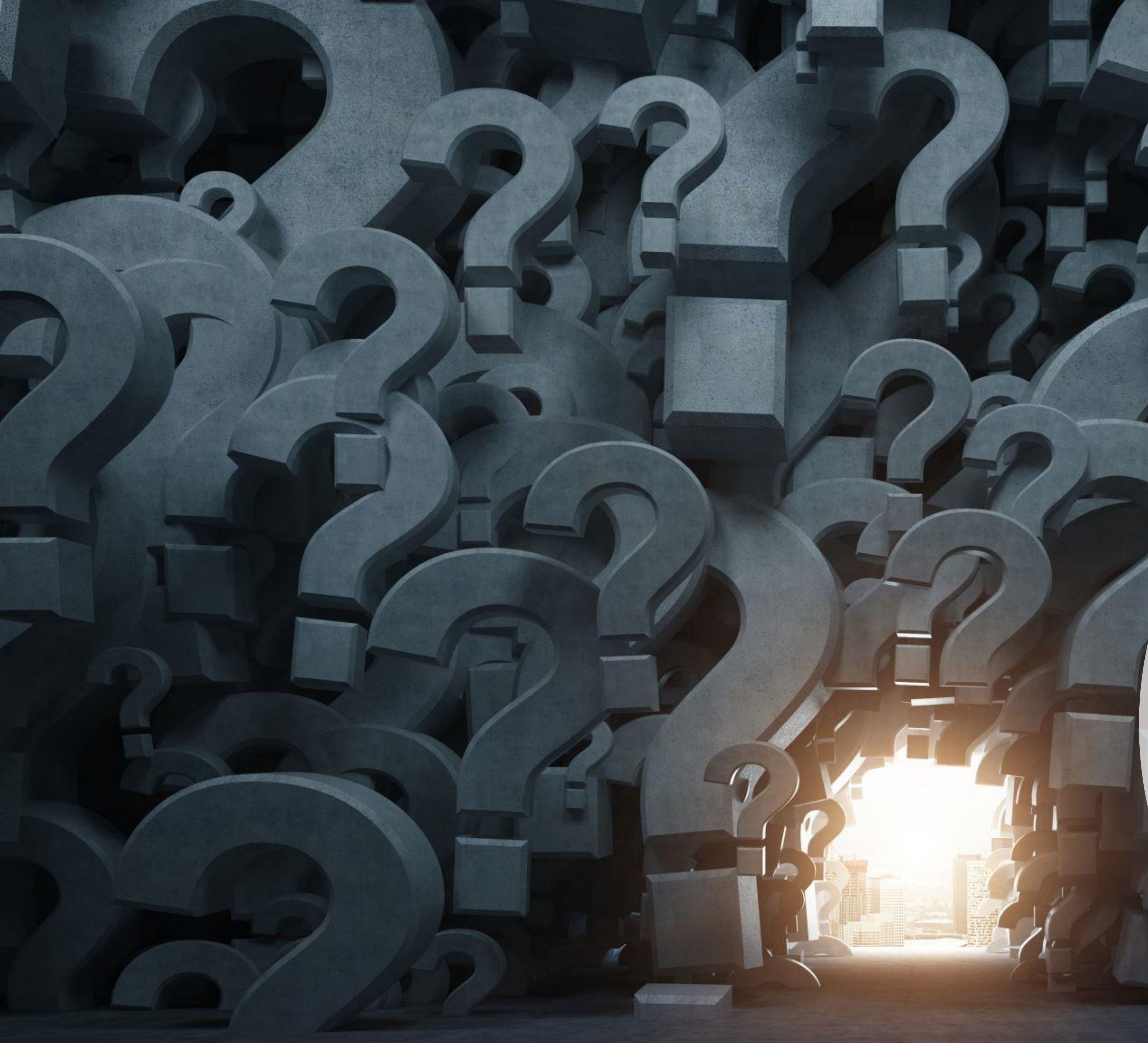


Accountability

- Review authors are responsible for the selection and use of an AI tool (it cannot be accountable for anything)
- We shouldn’t take on trust marketing materials that promote specific tools
- Important reviewers understand (at least up to a point) how a tool works, so they can gauge its risk in their review

Recommendations for

- Systematic reviewers
- Tool developers
- Systematic review organisations



Questions and discussion

What do you think should be in the guidance?

Research integrity

- What would you like guidance on in terms of using AI in systematic reviews?
  - Rigour
  - Honesty
  - Transparency and open communication
  - Care and respect
  - Accountability