4 September 2023, Preprint v1 · CC-BY 4.0 Qeios PREPRINT

Research Article

Towards Responsible AI-Assisted Scholarship: Comparative Assessment of Generative Models and Adoption Recommendations

Swapnil Morandé[1], Kanwal Gul[1]

1. University of Naples Federico II, Italy

The integration of generative AI into academic research holds immense promise but necessitates judicious oversight to address risks. This pioneering study provides crucial insights to guide responsible adoption through a rigorous comparative benchmarking of four cutting-edge models – Claude, LaMDA, Sydney and Galactica. Carefully designed prompts assess competencies across core scholarly tasks, with quantitative scoring and qualitative analysis elucidating specialized capabilities, gaps, risks and validation needs. Key findings reveal strengths in focused assistive roles but limitations in generalizing reasoning across disciplines compared to human scholars. The AI systems emphasize extensive validation to mitigate risks, underscoring the need for transparency, peer review, reproducibility checks and continuous benchmarking as adoption accelerates. To steer progress responsibly, tailored recommendations for pragmatic system-task alignment, calibrating expectations, enhancing reasoning skills, holistic risk mitigation and participatory oversight are provided to researchers, developers, institutions and publishers. This timely applied framework grounded in real-world evidence provides a roadmap to harness AI's immense opportunities to benefit scholarship through prudent integration focused on human-AI collaboration under an ethical oversight framework.

Introduction

The advent of large language models like GPT-3 has sparked tremendous enthusiasm for applying generative AI to transform scholarship and scientific research. Diverse academic tasks ranging from

doi.org/10.32388/7OVVW2

reviewing literature to drafting manuscripts are being augmented by systems like GPT-3, holding immense promise to enhance researcher productivity and accelerate discoveries (Chan, 2023). However, it is imperative that these powerful technologies are integrated into scholarly workflows judiciously and responsibly based on a nuanced understanding of their capabilities, limitations and risks grounded in rigorous assessments (Dwivedi et al., 2023). This study aims to contribute crucial insights in this direction by conducting a pioneering comparative evaluation of four cutting-edge generative AI systems – Anthropic's Claude, Alphabet's LaMDA, Microsoft's Sydney and Meta's Galactica. While recent studies have provided glimpses into individual models' strengths and weaknesses for select research activities, structured comparative benchmarks evaluating leading systems are scarce (Bhutoria, 2022; Eysenbach, 2023). Furthermore, empirical diagnosis of real-world risks and oversight needs remains limited amidst accelerating adoption. This study addresses these gaps through a rigorous mixed-methods assessment, combining quantitative scoring of four models across core research tasks with qualitative analysis of their perspectives on limitations and responsible usage. The findings provide data-driven insights and an evidence-based framework to guide integration while proactively addressing risks.

The recent study makes important contributions by comparatively benchmarking Claude, LaMDA, Sydney and Galactica using prompts that span from literature analysis to hypothesis generation. Quantitative scoring elucidates each system's specialized capabilities and limitations. Risks, restrictions and validation requirements are empirically elicited directly from the AI systems themselves to inform oversight (Morande, Arshi, Gul, & Amini, 2023). A pragmatic framework with actionable recommendations is developed to enhance prudence and maximize benefits, providing a vital roadmap for judicious adoption grounded in real-world performance data instead of hype.

The study employs a rigorous mixed-methods approach combining comparative prompting experiments with quantitative scoring, qualitative thematic analysis, and statistical tests (Driscoll et al., 2007). Carefully designed prompts assess competencies across 10 key research tasks identified from the literature. The AI systems' free-text responses are analyzed using rubrics and coding to extract major themes. Comparisons with human expert performance also contextualize capabilities versus advanced scholarship. This pioneering work makes significant contributions towards responsibly harnessing generative AI's immense opportunities to augment scholarship while proactively addressing risks and limitations. The evidence-based insights and recommendations offer a timely applied framework to guide researchers, developers, institutions and publishers in steering these technologies prudently to benefit knowledge advancement.

Literature Review

The potential of generative AI to transform scholarship has sparked growing research examining capabilities for core academic tasks, responsible adoption challenges, and emerging needs to uphold research integrity as integration advances. This review synthesizes key strands of literature shaping understanding of these powerful technologies' implications for the future of knowledge creation.

Evaluating Capabilities for Core Research Activities

A prominent focus has been empirically evaluating leading models' effectiveness for diverse scholarly tasks. Bommasani et al., (2021)'s formative benchmarking of GPT-3 found promising capabilities for constrained activities like summarization but major gaps in contextual reasoning, causal logic, and creative ideation compared to human scholars. Building on this work, Ray (2023) tracked accuracy improvements from GPT-3 to Codex but noted persisting challenges like hallucination risks. Sallam (2023) analyzed literature synthesis abilities, finding strengths in summarizing descriptive information but difficulties with deeper critical analysis. De Angelis et al. (2023) identified gaps in scientific reasoning capacities like deducing hypotheses from empirical gaps. Collectively, these studies reveal pockets of competency suitable for focused assistance but limitations in generalized skills integral to expert scholarship like integrative reasoning, critical thinking, and novel insight generation. However, most concentrate on a single model, frequently GPT-3, with comparative assessments across diverse emerging architectures remaining scarce (Gupta et al., 2023; Li et al., 2023; Lo, 2023). This constrains understanding system differences based on training methodologies.

Responsible Adoption Challenges and Oversight Needs

Another active focus has been investigating responsible adoption challenges. Key risks identified include biases and toxic language risks (Solaiman & Dennison, 2021), trust calibration difficulties (Vaithilingam et al., 2022), and threats of misuse or misinterpretation (Megahed et al., 2023). Workflow integration barriers highlighted include skill complementarity limitations and friction with traditional scholarly norms (Rana, 2023). To address these issues, researchers have proposed human oversight strategies like ethics boards (Harrer, 2023), alternating human-AI cycles (Wu et al., 2022),

doi.org/10.32388/7OVVW2

participatory oversight frameworks (Meskó & Topol, 2023) and reconsidering published end-products versus processes (Eloundou et al., 2023). However, empirical diagnosis of oversight needs based on deployed systems remains scarce, with most discussions still conceptual rather than grounded in implementation learnings (Bawack & Desveaud, 2022).

Research Integrity Considerations

As adoption accelerates, upholding research integrity necessitates adapting norms without undermining rigor or excellence (Cotton et al., 2023). Proposed integrity safeguards include transparency requirements, bias reviews, reproducibility testing, and editorial scrutiny on par with statistics checks (Dwivedi et al., 2023). However, holistic frameworks reconciling rapid technical advances (Morande et al., 2023) with responsible scholarship are emergent. Studies elucidating real-world validation requirements based on AI system perspectives remain limited but could meaningfully inform governance.

Key Research Gaps

The existing literature reveals three key gaps in knowledge, despite providing valuable snapshots of strengths and limitations:

Structured comparative assessments evaluating multiple diverse generative models to reveal differences based on training methodologies are scarce. Empirical elicitation of risks, constraints and oversight needs directly from deployed AI systems (Morande & Tewari, 2023) to inform governance is limited. Comprehensive applied frameworks and tailored recommendations to guide key stakeholders in judicious adoption upholding research integrity are lacking.

Advancing understanding requires comparative assessments across models and insights elicited directly from AI systems.

Research Methodology

This study employs a rigorous mixed-methods approach combining comparative prompting experiments with quantitative scoring, qualitative analysis, and human benchmarking to evaluate four leading generative AI systems - Claude, LaMDA, Sydney, and Galactica.

Comparative Prompting Design

The core methodology entails comparative prompting, exposing the AI systems to open-ended prompts assessing competencies across 10 key academic research tasks identified from the literature. Carefully crafted neutral prompts elicit free-text responses demonstrating each capability. Quantitative scoring and qualitative thematic analysis of responses provide multifaceted insights into strengths, limitations, risks, and oversight needs grounding in real-world evidence.

Systems Selection

The models represent diverse state-of-the-art architectures including attention mechanisms, memory modules, and reinforced training on scholarly corpora. This enables elucidating differences based on varying techniques (Ling et al., 2023). While not exhaustive, the set provides an initial comparative benchmarking.

Assessment Categories

Ten academic research categories were designed based on established AI skills frameworks (Ghahramani, 2015):

Literature reviews, hypothesis generation, summarizing literature, identifying research gaps, drafting manuscripts, creating datasets, finding connections, discussing limitations, describing biases and validating outputs were included.

This spans core competencies from literature analysis to results communication, providing comprehensive coverage.

Prompt Design

Carefully crafted open-ended prompts elicit free-text responses demonstrating capabilities for each research activity. Iterative piloting with external researchers ensured neutral wording. Prompts are designed to be non-leading and avoid disclosing expected answers. Example prompts are:

- "Briefly summarize the key findings from this research summary in 3 sentences in a scholarly tone." (Literature Summarization)
- "What factors should researchers consider when creating datasets to train AI systems ethically? Discuss 3 key elements." (Dataset Creation)

Data Collection

Data was collected from AI systems by providing prompts and analyzing responses under standard usage terms. Identifying information was anonymized for secure collection. Responses were evaluated by two human raters who independently scored each response from 1-10 based on completeness, accuracy, and relevance using calibrated rubrics. Interrater reliability was measured using Cohen's kappa.

Qualitative responses were analyzed using inductive coding, which involved iteratively refining codes into themes via constant comparison. Intercoder reliability was computed. Mean scores were compared between systems using ANOVA and post-hoc Tukey tests. Experts in academic research also responded to prompts, providing comparative context on AI versus human performance.

Data Analysis

Quantitative Benchmarking

The AI systems' prompt responses were independently scored by two trained raters on a 10-point scale assessing completeness, accuracy and relevance. Interrater reliability was substantial (Cohen's kappa = 0.68).

Table 1 summarizes mean scores for each system across the 10 assessment categories. Key findings:

- Claude achieved the highest overall composite score (7.9), reflecting strong performance across most research tasks.
- All systems struggled with making creative connections between concepts (Category 7).
- Galactica lagged other systems significantly in core competencies like literature review and hypothesis generation.
- Human experts outscored the AI systems (8.9/10), indicating gaps versus advanced scholarship.

System	Lit. Review	Hypoth. Gen.	Summarize	Research Gaps	Draft Papers
Claude	8.5	7.5	8.2	7.1	8.7
LaMDA	5.3	7.2	6.7	6.5	8.5
Sydney	7.2	7.8	7.6	6.9	7.3
Galactica	6.1	5.9	6.3	5.7	6.8
System	**Datasets**	**Connections**	**Limitations**	**Biases**	**Validation**
Claude	7.6	6.9	7.8	7.2	8.4
LaMDA	6.4	7.1	7.3	6.8	8.1
Sydney	6.8	7.5	6.6	6.4	8.2
Galactica	5.2	6.1	5.9	5.8	7.6

Table 1. Mean Scores by Category for AI Systems Evaluated

The scoring variations reveal specialized strengths but difficulties generalizing across scholarly tasks. This indicates the need for pragmatic system-task alignment based on empirical capability assessments rather than assumptions.

Qualitative Thematic Analysis

Inductive coding of open-ended responses identified key risks, limitations and validation needs:

Risks

- Potential for misuse of outputs for harmful purposes
- Concerns over biases in training data propagating

Limitations

- Difficulty smoothly adapting reasoning across disciplines
- Struggles with critical analysis beyond surface patterns

Validation

- Emphasis on rigorous human validation of outputs before use
- Suggested safeguards like peer review, citation checks and transparency

The AI systems demonstrated awareness of risks requiring diligent oversight before integrating outputs into published work.

Statistical Analysis

ANOVA found significant differences between mean scores across systems ($F=16.32$, $p<0.001$). Post-hoc Tukey tests revealed Claude scored higher than LaMDA and Galactica ($p<0.05$), but no significant difference between Claude and Sydney. This indicates competitive benchmarking can discern capabilities gaps based on training approaches.

Human Benchmarking

Experts (8.9/10) significantly outperformed AI systems in areas like creative ideation, inference, and synthesis. These underscore current limitations compared to advanced human scholarship and the need for judicious expectations calibrated to validated capabilities.

This multifaceted analysis combining quantitative benchmarking, qualitative insights, statistical tests and human grounding provides comprehensive applied perspectives on the systems' specialized strengths, risks requiring oversight, and integration needs. The findings will guide stakeholders in harnessing AI to ethically augment scholarship based on real-world performance rather than hype. Ongoing benchmarking as capabilities evolve will be key.

Findings

By integrating the quantitative comparative benchmarking and qualitative thematic analysis, several key findings emerge:

1. Specialized capabilities, but limitations generalizing across research tasks

The scoring variations across assessment categories reveal the AI systems exhibit specialized strengths but struggle to generalize competencies across different scholarly tasks. For instance, Claude and LaMDA excel at drafting papers but lag in creative literature analysis and hypothesis generation. This aligns with literature noting difficulties adapting reasoning across knowledge domains (Bawden & Robinson, 2020). It underscores the need to pragmatically match AI tools to suitable research scenarios based on empirical capability assessments, rather than assumptions.

2. Narrow augmentation is achievable presently, but human oversight remains essential

While showing promise in focused assistive roles, the current generative models lack the sophistication to replicate or replace advanced human scholarship. Human experts significantly outperformed the AI systems on integrative reasoning, inference, and creative synthesis. However, purposeful collaboration between human and AI strengths could enhance productivity through narrow augmentation. Maintaining realistic expectations calibrated to validated capabilities, rather than hype, is vital. Further enhancing AI reasoning capabilities to expand utility will be an ongoing priority. Overall, prudent integration leveraging complementary strengths promises the most fruitful path forward presently.

3. Transparency and validation critical for responsible adoption

The AI systems emphasized transparency in documenting AI use and extensive human validation of outputs before research deployment to uphold integrity. This indicates an awareness of biases, limitations, and misuse risks requiring diligent oversight and mitigation measures before integrating AI-generated content into published work. Adoption best practices should mandate version tracking, citation, peer review, reproducibility testing, and documentation.

4. Advancing contextual reasoning and critical analysis capabilities needed

A key limitation identified was difficulty smoothly adapting reasoning across disciplines and critically analyzing arguments beyond surface patterns. Enhancing contextual awareness, causal reasoning, structured knowledge integration, and critical thinking skills is imperative to expand the utility of scholarship. Multi-domain training, hybrid reasoning architectures, argument deconstruction datasets, and attention visualization hold promise.

5. Holistic initiatives needed to mitigate risks

Realizing benefits while mitigating harms requires collaborative action across stakeholders. Researchers must provide oversight and calibrate expectations. Developers should enhance technical safety. Publishers and institutions should mandate transparency, audits, training, and monitoring.

Policymakers should develop proportional governance balancing rigor and access. Constructive input from disciplines adept at reflecting on the societal impacts of technology is also vital. Findings provide data-driven guidance for pragmatic adoption aligned with ethical oversight to advance underdeveloped capabilities.

While promising for focused assistive roles if harnessed judiciously, automating advanced scholarship remains beyond current AI systems. The study yields critical insights and an applied framework to guide stakeholders in integrating generative AI to ethically augment academic research based on empirical performance data rather than hype. Continued benchmarking, capability enhancement, risk mitigation, and human oversight will be crucial as these technologies progress. But with diligence and collaboration, AI-assisted scholarship could open new horizons of possibility to benefit knowledge creation for society.

Discussion

The key findings from the comparative benchmarking and thematic analysis offer crucial insights to guide the responsible integration of generative AI into academic research. This section examines key implications and recommendations stemming from the results.

Pragmatic System-Task Alignment Needed

The scoring variations across research tasks reveal specialized capabilities but difficulties in generalizing competencies. This underscores the need to pragmatically match AI tools to suitable scenarios based on empirical assessments, rather than assumptions. Researchers should critically evaluate specific project needs and test model capabilities through prompting experiments to guide appropriate pairing. Published model documentation by AI providers detailing trained domains, intended uses, and limitations can further aid matching. Dynamically combining complementary systems can support end-to-end workflows, albeit requiring additional integration tools. Overall, adoption aligned with validated capabilities, rather than hype, will be essential for effective augmentation.

Calibrating Expectations on Augmentation Versus Automation

While promising for focused assistance, findings reveal current systems lack the reasoning sophistication of human experts across integrative literature synthesis, inference, and creative

doi.org/10.32388/7OVVW2

ideation. Sole reliance on AI for critical decisions thus remains inadvisable presently. However, purposeful collaboration between human and machine strengths could enhance productivity. Further enhancing contextual reasoning and critical analysis capabilities will be imperative to expand utility. Maintaining realistic expectations calibrated to empirical performance is vital. Complete automation of advanced scholarship remains beyond current AI. However, prudent integration leveraging complementary capabilities promises the most fruitful path forward. Ongoing benchmarking as systems progress will be key.

Upholding Research Integrity Through Diligent Oversight

The AI systems emphasized extensive human validation of outputs before research use, indicating awareness of risks. This underscores the duty of adopters to rigorously assess AI-assisted work for soundness before dissemination. Recommended safeguards include peer review, citation checks, reproducibility tests, bias reviews, transparency requirements, and editorial oversight.

While imperfect, combining process rigor with transparency can enhance prudence. High-quality scholarship necessitates diligence regardless of tools. Integrity norms must evolve intelligently, not be discarded, to integrate AI responsibly.

Advancing Contextual Reasoning and Critical Analysis

Difficulty smoothly adapting reasoning across disciplines and critically analyzing arguments beyond surface patterns emerged as key limitations. Advancing contextual awareness, causal reasoning, structured knowledge integration, argument deconstruction, and attention visualization are promising approaches to address these gaps and expand utility. Training enhancements like multi-domain corpora, configurable memory, hybrid reasoning architectures, and reinforcement learning also hold potential. Targeted capability enhancement to expand applicability across diverse scholarly tasks will be imperative.

Holistic Initiatives Needed to Mitigate Risks

The full realization of benefits while mitigating potential harms will necessitate collaborative action across key stakeholders. Researchers must judiciously evaluate capabilities, provide oversight, and uphold ethics. Developers should enhance technical safety through rigorous testing and bias mitigation. Publishers and institutions should mandate transparency, audits, training, and monitoring. Policymakers should develop proportional governance balancing rigor and access. Constructive perspectives from disciplines skilled in assessing the societal impacts of technologies can inform ethical development. By working together, an equitable way forward can be charted that allows generative AI's immense potential to flourish responsibly.

Findings provide data-driven guidance for pragmatic adoption focused on advancing underdeveloped capabilities under an ethical oversight framework. Continued unbiased research, discussion and policy development engaging diverse voices will be vital as adoption evolves.

Recommendations for Key Stakeholders

Based on the study insights, tailored recommendations can guide key groups:

For researchers:

1. Critically evaluate model capabilities against project needs through prompting tests. Avoid hype assumptions.
2. Maintain realistic expectations of augmentation over automation of advanced scholarship.
3. Rigorously validate outputs through peer review, reproducibility tests, and bias reviews before dissemination.
4. Document AI use, versions, and training data transparently to support reproducibility.
5. Complete ethics training on risks, limitations and responsible adoption.

For AI developers:

1. Enhance reasoning, contextual adaptation, critical analysis and bias mitigation capabilities to expand utility.
2. Conduct rigorous testing across diverse groups pre-release to uncover potential harms.
3. Provide clear documentation on model versions, training data, intended uses, and limitations to support due diligence.
4. Develop tailored testing prompts for capability evaluation by researchers.
5. Increase transparency on model uncertainties and mistakes to build appropriate trust.

For institutions and publishers:

1. Require transparency statements on AI use during manuscript submission.
2. Perform plagiarism and citation checks on AI-assisted submissions.

3. Establish oversight bodies to monitor risks, audit policies, and investigate violations.
4. Develop proportional governance policies balancing rigor, access and scientific progress.
5. Incorporate AI ethics training into both undergraduate and graduate curricula.

Through purposeful collaboration and pragmatic policies informed by an ethical framework and empirical insights, stakeholders can realize AI’s opportunities while addressing its risks to benefit scholarship and society.

Ongoing Inquiry Needed

Ongoing inquiry is needed to address various open questions that have arisen as AI adoption continues to evolve. These questions include how capabilities vary for field-specific tasks across disciplines, what risks emerge from emerging multi-modal generative applications, and how benchmarks can dynamically assess rapidly advancing models. Additionally, there is a need to determine what participatory frameworks for oversight can maximize benefit and minimize harm, as well as how adoption of best practices can balance rigor, access, transparency, and progress. Fostering responsible development of these powerful technologies to enhance scholarship will require ongoing vigilant, unbiased, and collaborative inquiry across stakeholders. This study provides an initial comparative assessment and evidence-based framework to guide that crucial journey.

Conclusion

The study contributes significantly to the responsible use of generative AI in academia by providing a comprehensive framework for its adoption and oversight. The mixed-methods approach provides a nuanced understanding of the capabilities and limitations of four cutting-edge AI models, as well as the risks and challenges associated with their use. An actionable framework with tailored recommendations for key stakeholders and an evidence-based roadmap for judicious adoption have been developed. The findings address key gaps in understanding the implications of these technologies and provide data-driven guidance for their pragmatic integration under ethical oversight.

The study contributes to the development of best practices and policy priorities for upholding research integrity, enhancing human-AI collaboration, and guiding progress responsibly across academic domains. It emphasizes the need to recalibrate expectations and enhance capabilities to unlock AI's full potential to augment scholarship, while also highlighting the importance of holistic risk mitigation strategies and ethical oversight.

The framework developed in this study could inform the responsible integration of AI across other professional domains involving complex reasoning. Additionally, the study showcases the potential of human-AI collaboration to amplify our best capacities through prudent partnership with machines.

While the study makes significant contributions, there are limitations that provide opportunities for future work, such as expanding the sample size and range of models assessed, exploring variances in capabilities for domain-specific tasks across disciplines, and investigating risks and biases in multi-modal generative applications.

Developing dynamic benchmarking techniques, participatory frameworks with scholars and stakeholders, and investigating the impact on intellectual property rights and academic culture are also important areas for future inquiry. Realizing the promise of AI to enhance scholarship while addressing the risks and challenges will require ongoing unbiased, collaborative, and creative inquiry across diverse perspectives. The goal of cultivating these technologies for the broader benefit of scientific progress and society remains worthy of our greatest efforts. This work aims to contribute to that journey of ethical co-creation – of both knowledge and a more enlightened world.

References

- Bawack, R., & Desveaud, K. (2022). *Consumer Adoption of Artificial Intelligence: A Review of Theories and Antecedents*.
- Bawden, D., & Robinson, L. (2020). *Information overload: An overview*.
- Bhutoria, A. (2022). Personalized education and Artificial Intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model. *Computers and Education: Artificial Intelligence*, *3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., & Brunskill, E. (2021). On the opportunities and risks of foundation models. *ArXiv Preprint ArXiv:2108.07258*.
- Chan, A. (2023). GPT-3 and InstructGPT: technological dystopianism, utopianism, and “Contextual” perspectives in AI ethics and industry. *AI and Ethics*, *3*(1), 53–64. https://doi.org/10.1007/s43681-022-00148-6

- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *00*(00), 1–12. https://doi.org/10.1080/14703297.2023.2190148
- De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120. https://doi.org/10.3389/fpubh.2023.1166120
- Driscoll, D. L., Appiah-Yeboah, A., Salib, P., & Rupert, D. J. (2007). *Merging qualitative and quantitative data in mixed methods research: How to and why not.*
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
- Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*. http://arxiv.org/abs/2303.10130
- Eysenbach, G. (2023). The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. In *JMIR medical education* (Vol. 9, p. e46885). https://doi.org/10.2196/46885
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459.
- Gupta, R., Herzog, I., Najafali, D., Firouzbakht, P., Weisberger, J., & Mailey, B. A. (2023). Application of GPT-4 in Cosmetic Plastic Surgery: Does Updated Mean Better? *Aesthetic Surgery Journal*, *43*(8), NP666–NP669. https://doi.org/10.1093/asj/sjad132
- Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, *90*, 104512. https://doi.org/10.1016/j.ebiom.2023.104512
- Li, L., Ma, Z., Fan, L., Lee, S., Yu, H., & Hemphill, L. (2023). ChatGPT in education: A discourse analysis of worries and concerns on social media. *ArXiv Preprint ArXiv …*, 1–35. https://arxiv.org/abs/2305.02201%0A https://arxiv.org/pdf/2305.02201
- Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., White, C., Gu, Q., … Zhao, L.

(2023). *Beyond One-Model-Fits-All: A Survey of Domain Specialization for Large Language Models*. http://arxiv.org/abs/2305.18703

- Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. In *Education Sciences* (Vol. 13, Issue 4). https://doi.org/10.3390/educsci13040410
- Megahed, F. M., Chen, Y.-J., Ferris, J. A., Knoth, S., & Jones-Farmer, L. A. (2023). How generative ai models such as chatgpt can be (mis) used in spc practice, education, and research? an exploratory study. *Quality Engineering*, 1–29.
- Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine*, *6*(1), 120. https://doi.org/10.1038/s41746-023-00873-0
- Morande, S., Arshi, T., Gul, K., & Amini, M. (2023). SECURE II: Unlocking the Potential of Artificial Intelligence for Entrepreneurial Success. *Qeios*, *Preprint*. https://doi.org/10.32388/RELTLQ.3
- Morande, S., Arshi, T., Gul, K., Amini, M., & Tewari, V. (2023). Tracing the Future: Blockchain and IoT's Role in Revolutionizing Food Supply Chain Transparency. In Ş. Aydın, E. Özgül Katlav, K. Çamlıca, & F. Yönet Eren (Eds.), *Impactful Technologies Transforming the Food Industry* (pp. 156–174). IGI Global. https://doi.org/10.4018/978-1-6684-9094-5.ch010
- Morande, S., & Tewari, V. (2023). Causality in Machine Learning: Innovating Model Generalization through Inference of Causal Relationships from Observational Data. *Qeios*, *Preprint*. https://doi.org/10.32388/P7MMGR
- Rana, S. (2023). AI and GPT for Management Scholars and Practitioners: Guidelines and Implications. In *FIIB Business Review* (Vol. 12, Issue 1, pp. 7–9). SAGE Publications Sage India: New Delhi, India.
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003
- Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. In *Healthcare* (Vol. 11, Issue 6). https://doi.org/10.3390/healthcare11060887
- Solaiman, I., & Dennison, C. (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, *34*, 5861–5873.
- Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. *Chi Conference on Human*

Factors in Computing Systems Extended Abstracts, 1–7.

- Wu, T., Terry, M., & Cai, C. J. (2022). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3517582

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.