Evidence creation, gathering, & synthesis with the tools of AI: potential & considerations

Chirag J Patel

NASEM: Exploring the types of evidence synthesis and communications in diet and chronic disease relationships

July 10, 2025

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org

1

Disclosures

- Research funding (had no role in the production of research or this presentation)
  - NIH (NIDDK, NIEHS, NIAID, NIA)
  - ARPA-H
  - NSF
  - Vranos Foundation
  - Steven and Alexandra Cohen Foundation
- Compute credits: Google, Microsoft, Amazon, Oracle
- Consulting: Sanofi, Janssens

2

Emerging types of AI and high level use cases in biomedical research and applications

Supervised
Practice of medicine:
Faster, cheaper, more accurate diagnoses/ decisions?

Unsupervised
Precision medicine:
Discovering new actionable disease states

Generative
Creating medicine:
Proposing new drugs
Chatting with a virtual doctor/expert

Reinforcement
"Problem solving":
Thinking through problems to "reason"

3

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

JAMA 2016

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

IMPORTANCE Deep learning is a family of computational methods that allow an algorithm to program itself by learning from a large set of examples that demonstrate the desired behavior, removing the need to specify rules explicitly. Application of these methods to medical imaging requires further assessment and validation.

OBJECTIVE To apply deep learning to create an algorithm for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

DESIGN AND SETTING A specific type of neural network optimized for image classification called a deep convolutional neural network was trained using a retrospective development data set of 128 175 retinal images, which were graded 3 to 7 times for diabetic retinopathy, diabetic macular edema, and image gradability by a panel of 54 US licensed ophthalmologists and ophthalmology senior residents between May and December 2015. The resultant algorithm was validated in January and February 2016 using 2 separate data sets, both graded by at least 7 US board-certified ophthalmologists with high intragrader consistency.

EXPOSURE Deep learning–trained algorithm.

MAIN OUTCOMES AND MEASURES The sensitivity and specificity of the algorithm for detecting referable diabetic retinopathy (RDR), defined as moderate and worse diabetic retinopathy, referable diabetic macular edema, or both, were generated based on the reference standard of the majority decision of the ophthalmologist panel. The algorithm was evaluated at 2 operating points selected from the development set, one selected for high specificity and another for high sensitivity.

RESULTS The EyePACS-1 data set consisted of 9963 images from 4997 patients (mean age, 54.4 years; 62.2% women; prevalence of RDR, 683/8878 fully gradable images [7.8%]); the Messidor-2 data set had 1748 images from 874 patients (mean age, 57.6 years; 42.6% women; prevalence of RDR, 254/1745 fully gradable images [14.6%]). For detecting RDR, the algorithm had an area under the receiver operating curve of 0.991 (95% CI, 0.988-0.993) for EyePACS-1 and 0.990 (95% CI, 0.986-0.995) for Messidor-2. Using the first operating cut point with high specificity, for EyePACS-1, the sensitivity was 90.3% (95% CI, 87.5%-92.7%) and the specificity was 98.1% (95% CI, 97.8%-98.5%). For Messidor-2, the sensitivity was 87.0% (95% CI, 81.1%-91.0%) and the specificity was 98.5% (95% CI, 97.7%-99.1%). Using a second operating point with high sensitivity in the development set, for EyePACS-1 the sensitivity was 97.5% and specificity was 93.4% and for Messidor-2 the sensitivity was 96.1% and specificity was 93.9%.

Key Points

Question How does the performance of an automated deep learning algorithm compare with manual grading by ophthalmologists for identifying diabetic retinopathy in retinal fundus photographs?

Finding In 2 validation sets of 9963 images and 1748 images, at the operating point selected for high specificity, the algorithm had 90.3% and 87.0% sensitivity and 98.1% and 98.5% specificity for detecting referable diabetic retinopathy, defined as moderate or worse diabetic retinopathy or referable macular edema by the majority decision of a panel of at least 7 US board-certified ophthalmologists. At the operating point selected for high sensitivity, the algorithm had 97.5% and 96.1% sensitivity and 93.4% and 93.9% specificity in the 2 validation sets.

Meaning Deep learning algorithms had high sensitivity and specificity for detecting diabetic retinopathy and macular edema in retinal fundus photographs.

Author Affiliations: Google Inc, Mountain View, California (Gulshan, Peng, Coram, Stumpe, Wu, Narayanaswamy, Venugopalan, Widner, Madams, Nelson, Webster); Department of Computer Science, University of Texas, Austin (Venugopalan); EyePACS LLC, San Jose, California (Cuadros); School of Optometry, Vision Science Graduate Group, University of California, Berkeley (Cuadros); Aravind Medical Research Foundation, Aravind Eye Care System, Madurai, India (Kim); Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya

Deep vision model (CNN) with high AUC
Clear clinical bottleneck articulated
Cross continental validation
Comparison with human raters

4

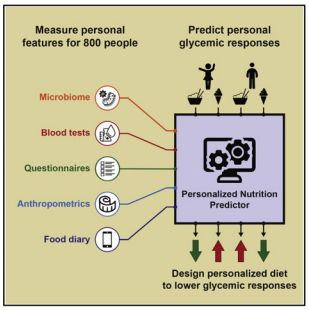
Examples in the precision nutrition space:
But need prospective data to learn of benefits

Article

Cell

Personalized Nutrition by Prediction of Glycemic Responses

Graphical Abstract



Authors
David Zeevi, Tal Korem, Niv Zmora, ..., Zamir Halpern, Eran Elinav, Eran Segal

Correspondence
eran.elinav@weizmann.ac.il (E.E.), eran.segal@weizmann.ac.il (E.S.)

In Brief
People eating identical meals present high variability in post-meal blood glucose response. Personalized diets created with the help of an accurate predictor of blood glucose response that integrates parameters such as dietary habits, physical activity, and gut microbiota may successfully lower post-meal blood glucose and its long-term metabolic consequences.

Cell 2015

Highlights

- High interpersonal variability in post-meal glucose observed in an 800-person cohort
- Using personal and microbiome features enables accurate glucose response prediction
- Prediction is accurate and superior to common practice in an independent cohort
- Short-term personalized dietary interventions successfully lower post-meal glucose

5

Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables

Emma Ahlqvist, Petter Storm, Annemari Käräjämäki, Mats Martinell*, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, Rashmi B Prasad, Dina Mansour Aly, Peter Almgren, Ylva Wessman, Nael Shaat, Peter Spégel, Hindrik Mulder, Eero Lindholm, Olle Melander, Ola Hansson, Ulf Malmqvist, Åke Lernmark, Kaj Lahti, Tom Forsén, Tiinamaija Tuomi, Anders H Rosengren, Leif Groop*

Lancet Diabetes and Endocrinology 2019

Are there more than 3 major types of diabetes?

… and do they matter clinically?

Unsupervised learning on diabetics: f(HbA1C, BMI, HOMA-IR, HOMA-B, age of diagnosis, glutamate carboxylate antibodies) highlight new phenotypes with potentially new risk factors

6

The Large Language Model ("ChatGPT") revolution: Vast feature set and getting better

- Summarization of text
- Creation of text, images, and videos
- Programming and data processing
- "Thinking": solving problems and searching for abstract ideas
- *Hallucination and references to false literature*
- *Proliferation of false positives?*

7

FDA NEWS RELEASE

FDA Launches Agency-Wide AI Tool to Optimize Performance for the American People

For Immediate Release: June 02, 2025

The U.S. Food and Drug Administration (FDA) today launched Elsa, a generative Artificial Intelligence (AI) tool designed to help employees—from scientific reviewers to investigators—work more efficiently. This innovative tool modernizes agency functions and leverages AI capabilities to better serve the American people.

"Following a very successful pilot program with FDA's scientific reviewers, I set an aggressive timeline to scale AI agency-wide by June 30," **said FDA Commissioner Marty Makary, M.D., M.P.H.** "Today's rollout of Elsa is ahead of schedule and under budget, thanks to the collaboration of our in-house experts across the centers."

Built within a high-security GovCloud environment, Elsa offers a secure platform for FDA employees to access internal documents while ensuring all information remains within the agency. The models do not train on data submitted by regulated industry, safeguarding the sensitive research and data handled by FDA staff.

"Today marks the dawn of the AI era at the FDA with the release of Elsa, AI is no longer a distant promise but a dynamic force enhancing and optimizing the performance and potential of every employee," **said FDA Chief AI Officer Jeremy Walsh.** "As we learn how employees are using the tool, our development team will be able to add capabilities and grow with the needs of employees and the agency."

The agency is already using Elsa to accelerate clinical protocol reviews, shorten the time needed for scientific evaluations, and identify high-priority inspection targets.

8

Can *Large Language Models* enhance and/or reduce the bottlenecks of the evidence synthesis process?
Specific Questions:

- What are the **"causal machinery"** behind LLMs? [**unknown**]
  - What data were they trained on?
  - How can these be refined (e.g., nutritional or interventional studies and papers)?
- How good are they at information retrieval?: millions of papers to traverse
- Can they distinguish study design artifacts from genuine biases? [**to be evaluated**]
  - Can they reduce publication biases by uncovering file-drawer studies?
  - Do they know what is "high quality" vs. "low quality"? Randomized vs observational evidence?
  - Can they infer across biological models (cell line vs. mouse vs human?)
- Can they predict reproducibility? [**to be evaluated**]
  - .. or understand concepts such as *triangulation* or *Bradford-Hill* criteria?
  - … or inflate false positive findings? [**yes**]
- Can they extract summary statistics? [**to be evaluated; getting better**]

9

NEJM AI

DOI: 10.1056/AIoa2400555

ORIGINAL ARTICLE

Autonomous LLM-Driven Research—from Data to Human-Verifiable Research Papers

Tal Ifargan, BSc.,[1] Lukas Hafner, Ph.D.,[2] Maor Kern,[1] Ori Alcalay, BSc.,[1] and Roy Kishony, Ph.D.[2,4,5]

Received: June 4, 2024; Revised: October 10, 2024; Accepted: October 17, 2024; Published: December 3, 2024

BACKGROUND Artificial intelligence (AI) promises to accelerate scientific discovery, but it remains unclear whether AI systems can perform fully autonomous research, and whether they can do so while adhering to key scientific values, such as transparency, traceability, and verifiability. The aim of this study was to develop and evaluate an AI-automation platform that performs transparent, traceable, and human-verifiable scientific research.

METHODS To mimic human scientific practices, we built "data-to-paper," an automation platform that guides interacting large language model (LLM) agents through a complete stepwise research process that starts with annotated data and results in comprehensive research papers, while programmatically backtracing information flow and allowing human oversight and interactions. The platform can run fully autonomously (in autopilot mode) or with human intervention (in copilot mode).

RESULTS In autopilot mode, provided only with annotated data, data-to-paper raised hypotheses; designed research plans; wrote and debugged analysis codes; generated and interpreted results; and created complete, information-traceable research papers. Even though the research novelty of manuscripts created by data-to-paper was relatively limited, the process demonstrated the autonomous generation of de novo quantitative insights from data, such as unraveling associations between health indicators and clinical outcomes. For simple research goals and datasets, a fully autonomous cycle can create manuscripts that independently recapitulate the findings of peer-reviewed biomedical publications without major errors in about 80 to 90% of cases. Yet, as goal or data complexity increases, human copiloting becomes critical for ensuring accuracy and overall quality. By tracking information flow through the steps, the platform creates "data-chained" manuscripts, in which downstream results are programmatically linked to upstream code and data, thus setting a new standard for the verifiability of scientific outputs.

CONCLUSIONS Our work demonstrates the potential for AI-driven acceleration of scientific discovery in data-driven biomedical research and beyond, while enhancing, rather than jeopardizing, traceability, transparency, and verifiability.

Open Goal/"autopilot":
use of public data (e.g., CDC BRFSS)
1 hour ~ hypothesis generation to code, run analyses, and generate the paper; run 5 times

Fixed Goal: replication of existing paper with negative or positive findings; or findings with multiple steps (published after the GPT was trained)
Provided the goal of the paper; ran 10 times

Input annotated data (e.g., data dictionary) and produces a traceable paper

1. Supply the goal (closed) or have the system create a hypothesis (autopilot or with some human intervention)
2. Literature search
3. Plan for testing hypothesis
4. Table design
5. Literature search
6. Human review

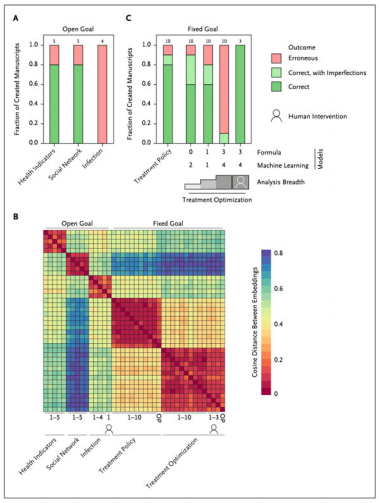


Figure 3. Data-to-Paper Is Able to Autonomously Create Correct Papers for Simple Research Goals and Datasets While Human Copiloting Is Required to Ensure Accuracy in More Complex Settings.

10

nature

Explore content · About the journal · Publish with us · Subscribe

PLOS Biology

nature > news > article

NEWS | 21 May 2025

AI linked to explosion of low-quality biomedical research papers

Analysis flags hundreds of studies that seem to follow a template, reporting correlations between complex health conditions and single variables based on publicly available data sets.

By Miryam Naddaf

Health data from thousands of people is publicly available and ready to plug into AI systems for analysis. Credit: BSIP/UIG Via Getty

The scientific literature is at risk of becoming flooded with papers that make misleading health claims based on openly available data that are easy to process using artificial intelligence (AI) tools, researchers have warned.

In a study published in *PLoS Biology* on 8 May[1], scientists analysed more than 300 papers that used data from the US National Health and Nutrition Examination Survey (NHANES), an open data set of health records. The papers all seemed to

Check for updates

OPEN ACCESS

Citation: Suchak T, Aliu AE, Harrison C, Zwiggelaar R, Geifman N, Spick M (2025) Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. PLoS Biol 23(5): e3003152. https://doi.org/10.1371/journal.pbio.3003152

Academic Editor: Marcus Munafò, University of Bristol, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: January 15, 2025

Accepted: April 4, 2025

Published: May 8, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3003152

META-RESEARCH ARTICLE

Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database

Tulsi Suchak[1], Anietie E. Aliu[1], Charlie Harrison[2], Reyer Zwiggelaar[2], Nophar Geifman[1], Matt Spick[1]*

1 School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, United Kingdom, **2** Department of Computer Science, Aberystwyth University, Ceredigion, United Kingdom

* matt.spick@surrey.ac.uk

Abstract

With the growth of artificial intelligence (AI)-ready datasets such as the National Health and Nutrition Examination Survey (NHANES), new opportunities for data-driven research are being created, but also generating risks of data exploitation by paper mills. In this work, we focus on two areas of potential concern for AI-supported research efforts. First, we describe the production of large numbers of formulaic single-factor analyses, relating single predictors to specific health conditions, where multifactorial approaches would be more appropriate. Employing AI-supported single-factor approaches removes context from research, fails to capture interactions, avoids false discovery correction, and is an approach that can easily be adopted by paper mills. Second, we identify risks of selective data usage, such as analyzing limited date ranges or cohort subsets without clear justification, suggestive of data dredging, and post-hoc hypothesis formation. Using a systematic literature search for single-factor analyses, we identified 341 NHANES-derived research papers published over the past decade, each proposing an association between a predictor and a health condition from the wide range contained within NHANES. We found evidence that research failed to take account of multifactorial relationships, that manuscripts did not account for the risks of false discoveries, and that researchers selectively extracted data from NHANES rather than utilizing the full range of data available. Given the explosion of AI-assisted productivity in published manuscripts (the systematic search strategy used here identified an average of 4 papers per annum from 2014 to 2021, but 190 in 2024–9 October alone), we highlight a set of best practices to address these concerns, aimed at researchers, data controllers, publishers, and peer reviewers, to encourage improved statistical practices and mitigate the risks of paper mills using AI-assisted workflows to introduce low-quality manuscripts to the scientific literature.

11

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

Follow this preprint

The architecture of exposome-phenome associations

Chirag J Patel, John PA Ioannidis, Arjun K Manrai

doi: https://doi.org/10.1101/2025.06.05.25329055

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

Non-genetic exposures—including nutrients, lifestyle factors, consumables, and pollutants—substantially contribute to phenotypic variation. Most studies assess only a few exposures or phenotypes, yielding fragmented exposome-phenome relationships. Systematic approaches are needed to quantify how the exposome—the totality of environmental exposures—relates broadly to clinically relevant phenotypes. We developed a resource benchmarking the exposome's role using data from the National Health and Nutrition Examination Survey (NHANES), cataloging 619 exposures and 278 phenotypes, and systematically testing associations (Phenotype-exposure-wide association study [P-ExWAS]). Among ~119k associations, 5% (n=5,661) were Bonferroni significant, and 40% replicated across independent population samples. Single exposures explained modest variance (median R^2=0.5%; interquartile range [IQR]: 0.27–1.10%). Twenty simultaneous exposome factors increased median variance explained to 3.5% (IQR: 1.8–7.8%), comparable to 1M genetic variants. The exposome-phenome atlas is available at: http://apps.chiragjpgroup.org/pe_atlas

Competing Interest Statement

The authors have declared no competing interest.

Funding Statement

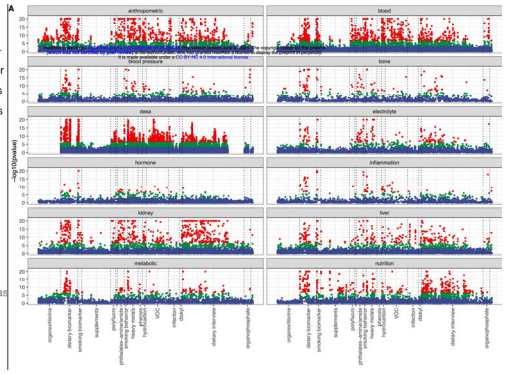
This study was funded by NIEHS R01ES032470 (CJP, JPAI, AKM), NIDDK R01DK137993 (CJP, AKM), NIEHS U24ES036819 (CJP)

Can an AI be taught to go "domain-wide"?

119k total associations:
~6k significant (by FDR or family wise correction, 5%)

~40k nominally significant ($p < 0.05$): **36%!**

What are the "true positives"?!



12

What makes an AI application successful?: Clear benchmarks to gauge advances

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[1,2☒], Richard Evans[1,2], Alexander Pritzel[1,2], Tim Green[1,2], Michael Figurnov[1,2], Olaf Ronneberger[1,2], Kathryn Tunyasuvunakool[1,2], Russ Bates[1,2], Augustin Žídek[1,2], Anna Potapenko[1,2], Alex Bridgland[1,2], Clemens Meyer[1,2], Simon A. A. Kohl[1,2], Andrew J. Ballard[1,2], Andrew Cowie[1,2], Bernardino Romera-Paredes[1,2], Stanislav Nikolov[1,2], Rishub Jain[1,2], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[3,4], 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[1,2☒]

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

The development of computational methods to predict three-dimensional (3D) protein structures from the protein sequence has proceeded along two complementary paths that focus on either the physical interactions or the evolutionary history. The physical interaction programme heavily integrates our understanding of molecular driving forces into either thermodynamic or kinetic simulation of protein physics[16] or statistical approximations thereof[17]. Although theoretically very appealing, this approach has proved highly challenging for even moderate-sized proteins due to the computational intractability of molecular simulation, the context dependence of protein stability and the difficulty of producing sufficiently accurate models of protein physics. The evolutionary programme has provided an alternative in recent years, in which the constraints on protein structure are derived from bioinformatics analysis of the evolutionary history of proteins, homology to solved structures[18,19] and pairwise evolutionary correlations[20–24]. This bioinformatics approach has benefited greatly from the steady growth of experimental protein structures deposited in the Protein Data Bank (PDB)[5], the explosion of genomic sequencing and the rapid development of deep learning techniques to interpret these correlations. Despite these advances, contemporary physical and evolutionary-history-based approaches produce predictions that are far short of experimental accuracy in the majority of cases in which a close homologue has not been solved experimentally and this has limited their utility for many biological applications.

In this study, we develop the first, to our knowledge, computational approach capable of predicting protein structures to near experimental accuracy in a majority of cases. The neural network AlphaFold that we developed was entered into the CASP14 assessment (May–July 2020; entered under the team name 'AlphaFold2' and a completely different model from our CASP13 AlphaFold system[25]). The CASP assessment is carried out biennially using recently solved structures that have not been deposited in the PDB or publicly disclosed so that it is a blind test

[1]DeepMind, London, UK. [2]School of Biological Sciences, Seoul National University, Seoul, South Korea. [3]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea. [4]These authors contributed equally: John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Demis Hassabis. ☒e-mail: jumper@deepmind.com; dhcontact@deepmind.com

Challenge:
3D structure prediction from sequence is difficult and costly

Enabled by:
Critical Assessment of Protein Structure prediction

Database of "gold standard" labeled data and benchmarks for algorithms to exceed Experiments online since 1994-current day

https://predictioncenter.org/

What are the benchmarks for **evidence synthesis**?

13

Article | Open access | Published: 12 July 2023

Large language models encode clinical knowledge

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, ... Vivek Natarajan + Show authors

Nature **620**, 172–180 (2023) | Cite this article

340k Accesses | **1763** Citations | **1222** Altmetric | Metrics

A Publisher Correction to this article was published on 27 July 2023

This article has been updated

Abstract

Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high. Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Here, to address these limitations, we present MultiMedQA, a benchmark combining six existing medical question answering datasets spanning professional medicine, research and consumer queries and a new dataset of medical questions searched online, HealthSearchQA. We propose a human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possible harm and bias. In addition, we evaluate Pathways Language Model[1] (PaLM, a 540-billion parameter LLM) and its instruction-tuned variant, Flan-PaLM[2] on MultiMedQA. Using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice dataset (MedQA[3], MedMCQA[4], PubMedQA[5] and Measuring Massive Multitask Language Understanding (MMLU) clinical topics[6]), including 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%. However, human evaluation reveals key gaps. To resolve this, we introduce instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars. The resulting model, Med-PaLM, performs encouragingly, but remains inferior to clinicians. We show that comprehension, knowledge recall and reasoning improve with model scale and instruction prompt tuning, suggesting the potential utility of LLMs in medicine. Our human evaluations reveal limitations of today's models, reinforcing the importance of both evaluation frameworks and method development in creating safe, helpful LLMs for clinical applications.

Benchmarking by medical QA: LLMs clearly capture medical textbook knowledge

But how do we assess how they do in practice with new data, emerging from RCTs and observational studies?

What is their false positive or hallucination rate?

14

MultimedBench

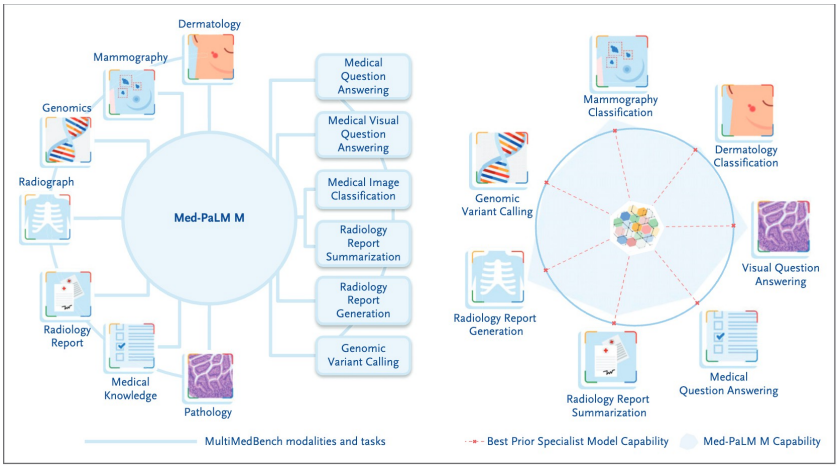


Figure 1. Med-PaLM M Overview.

A generalist biomedical artificial intelligence system should be able to handle a diverse range of biomedical data modalities and tasks, ideally using a single set of model weights to enable computationally efficient usage and elegant modeling of cross-modality interactions. To enable progress toward this overarching goal, we curated MultiMedBench, a benchmark spanning 14 diverse biomedical tasks, including question answering, visual question answering, image classification, radiology report generation and summarization, and genomic variant calling. Med-PaLM Multimodal (Med-PaLM M), our proof-of-concept for such a generalist biomedical artificial intelligence system (denoted by the shaded blue area), is competitive with or exceeds prior state-of-the-art results from specialist models (denoted by dotted red lines) on all tasks in MultiMedBench.

Pathways Learning Model (PaLM)

Sighal et al., Nature 2023
Tu et al., NEJM AI 2024

15

Table 1. Performance Comparison on MultiMedBench.*

Task Type	Modality	Dataset	Metric	SOTA†	PaLM-E (84B)	Med-PaLM M (84B)
Question answering	Text	Medical Question Answering (MedQA)	Accuracy	**86.50%**[16]	28.83%	46.11%
		Medical Multiple-Choice Question Answering (MedMCQA)	Accuracy	**72.30%**[16]	33.35%	47.60%
		PubMed Question Answering (PubMedQA)	Accuracy	**81.80%**[16]	64.00%	71.40%
Report summarization	Radiology	Medical Information Mart for Intensive Care (MIMIC)-III	ROUGE-L	**38.70%**[17]	3.30%	31.47%
			BLEU	**16.20%**[17]	0.34%	15.36%
			F1-RadGraph	**40.80%**[17]	8.00%	33.96%
Visual question answering	Radiology	Visual Question Answering Radiology (VQA-RAD)	BLEU-1	**71.03%**[18]	59.19%	69.38%
			F1	NA3	38.67%	**59.90%**
		Semantically-Labeled Knowledge-Enhanced Visual Question Answering (Slake-VQA)	BLEU-1	78.60%[19]	52.65%	**92.70%**
			F1	78.10%[19]	24.53%	**89.28%**
	Pathology	Pathology Visual Question Answering (Path-VQA)	BLEU-1	**70.30%**[19]	54.92%	70.16%
			F1	58.40%[19]	29.68%	**59.51%**
Report generation	Chest x-ray	MIMIC Chest X-ray (MIMIC-CXR)	Micro-F1-14	44.20%[20]	15.40%	**53.56%**
			Macro-F1-14	30.70%[20]	10.11%	**39.83%**
			Micro-F1-5	56.70%[21]	5.51%	**57.88%**
			Macro-F1-5	NA3	4.85%	**51.60%**
			F1-RadGraph	24.40%[14]	11.66%	**26.71%**
			BLEU-1	**39.48%**[20]	19.86%	32.31%
			BLEU-4	**13.30%**[21]	4.60%	11.31%
			ROUGE-L	**29.60%**[22]	16.53%	27.29%
			CIDEr-D	**49.50%**[23]	3.50%	26.17%
Image classification	Chest x-ray	MIMIC-CXR (5 conditions)	Macro-AUC	**81.27%**[24]	51.48%	78.35%
			Macro-F1	NA3	7.83%	**36.83%**
	Dermatology	PAD-UFES-20	Macro-AUC	NA2	63.37%	**97.27%**
			Macro-F1	NA2	1.38%	**84.32%**
	Mammography	VinDr-Mammo	Macro-AUC	64.50%[25]	51.49%	**71.76%**
			Macro-F1	NA3	16.06%	**35.70%**
		Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) (mass)	Macro-AUC	NA	47.75%	**73.09%**
			Macro-F1	NA	7.77%	**49.98%**
		Curated Breast Imaging Subset of DDSMDigital Database for Screening Mammography (CBIS-DDSM) (calcification)	Macro-AUC	NA3	40.67%	**82.22%**
			Macro-F1	**70.71%**[26]	11.37%	63.81%
	Genomics (variant calling)	PrecisionFDA (Truth Challenge V2)	Indel-F1	**99.40%**[27]	53.01%	97.04%
			SNP-F1	**99.70%**[27]	52.84%	99.32%

Tu et al., NEJM AI 2024

16

Finding and assimilating evidence is resource intensive (1 SR ~ 62 weeks, Borah et al BMJ Open 2017), with implications on resources on future studies

Can the paper screening and data extraction procedure be sped up or automated?

Elicit
Covidence
RevMan
DistillerSR

17

Automation of Systematic Reviews with Large Language Models

Christian Cao[1,†], Rohit Arora[2,†], Paul Cento[3,†], Katherine Manta[1], Elina Farahani[1], Matthew Cecere[1], Anabel Selemon[4], Jason Sang[3], Ling Xi Gong[2], Robert Kloosterman[1], Scott Jiang[5], Richard Saleh[1], Denis Margalik[1], James Lin[6], Jane Jomy[1], Jerry Xie[2], David Chen[1], Jaswanth Gorla[1], Sylvia Lee[8], Kelvin Zhang[7], Harriet Ware[9], Mairead Whelan[9], Bijan Teja[1,10], Alexander A. Leung[9], Lina Ghosn[11,12,13], Rahul K. Arora[9], Allen S. Detsky[1], Michael Noetel[14], David B. Emerson[15], Isabelle Boutron[11,12,13], David Moher[1,16,17], George Church[2], Niklas Bobrovitz[2]

[1]University of Toronto; [2]Harvard Medical School; [3]Independent Researcher; [4]McGill University; [5]University of British Columbia; [6]Massachusetts Institute of Technology; [7]University of Waterloo; [8]Mount Sinai Hospital; [9]University of Calgary; [10]St. Michael's Hospital; [11]Université Paris Cité; [12]Université Sorbonne Paris Nord; [13]Cochrane France; [14]The University of Queensland; [15]Vector Institute; [16]Ottawa Hospital Research Institute; [17]University of Ottawa.

Abstract

Systematic reviews (SRs) inform evidence-based decision making. Yet, they take over a year to complete, are prone to human error, and face challenges with reproducibility; limiting access to timely and reliable information. We developed *otto-SR*, an end-to-end agentic workflow using large language models (LLMs) to support and automate the SR workflow from initial search to analysis. We found that *otto-SR* outperformed traditional dual human workflows in SR screening (*otto-SR*: 96.7% sensitivity, 97.9% specificity; human: 81.7% sensitivity, 98.1% specificity) and data extraction (*otto-SR*: 93.1% accuracy; human: 79.7% accuracy). Using *otto-SR*, we reproduced and updated an entire issue of Cochrane reviews (n=12) in two days, representing approximately 12 work-years of traditional systematic review work. Across Cochrane reviews, *otto-SR* incorrectly excluded a median of 0 studies (IQR 0 to 0.25), and found a median of 2.0 (IQR 1 to 6.5) eligible studies likely missed by the original authors. Meta-analyses revealed that *otto-SR* generated newly statistically significant findings in 2 reviews and negated significance in 1 review. These findings demonstrate that LLMs can rapidly conduct and update systematic reviews with superhuman performance, laying the foundation for automated, scalable, and reliable evidence synthesis.

18

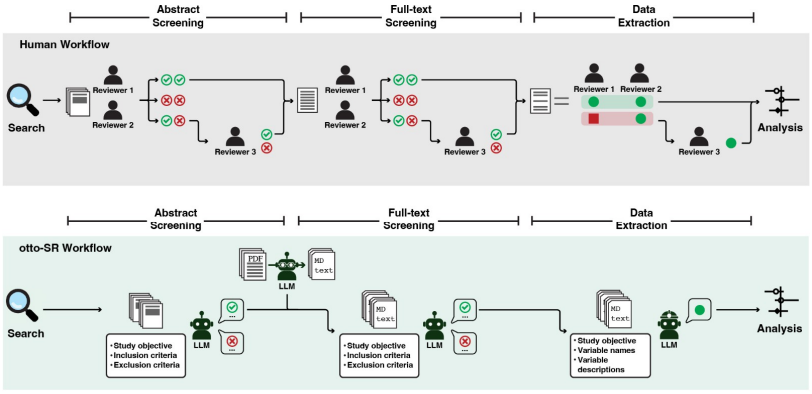


Figure 1: An automated systematic review workflow using LLMs. Infographic displaying the end-to-end SR process for humans (grey) and *otto-SR* (green). Abbreviations: Markdown file (MD)

19

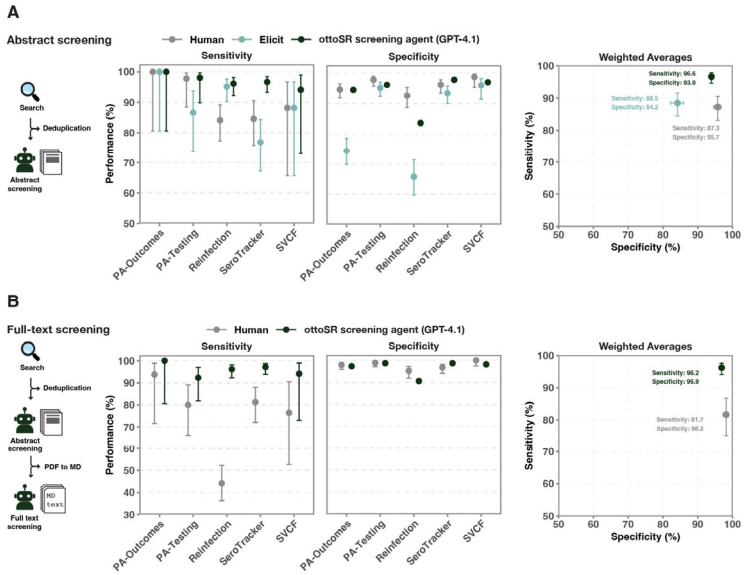


Figure 2: *otto-SR* screening agent (GPT-4.1) achieves superhuman screening sensitivity and specificity A. Diagram of *otto-SR* abstract screening agent (left), sensitivity, specificity of *otto-SR* screening agent, dual human reviewers, and Elicit, for abstract screening evaluated across five reviews (middle). Weighted averages for sensitivity and specificity across comparator groups (right). Error bars indicate 95% confidence intervals. **B.** diagram of *otto-SR* full-text screening agent (left), sensitivity, specificity, and accuracy of *otto-SR* screening agent evaluated across five reviews, and dual human reviewers for full-text screening evaluated across five reviews (middle). Weighted average for sensitivity and specificity across *otto-SR* (five reviews) and dual human (four reviews) (right). Error bars indicate 95% confidence intervals.

20

NEJM AI

NEJM AI 2024; 1 (8)
DOI: 10.1056/AIoa2400196

ORIGINAL ARTICLE

Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis

Weixin Liang, M.S.,[1] Yuhui Zhang, M.S.,[1] Hancheng Cao, Ph.D.,[1] Binglu Wang, M.S.,[2] Daisy Yi Ding, M.S.,[1] Xinyu Yang, B.E.,[4] Kailas Vodrahalli, M.S.,[1] Siyu He, Ph.D.,[2] Daniel Scott Smith, Ph.D.,[5] Yian Yin, Ph.D.,[6] Daniel A. McFarland, Ph.D.,[3] and James Zou, Ph.D.[1,2,7]

Received: February 22, 2024; Revised: May 17, 2024; Accepted: May 28, 2024; Published: July 17, 2024

Abstract

BACKGROUND Expert feedback lays the foundation of rigorous research. However, the rapid growth of scholarly production challenges the conventional scientific feedback mechanisms. High-quality peer reviews are increasingly difficult to obtain.

METHODS We created an automated pipeline using Generative Pretrained Transformer 4 (GPT-4) to provide comments on scientific papers. We evaluated the quality of GPT-4's feedback through two large-scale studies. We first quantitatively compared GPT-4's generated feedback with human peer reviewers' feedback in general scientific papers from 15 *Nature* family journals (3096 papers in total) and the International Conference on Learning Representations (ICLR) machine learning conference (1709 papers). To specifically assess GPT-4's performance on biomedical papers, we also analyzed a subset of 425 health sciences papers from the *Nature* portfolio and a random sample of 666 submissions to *eLife*. Additionally, we conducted a prospective user study with 308 researchers from 110 institutions in the fields of artificial intelligence and computational biology to understand how researchers perceive feedback generated by our system on their own papers.

RESULTS The overlap in the points raised by GPT-4 and by human reviewers (average overlap of 30.85% for *Nature* journals and 39.23% for ICLR) is comparable with the overlap between two human reviewers (average overlap of 28.58% for *Nature* journals and 35.25% for ICLR). Results on *eLife* and a subset of health sciences papers as categorized by the *Nature* portfolio show similar patterns. In our prospective user study, more than half (57.4%) of the users found GPT-4-generated feedback helpful/very helpful, and 82.4% found it more beneficial than feedback from at least some human reviewers. We also identify several limitations of large language model (LLM)-generated feedback.

Mr. Liang, Mr. Zhang, and Dr. Cao contributed equally to this article.

The author affiliations are listed at the end of the article.

Dr. Zou can be contacted at jamesz@stanford.edu or at 350 Jane Stanford Way, Room 369, Stanford, CA 94305.

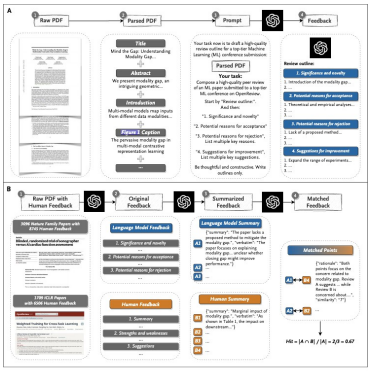


Figure 1. Characterizing the Capability of LLMs in Providing Helpful Feedback to Researchers.

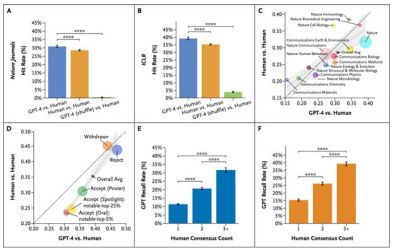


Figure 2. Retrospective Analysis of LLMs and Human Scientific Feedback.

21

Key observations:
AI is getting better, but advances are resource and capital intensive

- Much of the reduction of practice by industry (e.g., access to infrastructure, energy, and human resources)
- Closed vs. open source LLMs: closed versions have dominated the market, but gap shrinking. What are the implications in evaluation?
- "Human in the loop" approaches are proliferating
- Automating procedures may lead to more false leads in the literature
- Success in AI deployment is related to availability of clear benchmarks (e.g., protein folding, disease screening)
- How will we evaluate in research and practice?
- How will consumers use and evaluate new technologies? (Not discussed here)

22