Can We Trust AI Output? A Trustworthy AI Perspective for HEOR and RWE

Dr Rachael L. Fleurence

Senior Advisor, National Institutes of Health

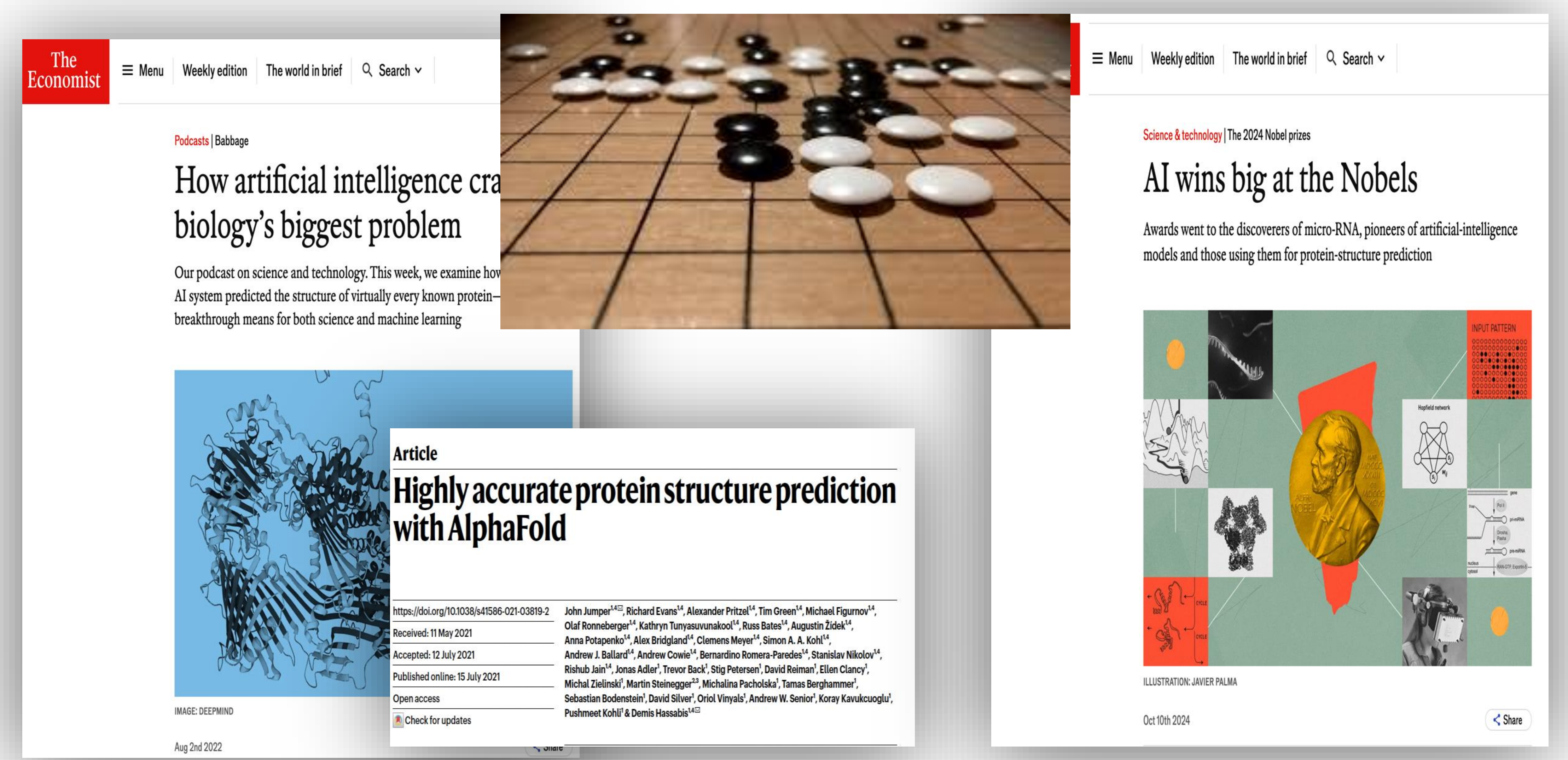
Senior Advisor, National Institute of Biomedical Imaging and Bioengineering

November 18, 2024

Outline

- The promise of generative AI and emerging HEOR applications
- The limitations of generative AI
- NICE Position statement on AI
- Existing frameworks for evaluating trustworthy AI
- Considerations for an evaluation framework in the context of HEOR

The Promise of Generative AI



The Economist

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

Podcasts | Babbage

How artificial intelligence cra biology's biggest problem

Our podcast on science and technology. This week, we examine hov AI system predicted the structure of virtually every known protein— 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]✉

☰ 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

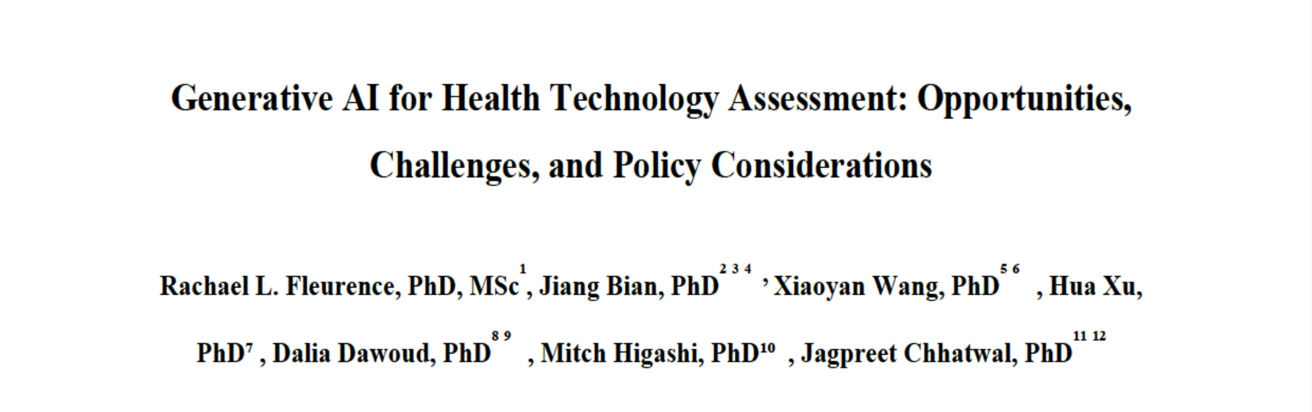
ILLUSTRATION: JAVIER PALMA

Oct 10th 2024

Share

Emerging Applications in HEOR

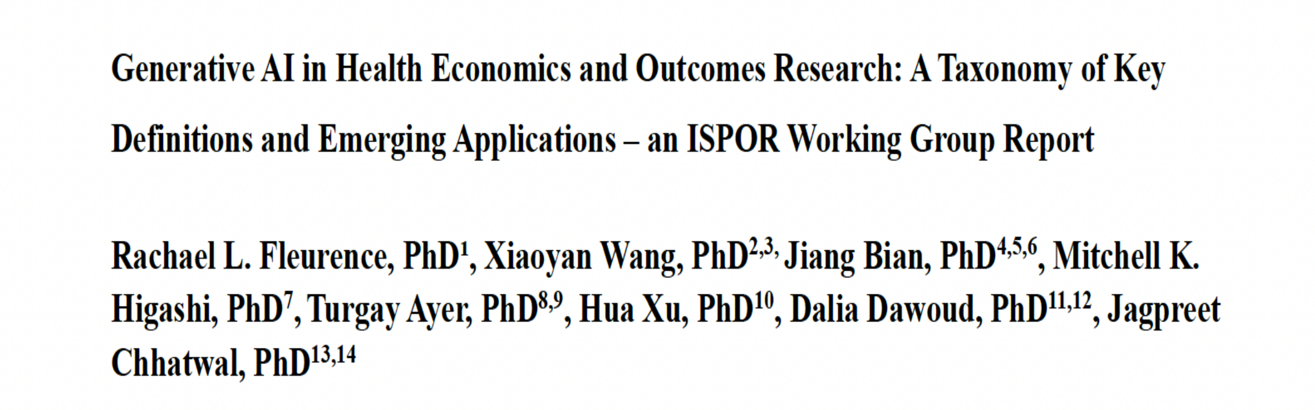
- Systematic Literature Reviews
- Health Economic Modeling
- Real World Evidence Generation
- Dossier Development



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



Generative AI in Health Economics and Outcomes Research: A Taxonomy of Key Definitions and Emerging Applications – an ISPOR Working Group Report

Rachael L. Fleurence, PhD[1], Xiaoyan Wang, PhD[2,3,] Jiang Bian, PhD[4,5,6], Mitchell K. Higashi, PhD[7], Turgay Ayer, PhD[8,9], Hua Xu, PhD[10], Dalia Dawoud, PhD[11,12], Jagpreet Chhatwal, PhD[13,14]

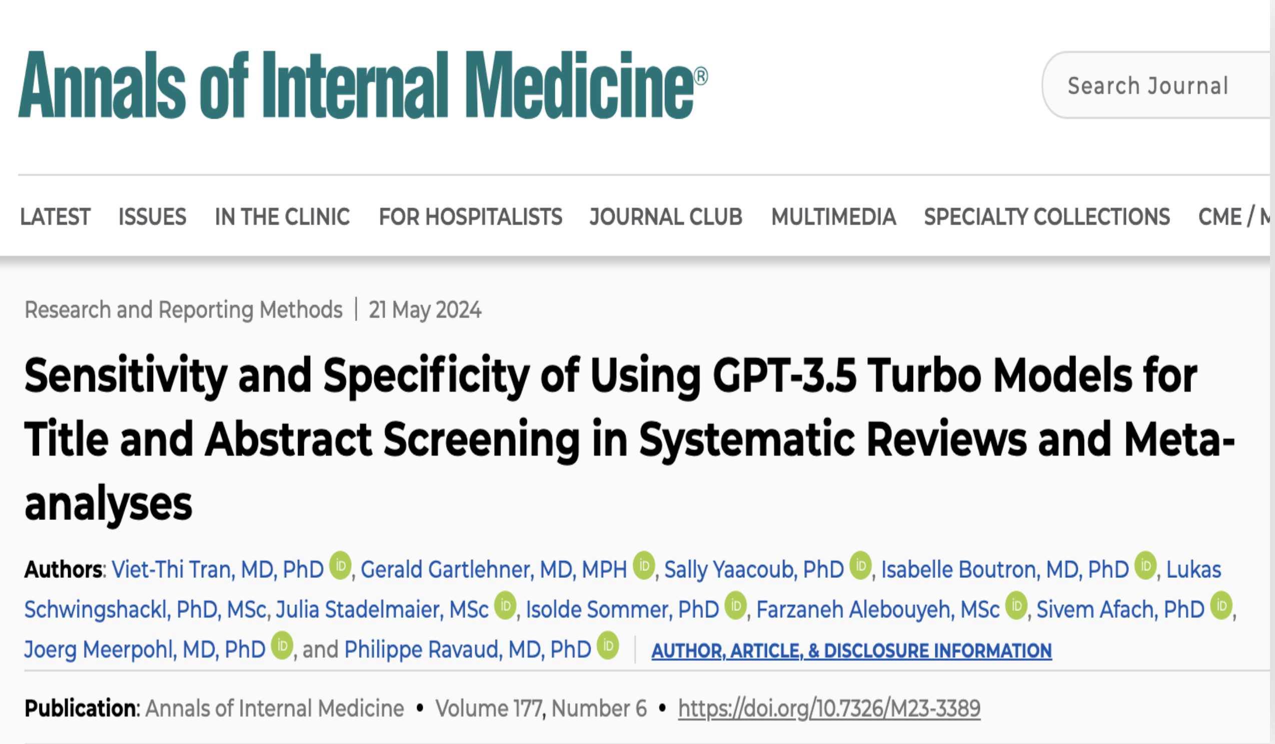
Fleurence et al. https://arxiv.org/abs/2410.20204

EXAMPLE !

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



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

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

Some Limitations of Foundation Models and LLMs

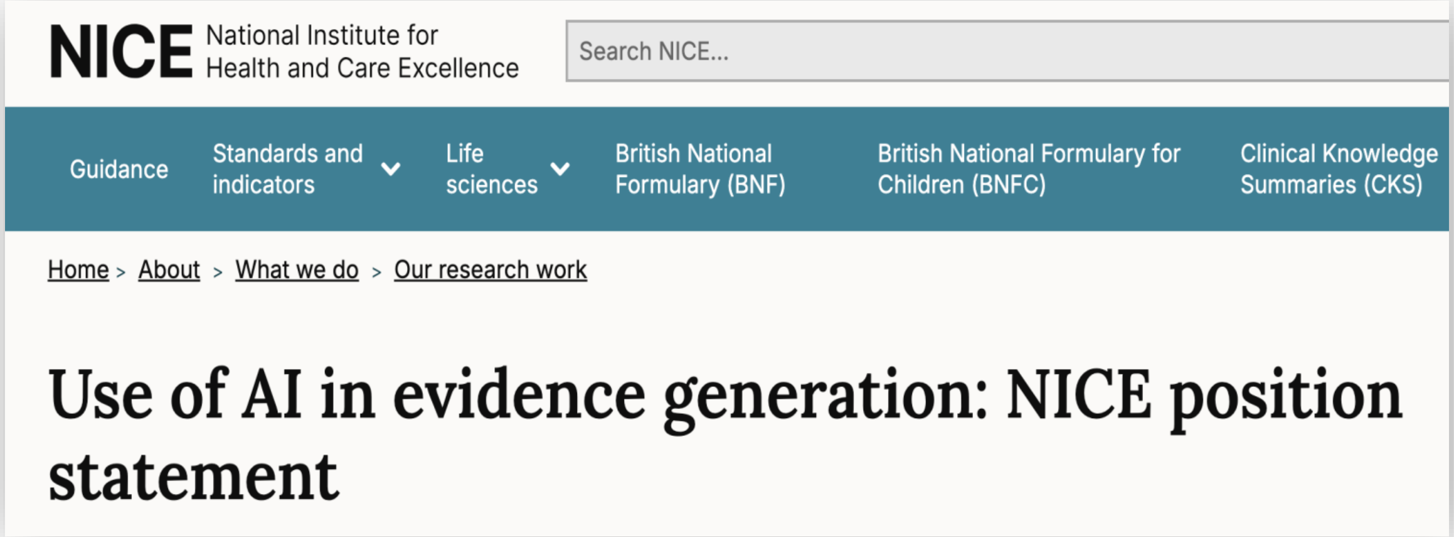
- **Accuracy Concerns**: LLMs can produce errors in tasks such as abstract classification and data extraction. There's also the risk of hallucinations (e.g. non-existent citations).
- **Human Oversight is Essential**: While some studies suggest that LLMs can achieve accuracy levels comparable to human 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.

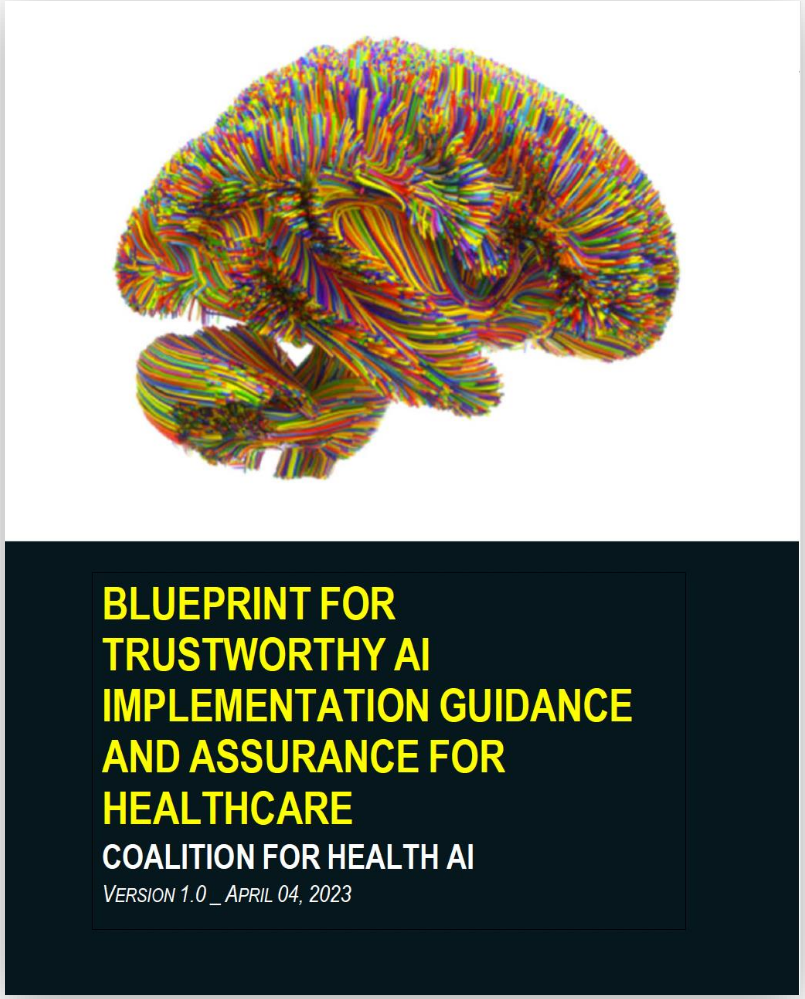

NICE National Institute for Health and Care Excellence
Search NICE...
Guidance
Standards and indicators
Life sciences
British National Formulary (BNF)
British National Formulary for Children (BNFC)
Clinical Knowledge Summaries (CKS)
Home > About > What we do > Our research work

Use of AI in evidence generation: NICE position statement

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

Frameworks for Trustworthy AI: Coalition for Health AI (CHAI)

- Transparency & Accountability:
- Bias Management
- Safety & Reliability
- Security & Privacy
- Continuous Monitoring



BLUEPRINT FOR TRUSTWORTHY AI IMPLEMENTATION GUIDANCE AND ASSURANCE FOR HEALTHCARE

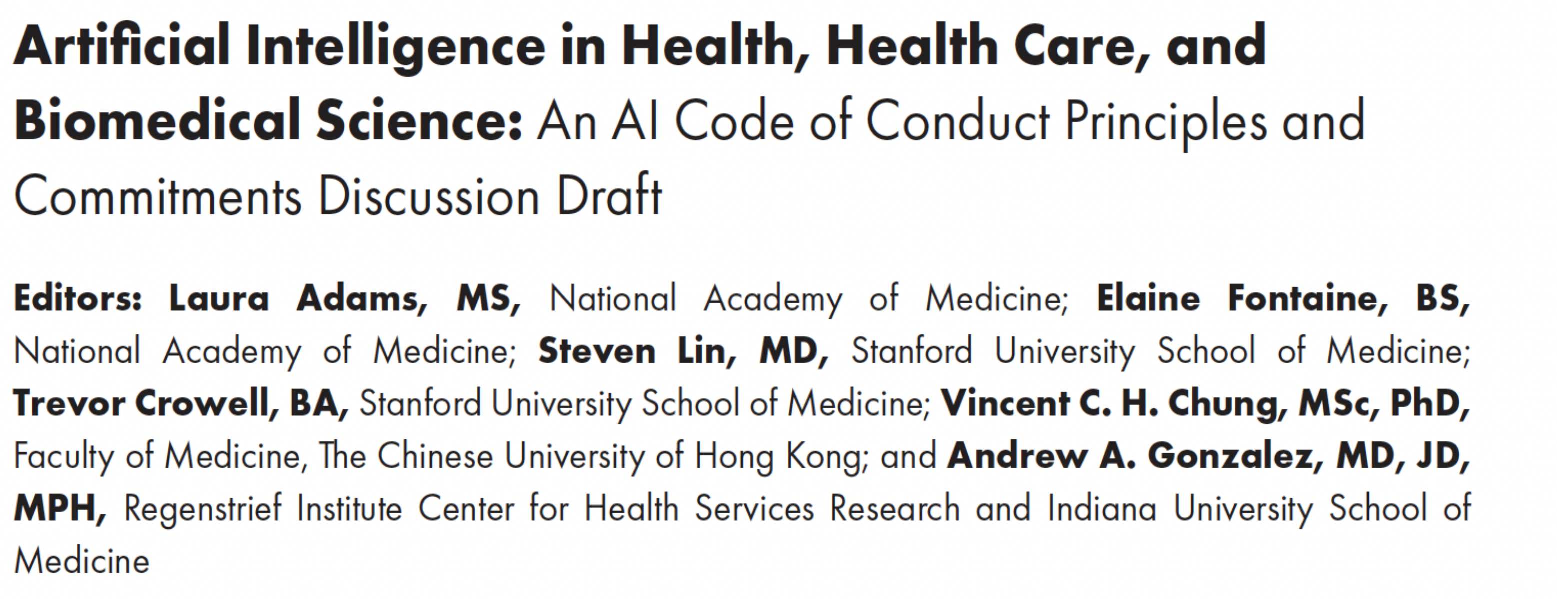
COALITION FOR HEALTH AI

VERSION 1.0 _ APRIL 04, 2023

Reference: Blueprint for trustworthy AI implementation guidance and assurance for healthcare (CHAI, 2023)

Frameworks for Trustworthy AI: National Academies of Medicine

- Engagement and Inclusiveness
- Safety and Accountability
- Equity and Fairness
- Transparency and Explainability
- Sustainability and Efficiency

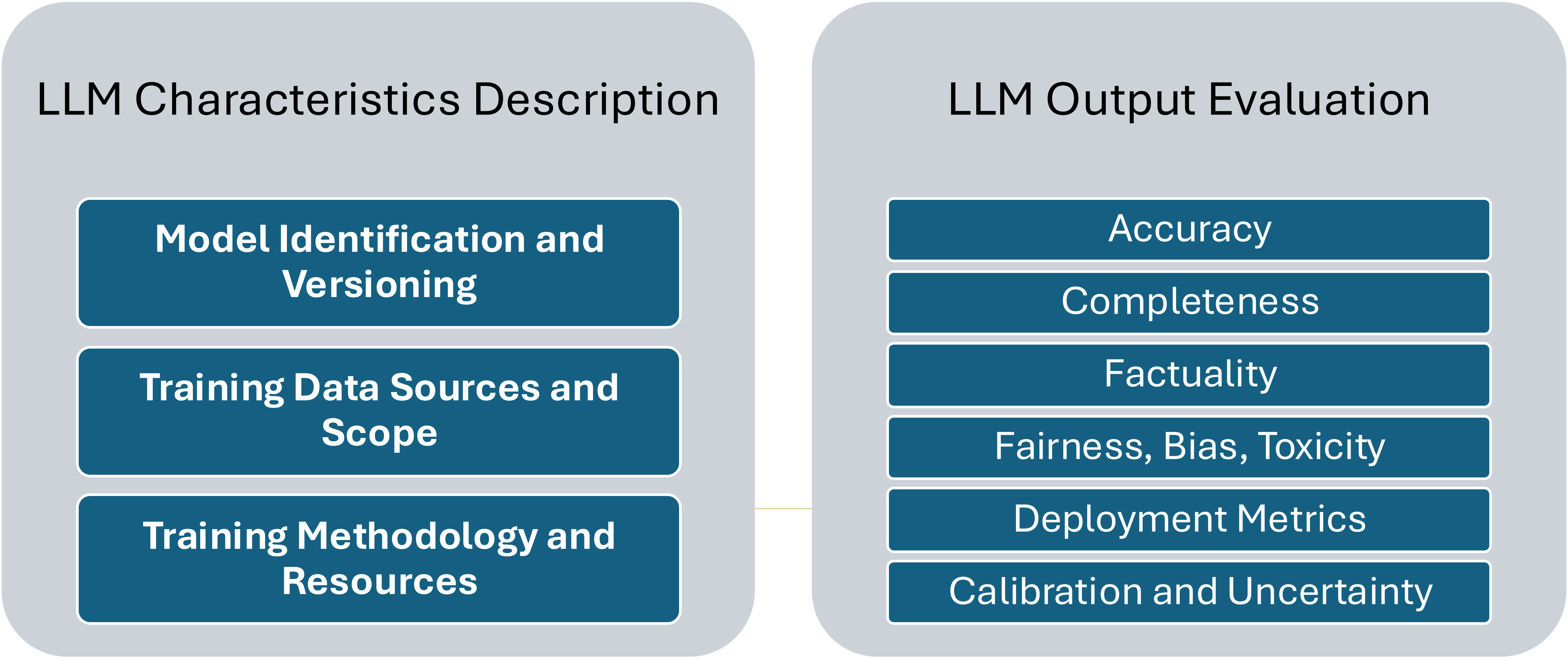


Artificial Intelligence in Health, Health Care, and Biomedical Science: An AI Code of Conduct Principles and Commitments Discussion Draft

Editors: Laura Adams, MS, National Academy of Medicine; **Elaine Fontaine, BS,** National Academy of Medicine; **Steven Lin, MD,** Stanford University School of Medicine; **Trevor Crowell, BA,** Stanford University School of Medicine; **Vincent C. H. Chung, MSc, PhD,** Faculty of Medicine, The Chinese University of Hong Kong; and **Andrew A. Gonzalez, MD, JD, MPH,** Regenstrief Institute Center for Health Services Research and Indiana University School of Medicine

AI in health, healthcare and biomedical science: an AI code of conduct (Adams et al. 2024)

Possible Domains for 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

ISPOR Working Group on Generative AI - Work in Progress, November 2024

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.