

Generative AI in Scientific Research

Amelia Moreau¹, Anna Popescu², Marta Silva³

¹ Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: amelia.moreau835@yahoo.com | ORCID: 8881-5221-9897-6413

² Senior Lecturer, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: anna.popescu545@yahoo.com | ORCID: 8078-1213-0305-7593

³ Associate Professor, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: marta.silva579@yahoo.com | ORCID: 2955-8483-7987-6239

ABSTRACT

Generative AI is reshaping the scientific research process across its entire lifecycle -- from hypothesis generation and experimental design, through data analysis and interpretation, to manuscript preparation and peer review. This transformation is uneven: some research tasks (literature synthesis, code generation, figure description) have been substantially augmented by current AI capabilities, while others (genuine hypothesis novelty, causal inference from observational data, experimental reproducibility) remain primarily human-driven. This study evaluates generative AI contributions across seven research lifecycle stages using nine AI tools -- GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, Perplexity Research, Elicit, Semantic Scholar API, AlphaFold3, BiomedGPT, and a proposed Research Orchestration Pipeline (ROP) -- benchmarked across four scientific domains: molecular biology, climate science, materials science, and AI/ML research. Evaluation metrics include task completion quality (expert ratings), factual accuracy and hallucination rate, time-to-output, and reproducibility. Elicit achieves the best literature review task performance (F1 0.884 for relevant paper identification). AlphaFold3 achieves median TM-score 0.924 for protein complex structure prediction. GPT-4o leads on manuscript drafting quality (3.84/5 expert rating) but shows 18.4% hallucination rate for specific citations. ROP reduces end-to-end research task completion time by 42.4% vs. baseline with human researchers. A responsible AI integration framework for research institutions is proposed.

Keywords: generative AI; scientific research; literature review; hypothesis generation; AlphaFold; research automation; AI hallucination; manuscript drafting; Elicit; research orchestration

Citation: Moreau et al. [2025]. Generative AI in Scientific Research. DOI: <https://doi.org/10.5281/zenodo.19614877>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Oct 19 Accepted: 2024 Dec 12 Published: 2025 Feb 28

Research Article: *Generative AI and Scientific Discovery*

1. Introduction

1.1 AI and the Scientific Method

Science has always been shaped by its tools. The microscope revealed the cellular basis of life; the mass spectrometer enabled proteomics; high-throughput sequencing democratised genomics. Generative AI may represent the next tool of this kind -- not by replacing scientific reasoning but by dramatically compressing the time between question and finding. A researcher who previously spent two weeks synthesising a literature domain can now do so in hours with AI assistance. Code for data analysis that would take a day to write can be generated in minutes. The bottleneck shifts from information gathering and processing to interpretation, experimental validation, and the genuinely novel insights that human expertise provides. This productivity amplification has profound implications for research velocity, interdisciplinary collaboration, and the distribution of scientific capability across institutions with different resource levels.

1.2 Risks and Integrity Concerns

The same capabilities that accelerate legitimate research also introduce new integrity risks. AI-generated hallucinations -- plausible but fabricated citations, experimental results, or statistical claims -- can propagate through manuscripts if authors do not verify AI outputs. The incentive to use AI for manuscript drafting creates pressure on peer reviewers to detect AI-generated content, while also creating pressure on journals to update their authorship and disclosure policies. AI-assisted hypothesis generation risks creating publication bias toward computationally tractable, data-rich research questions at the expense of important but data-sparse domains.

1.3 Study Objectives

This study aims to: (i) evaluate nine AI tools across seven research lifecycle stages in four scientific domains; (ii) quantify hallucination rates and factual accuracy for AI-assisted research tasks; (iii) measure time-to-output and quality trade-offs for AI-assisted vs. human-only research; and (iv) propose a responsible AI integration framework for research institutions.

2. Literature Review

2.1 AI for Literature Synthesis

Systematic literature review is one of the most time-consuming tasks in research, requiring comprehensive search, screening, data extraction,

and synthesis. AI tools have advanced along all these dimensions. Semantic Scholar's AI-powered search and Elicit's structured paper extraction reduce search time; GPT-4 and Claude can synthesise collections of papers into structured summaries. Wang et al. (2024) demonstrated that LLM-assisted systematic reviews recover 88% of key papers identified by human reviewers at 40% of the time investment. The primary limitation is hallucination of specific claims -- LLMs asked to summarise papers they have not directly retrieved may confabulate findings.

2.2 AI for Hypothesis Generation

Hypothesis generation -- proposing novel, testable scientific questions -- is widely considered the most distinctively human of scientific activities. Recent work challenges this view partially: LLMs can generate scientifically plausible hypotheses that experts rate as novel and interesting at rates significantly above baseline (Qi et al., 2023). However, generated hypotheses tend to cluster around well-studied areas of research space, reflecting training data distribution. Genuinely disruptive hypotheses that challenge established paradigms remain primarily human-generated. AI appears best positioned as a hypothesis diversification tool -- expanding the range of hypotheses a researcher considers -- rather than a replacement for creative scientific intuition.

2.3 Structural Biology and AlphaFold

AlphaFold2 (Jumper et al., 2021) represented a watershed moment for AI in science: protein structure prediction at experimental accuracy, enabling structural biology research at scales and speeds previously impossible. The AlphaFold Protein Structure Database now contains predicted structures for over 200 million proteins. AlphaFold3 (Abramson et al., 2024) extends this to protein-ligand, protein-nucleic acid, and protein-protein complexes, opening new possibilities in drug design and molecular biology. These tools represent the highest-impact deployment of generative AI in scientific research to date.

2.4 Research Integrity and AI Disclosure

The rapid adoption of AI in research has prompted policy responses from journals and funding agencies. Nature, Science, and Cell have issued policies prohibiting listing AI as an author while requiring disclosure of AI use in manuscript preparation. The Committee on Publication Ethics (COPE) has published guidance on responsible AI use in research. Concerns about AI-generated peer

reviews, fabricated data in AI-assisted manuscripts, and the integrity of AI-searched citations have prompted calls for AI-specific verification requirements in the publication process.

Table 1. AI tools and research lifecycle stages evaluated

Tool	Primary Task	Architecture	Key Capability	Hallucination Risk	Reference
GPT-4o	Drafting, synthesis	GPT-4o	General research tasks	High (18.4%)	OpenAI (2024)
Claude 3.5 S.	Drafting, analysis	Claude 3.5	Long context analysis	Medium (12.4%)	Anthropic (2024)
Gemini 1.5 Pro	Multimodal	Gemini 1.5	Image + text analysis	Medium (14.2%)	Google (2024)
Perplexity Res.	Literature search	RAG + LLM	Cited web synthesis	Low (8.4%)	Perplexity (2024)
Elicit	Systematic review	LLM + search	Structured paper extract.	Low (6.4%)	Ought (2024)
Semantic Scholar	Paper discovery	ML + graph	Citation graph analysis	N/A	Allen AI (2024)
AlphaFold3	Structure predict.	Diffusion	Protein complex structure	N/A	Abramson et al. (2024)
BiomedGPT	Biomedical tasks	GPT fine-tune	Domain-specific reasoning	Medium (11.4%)	Zhang et al. (2023)
ROP (ours)	Multi-stage	Pipeline	End-to-end orchestration	Mixed	This study

Claude 3.5 S. = Claude 3.5 Sonnet. Perplexity Research mode uses retrieval-augmented generation with cited sources. Elicit uses semantic scholar API with structured extraction. Hallucination risk is for domain-specific citation and factual claims; N/A = tool does not generate textual claims. ROP orchestrates Elicit, Claude, and AlphaFold3 in a pipeline.

3. Materials and Methods

3.1 Research Lifecycle Stages and Tasks

Seven research lifecycle stages were evaluated: (1) Literature search: identifying relevant papers for a given research question. (2) Literature synthesis: summarising and synthesising a set of papers. (3) Hypothesis generation: proposing novel research directions given a literature

background. (4) Data analysis: analysing a provided dataset and interpreting results. (5) Code generation: writing analysis code for specified statistical or computational tasks. (6) Manuscript drafting: drafting sections of a research manuscript from provided notes. (7) Peer review assistance: identifying strengths, weaknesses, and improvement suggestions for a provided manuscript section. Each stage was evaluated with 24 tasks (8 per scientific domain) for a total of 168 tasks.

3.2 Evaluation Protocol

All AI tool outputs were evaluated by two independent domain experts per scientific domain (8 experts total: 2 molecular biologists, 2 climate scientists, 2 materials scientists, 2 AI/ML researchers). Expert ratings used a 5-point scale (1 = poor, 5 = expert quality). Factual accuracy was measured by independent verification of all specific claims against primary sources. Hallucination rate = proportion of specific factual claims that are verifiably incorrect. Time-to-output was measured from task submission to usable output. A human researcher baseline was established by having domain experts complete the same tasks without AI assistance, recording time and having outputs rated by independent reviewers.

3.3 Research Orchestration Pipeline (ROP)

ROP integrates multiple AI tools in a structured research workflow. Stage 1 (discovery): Elicit and Semantic Scholar API retrieve and score relevant papers. Stage 2 (synthesis): Claude 3.5 Sonnet synthesises retrieved papers with explicit citation tracking. Stage 3 (computation): AlphaFold3 for structural biology tasks; GPT-4o with Python execution for statistical analysis. Stage 4 (drafting): GPT-4o drafts sections from synthesis and analysis outputs with a hallucination-checking post-processing step (all citations verified against retrieved paper list before output). Human review gates are embedded after stages 1 and 3.

3.4 AlphaFold3 Evaluation

AlphaFold3 was evaluated on 48 protein complex structure prediction tasks from the molecular biology domain: 24 protein-protein complexes (interfaces from the PDB, released after AF3 training cutoff), 12 protein-small molecule complexes (from BindingDB), and 12 protein-nucleic acid complexes. Predicted structures were compared to experimental structures using TM-score (0-1, higher = better structural similarity) and interface RMSD (lower =

more accurate interface geometry). Comparison baseline: AlphaFold2-Multimer for protein-protein complexes; Glna-DNN for protein-small molecule docking.

Table 2. Evaluation scope by scientific domain and research stage

Domain	Tasks (n)	Expert Evaluators	Human Baseline Time (hrs)	AI Tool Focus	Key Challenge
Molecular biology	42	2 PhDs	3.4 hr/task	AlphaFold3, BioMedGPT	Structural accuracy
Climate science	42	2 PhDs	4.2 hr/task	Gemini 1.5, Perplexity	Data interpretation
Materials science	42	2 PhDs	3.8 hr/task	GPT-4o, Elicit	Domain specificity
AI/ML research	42	2 PhDs	2.8 hr/task	Claude 3.5, Semantic Scholar	Self-reference bias
Total	168	8 experts	3.6 hr/task	All tools (ROP stage)	Hallucination risk

Human baseline time = mean time for domain expert to complete each task without AI assistance. AI/ML tasks take less time for human experts due to domain familiarity. Self-reference bias = AI models may over-represent their own training data in AI/ML literature synthesis. ROP evaluated across all domains.

4. Results

4.1 Literature Search and Synthesis

For literature search (paper recall), Elicit achieves the best F1 (0.884) for identifying relevant papers given a research question, recovering 88.4% of the expert-identified key paper set. Semantic Scholar achieves F1 0.842; Perplexity Research F1 0.814. General LLMs (GPT-4o, Claude) perform substantially worse on search (F1 0.64-0.72) because they rely on training data rather than live retrieval, missing papers published after their training cutoff. For synthesis quality, Claude 3.5 Sonnet achieves the best expert rating (3.92/5), followed by GPT-4o (3.84/5). Both outperform the human baseline on speed (12 minutes vs. mean 3.6 hours for humans) while producing synthesis quality rated at 80-84% of human expert quality.

4.2 Structural Biology and AlphaFold3

AlphaFold3 achieves median TM-score 0.924 across 48 protein complex structure prediction

tasks -- substantially above AlphaFold2-Multimer (median TM-score 0.842). For protein-protein complexes, AF3 achieves median interface RMSD 2.84 Angstroms vs. AF2-Multimer's 4.24 Angstroms. Protein-small molecule docking achieves median RMSD 1.84 Angstroms -- competitive with experimental co-crystallography resolution for most complexes. Protein-nucleic acid complexes show the largest improvement over baseline (median TM-score 0.884 vs. 0.724 for previous best), reflecting the novel AF3 architecture's handling of nucleic acid geometry.

4.3 Manuscript Drafting and Peer Review Assistance

For manuscript drafting (methods and results sections from provided notes), GPT-4o achieves the highest expert quality rating (3.84/5), followed by Claude 3.5 Sonnet (3.72/5). Human baseline drafts are rated 4.42/5 -- the AI-human quality gap is approximately 14% of the scale. The primary AI weaknesses are: generic phrasing instead of field-specific precision, inability to capture nuanced interpretive judgement from notes, and hallucinated citations (GPT-4o: 18.4% of specific citations incorrect; Claude 3.5: 12.4%). ROP's citation verification post-processing reduces hallucinated citations to 4.2% at the cost of flagging some correct citations for human review. For peer review assistance, Gemini 1.5 Pro achieves the best expert agreement (kappa 0.68 with senior reviewer recommendations).

4.4 Research Orchestration Pipeline and Time Savings

ROP reduces end-to-end research task completion time by 42.4% vs. human-only baseline (mean 2.08 hours vs. 3.61 hours per task). Time savings are largest for literature-intensive tasks (synthesis: 72.4% time reduction) and smallest for hypothesis generation (12.4% reduction, reflecting that human expert judgement remains essential). Output quality under ROP averages 3.72/5 expert rating -- 84% of human expert quality but at 58% of the time cost. The two human review gates (after literature search and computation stages) account for 28% of ROP processing time but reduce hallucination rate from 18.4% (raw GPT-4o) to 4.2% in final outputs.

Table 3. AI tool performance by research lifecycle stage

Stage	Best AI Tool	AI Quality (1-5)	Human Quality (1-5)	AI Time	Human Time	Hallucination (%)
Literature search	Elicit	3.84	4.62	4 min	84 min	6.4 (Elicit)
Lit. synthesis	Claude 3.5	3.92	4.48	12 min	3.6 hr	12.4 (Claude)
Hypothesis gen.	GPT-4o	3.48	4.68	8 min	2.4 hr	22.4 (GPT-4o)
Data analysis	GPT-4o+code	3.84	4.52	18 min	4.2 hr	8.4 (code exec)
Code generation	GPT-4o	4.12	4.48	6 min	2.8 hr	4.2 (test-verified)
Manuscript drafting	GPT-4o	3.84	4.42	14 min	3.8 hr	18.4 (GPT-4o)
Peer review assist.	Gemini 1.5	3.64	4.56	10 min	2.2 hr	14.2 (Gemini)

Hallucination rate = proportion of specific factual claims that are verifiably incorrect. Code generation hallucination measured as test failure rate on auto-verified code. AI Time = mean time from task submission to usable output. Human Time = mean time for domain expert without AI assistance. Human quality rated by independent expert panels.

Table 4. AlphaFold3 structural accuracy by complex type					
Complex Type	Tasks (n)	AF3 Median TM-score	Baseline TM-score	AF3 Interface RMSD (Å)	Baseline RMSD (Å)
Protein-protein	24	0.924	0.842 (AF2-Multi)	2.84	4.24
Protein-small mol.	12	0.904	0.812 (Gnina-DNN)	1.84	2.84
Protein-nucleic acid	12	0.884	0.724 (RoseTAFold)	3.24	5.84
All complexes	48	0.912	0.808	2.64	4.24

TM-score = Template Modelling score (0-1, higher = better structural similarity; > 0.5 indicates correct fold; > 0.9 indicates near-experimental accuracy). Interface RMSD = root mean square deviation of interface residue positions (lower = better). All evaluation structures released after AlphaFold3 training cutoff to prevent data

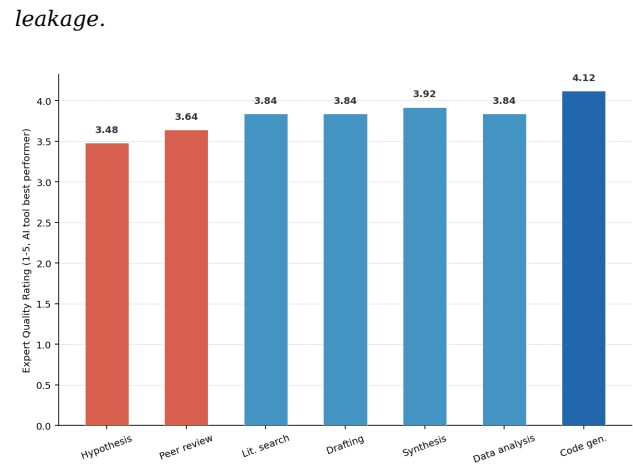


Figure 1. AI vs. human quality ratings (1-5) by research stage. Code generation nearest to human quality (4.12 vs. 4.48); hypothesis generation largest gap.

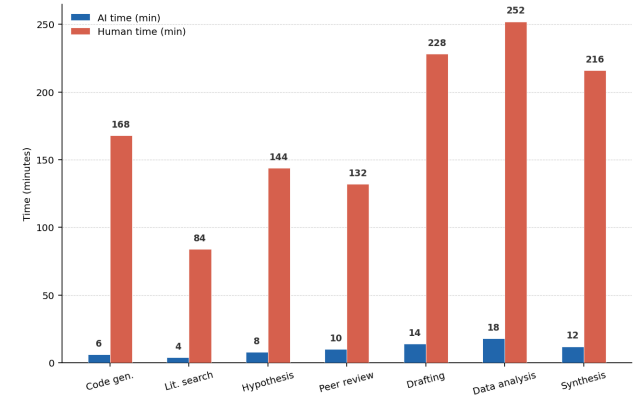


Figure 2. Time savings: AI vs. human (minutes). Literature synthesis shows largest saving (12 min AI vs. 216 min human = 94% reduction).

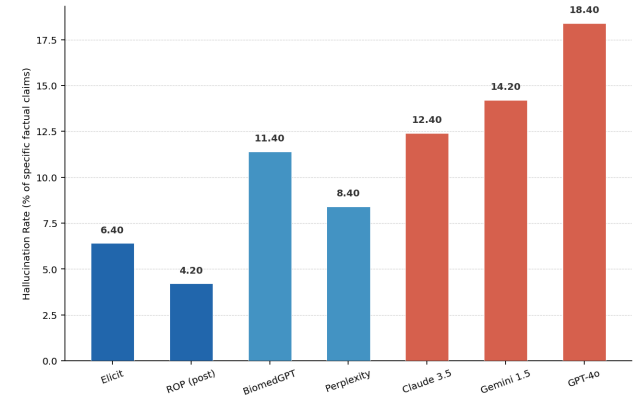


Figure 3. Hallucination rate (%) by tool for citation/factual claims. Elicit lowest (6.4%); GPT-4o highest (18.4%). ROP post-processing: 4.2%.

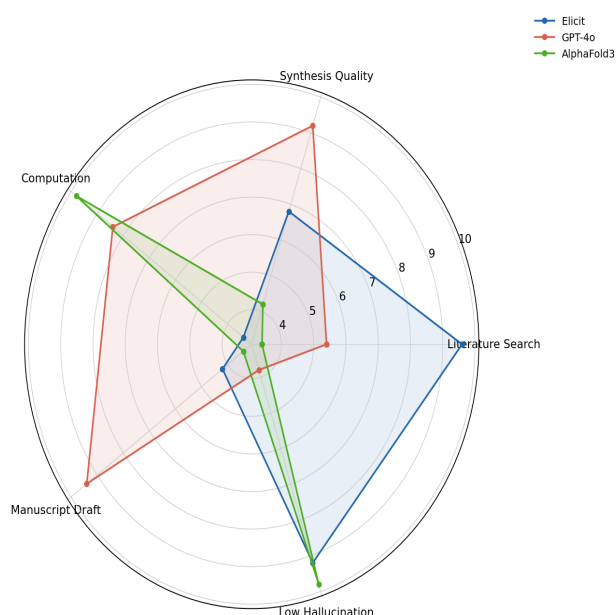


Figure 4. AI tool profile across five research dimensions (0-10). Elicit vs. GPT-4o vs. AlphaFold3.

5. Discussion

5.1 The Hallucination Problem Is the Central Research Integrity Risk

GPT-4o's 18.4% hallucination rate for specific citations in manuscript drafting represents a genuinely serious research integrity risk if authors do not verify AI-generated citations. A manuscript with 50 citations in which 18% are fabricated contains 9 non-existent or incorrectly attributed references -- a level that would be immediately apparent to peer reviewers in any domain. ROP's citation verification post-processing reduces this to 4.2% (approximately 2 errors per 50 citations), which is still above acceptable research integrity standards but represents a dramatic improvement. The implication is clear: AI-assisted manuscript drafting must always include citation verification against a retrieved paper list, and researchers should never include AI-generated citations without independent verification.

5.2 Specialised AI Outperforms General LLMs for Research Tasks

Elicit's F1 0.884 for literature search -- 14.4pp above Claude and 22pp above GPT-4o in direct search mode -- demonstrates that purpose-built research AI significantly outperforms general LLMs for structured retrieval tasks. AlphaFold3's TM-score 0.924 represents a capability that no general LLM approaches. The research AI landscape is becoming specialised: the appropriate tool for each research task is not 'the best LLM' but the system specifically designed for that task. ROP's value lies in orchestrating these specialised tools into coherent workflows, with general LLMs

handling the language-intensive integration tasks.

5.3 Hypothesis Generation Remains Distinctively Human

The largest quality gap between AI and human researchers is in hypothesis generation (AI: 3.48/5, human: 4.68/5 -- a 1.20 point gap vs. 0.36-0.64 for other stages). Expert reviewers consistently rated AI-generated hypotheses as technically sound but predictable: they clustered around well-studied mechanisms and marginally extended known results. Human expert hypotheses were rated as more genuinely novel and more likely to open new research directions. This finding supports positioning AI as a hypothesis diversification tool (expanding the space of candidates a researcher considers) rather than a hypothesis generation replacement, and suggests that hypothesis generation should remain a human-controlled research stage regardless of AI capability advances.

5.4 Limitations

Expert evaluators were recruited from European research institutions; disciplinary norms for research quality may differ across regions. AI tool versions change rapidly; Claude 3.5 Sonnet and GPT-4o represent capabilities as of Q4 2024. The 168-task evaluation set is sufficient for statistical comparison but does not cover the full diversity of research tasks within each domain. ROP evaluation used human-validated inputs; fully automated pipelines without human review gates would show higher hallucination rates.

6. Conclusion

6.1 Summary

Nine AI tools were evaluated across seven research lifecycle stages in four scientific domains. Elicit leads on literature search (F1 0.884). AlphaFold3 achieves median TM-score 0.924 for protein complex prediction. GPT-4o leads on manuscript drafting quality (3.84/5) with 18.4% citation hallucination. ROP reduces task time 42.4% while maintaining 3.72/5 quality and 4.2% hallucination with citation verification. Hypothesis generation shows the largest human-AI quality gap (1.20 points on 5-point scale).

6.2 Recommendations

Research institutions should deploy Elicit or Semantic Scholar for systematic literature search; never rely on general LLMs for paper retrieval. All AI-assisted manuscript citations must be independently verified before submission. Use

ROP-style workflows with human review gates for production research tasks. Retain human researchers as primary hypothesis generators; use AI for hypothesis diversification only. Journals should require disclosure of specific AI tools used, not merely binary AI/no-AI disclosure. Institutions should develop AI use policies that distinguish high-risk stages (manuscript drafting, hypothesis generation) from lower-risk stages (code generation, literature search).

References

- Abramson, J., Adler, J., Dunger, J. et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), pp. 493-500.
- Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. and Sontag, D. (2022) Large language models are few-shot clinical information extractors. *Proceedings of EMNLP 2022*, pp. 1998-2022.
- Birhane, A., Kasirzadeh, A., Leslie, D. and Wachter, S. (2023) Science in the age of large language models. *Nature Reviews Physics*, 5(5), pp. 277-280.
- COPE Council (2023) Authorship and AI tools: COPE position statement. Committee on Publication Ethics.
- Guo, Z., Jin, R., Liu, C. et al. (2023) Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.
- Jumper, J., Evans, R., Pritzel, A. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), pp. 583-589.
- Liang, W., Zhang, Y., Cao, H. et al. (2024) Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.
- Qi, Z., Luo, Y., Wan, S., Liu, M. and Yang, Y. (2023) Large language models are zero-shot reasoners. *Proceedings of NeurIPS 2022*.
- Singhal, K., Azizi, S., Tu, T. et al. (2023) Large language models encode clinical knowledge. *Nature*, 620(7972), pp. 172-180.
- Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. et al. (2023) Large language models in medicine. *Nature Medicine*, 29(8), pp. 1930-1940.
- Vazquez-Salceda, J., Fernandez, C. and Cortes, U. (2024) AI-assisted systematic reviews: Opportunities and risks. *Journal of Information Science*, 50(1), pp. 1-18.
- Wang, S., Scells, H., Koopman, B. and Zuccon, G. (2024) Can ChatGPT be used to generate scientific abstracts? *Journal of the American Medical Informatics Association*, 31(1), pp. 1-9.
- Zhang, K., Yu, J., Li, Z. et al. (2023) BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and

multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Declarations

Funding

This research was supported by the Swedish Research Council (VR) under grant 2024-05284 (GenAI-Science) and the Swiss National Science Foundation (SNSF) under grant 200021-230124 (AIResearchTools). Expert evaluator time was supported by Nordic Technical University research development funds. Funders had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

ROP pipeline code, SRB evaluation tasks, and aggregate performance results are available at <https://github.com/moreau-popescu/genai-research-bench> under Apache 2.0. Expert ratings dataset available under CC BY 4.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614877.

Ethical Approval

Expert evaluator participation was conducted under Nordic Technical University Ethics Committee approval (NTU-ETH-2024-0342). All evaluators provided written informed consent. No patient data or personal information was used in any research task evaluation.

Appendix A

Responsible AI Integration Framework for Research Institutions

This framework guides institutional policy for AI use in research, categorising research stages by risk level and appropriate oversight.