“LLM: Usage of ChatGPT and AI Tools. An academic analysis from an auditor's perspective.”

Pooja Singh,

University at Buffalo, Buffalo, NY, USA

psingh59@buffalo.edu

Introduction

Over the last few years, complex language models (LLMs) such as ChatGPT, Gemini, and Claude have developed in favor among both students and professionals. These tools can write code in response to prompts, explain concepts, and provide solutions to questions. Because they are easy to use and are usually available for free or for a fee, many individuals have quickly incorporated them into their work and study habits. The swift growth of LLMs brings both advantages and disadvantages. On the side, they help individuals quickly grasp difficult concepts and create original ideas. On the downside, they sometimes produce information and might encourage dependence or academic dishonesty when misused. As a result, teachers, managers, and lawmakers are actively debating ways to manage this technology. As a result Auditors are among the professionals most exposed to the rapidly developing capabilities of artificial intelligence (AI) systems (Eloundou, 2023). It emphasizes the fundamental benefits that consumers highlight, as well as major downsides, including prejudice, errors, and privacy concerns. Large language models are a category of systems designed to manage human’s repetitive work enhancement and fulfillment. LLMs have shown their potential to recreate knowledge and perform simple tasks, making them available to a diverse community of users through interfaces that require minimal coding . (Bommasani)

Audit firms have invested billions of dollars in developing their internal LLMs, due to the high costs associated with developing internal LLMs, small to medium companies and firms are likely to face challenges in building their own (Deloitte, 2024; EY, 2023; KPMG, 2023), forcing them to rely on external LLMs if they intend to leverage this technology, Hence the ChatGPT, Gemini, and other MSPs come into the picture for usability, governance, Technical controls, and Monitoring.

The use of LLM and why LLMs are everywhere

When students are confused where to start, they regularly use LLMs to enhance their language, and work. Irrespective of the nature of the work, it could be fine arts or it source is computer science where code writing is required. The end result is AI written text is sharper and simpler to read, furthermore, it injects the good sources in to the content to reflect how the ideas or frameworks for essays and projects are being used as best practices in the specific domain. Ultimately, helping the students & professionals with their assignments in the respective ecosystems.

Nevertheless, there are challenges associated with employing LLM in settings. Certain professionals might depend on the model to generate work and subsequently present the material as their own with polished expert advice. This complicates the task of differentiating between cheating and seeking help. Furthermore, the usage of AI is believed to be making humans less intelligent as it has started doing basic work completely automated and taking away the opportunity to brainstorm and thinking solutions originally own their own. For any job, understanding the basics and fundamentals knowledge are key elements of foundation building which is being weakened by AI due to widespread usage. LLMs are used in the workplace to save time and lessen routine effort, particularly in positions that require a lot of reading and writing. Knowledge workers use them to prepare emails, reports, and presentations, and to summarize extensive documents or meeting notes. In technical industries such as software development, LLMs can create example code, explain error messages, or propose solutions to fix bugs. But humans still need to carefully review any writing or code produced by AI, since the governance of AI is still in question.

AI systems frequently need to be responsive to a variety of parameters that are valuable to interested parties in order to be trusted. Negative AI risks can be decreased by methods that improve AI trustworthiness. The following qualities of reliable AI are outlined in this framework, along with recommendations for resolving them. The qualities of reliable AI includes valid and dependable, safe, secure, and robust, transparent and responsible, explainable and interpretable, privacy-enhanced, and equitable with harmful bias handled are some examples of systems. Each of these qualities must be balanced according to the AI system's intended application in order to produce trustworthy AI. Accountability and transparency are related to both the internal

and exterior processes and activities of an AI system, even if these qualities are socio-technical system traits (NIST AI, 2023).

Merits and Limitations of LLMs

Speed and productivity can be greatly increased with LLMs. With outlines, concept lists, or editable first drafts, tasks that used to take hours, like writing reports, can frequently be started in a matter of minutes. Additionally, they help non-native speakers by enhancing grammar, tone, and clarity, and they promote learning by offering explanations and examples in subjects like statistics, programming, and mathematics. These advantages do, however, come with significant hazards. If results are not independently validated, LLMs may produce confident but inaccurate content, which could result in bad conclusions. computers may also replicate bias or stereotypes since computers learn from human-produced material, raising issues with justice and reputation. Frequent dependence on LLMs can also erode autonomous thought and cause problems with integrity in both professional and academic contexts. Lastly, when users enter sensitive personal or organizational data into external AI technologies without fully understanding how that data is kept, processed, or retained, privacy and confidentiality problems emerge (Lazarus Elad Fotoh, 2025).

For auditors, accountability for data produced by LLM is a crucial ethical concern. Accountability is the foundation of audit transparency and mandates that auditors behave in the public interest, abide by the law, and refrain from actions that harm the profession (IESBA, 2024). Auditors are still in charge of the decisions made even whether LLMs aid in analysis or decision-making. This is a danger because LLM outputs could be erroneous, hard to link to reliable sources, and could lower audit quality when applied in highly autonomous ways. Organizations could improve transparency by revealing LLM capabilities, limitations, and error-handling expectations in order to address these issues. They should also clearly establish accountability for AI-assisted outputs.

Key elements an auditor should evaluate for LLMs usage

- Audits focused solely on compliance are inadequate: Because important hazards (hallucinations, bias, data leakage, rapid injection, and emergent behaviors) frequently fall outside of standard compliance testing, checking boxes against policies or regulations won't give LLMs any real assurance. (NIST: AI Risk Management, 2022) (ISO.: ISO/IEC 23894, 2023)

- Independent outside confirmation is required: In addition to establishing vendor accountability in the event of mishaps, irregularities, or harm, third-party audits assist in verifying that LLM installations are technically sound, ethically sound, and legally compliant. (Floridi, 2022) (Engler, 2023)
- Co-executing audits with suppliers is required: To ensure practicality and credibility, effective LLM audits require both independent auditor judgment and collaboration from technology vendors (access to model documentation, logs, evaluation findings, and system design). (Engler, 2023)
- An method that combines technical auditing and governance is crucial: Both technological controls (security, data pipelines, model behavior testing, monitoring) and governance controls (rules, roles, training, supervision, incident response) must be assessed by LLM assurance. (Raji, 2022)
- Redesigning technology audit techniques for LLM hazards is necessary. To evaluate LLM-specific risks, such as prompt-injection channels, tool/agent overreach, model drift, and opaque decision trails, conventional IT audit techniques must be modified. (Steed, 2022)
- Safer redesign and risk reduction are made possible by model audits: In order to minimize downstream damage, targeted model assessments can identify constraints and failure modes, make them evident to stakeholders, and encourage system redesign (guardrails, human review, constrained outputs). (OECD, 2022)
- Monitoring after deployment must be ongoing: To satisfy regulatory and assurance goals, audit frameworks must incorporate continuous, ex-post monitoring since LLM behavior and risk exposure might vary over time due to updates, new data, new prompts, and new integrations. (Ganguli, 2023) (Perez, 2022)

Understanding Large Language Models (LLMs)

Large language models (LLMs) are machine-learning algorithms that have been trained on vast text corpora, such as books, websites, and articles, in order to identify statistical patterns in language and provide replies by forecasting possible word sequences. They are usually accessed through conversational interfaces and integrated into programs like writing assistance, coding assistants, and customer service chatbots. They are implemented in technologies like ChatGPT, Gemini, and Claude. Although LLMs create outputs that are fluent and human-like, they do not "understand" in a human sense;

they might replicate biases seen in training data and generate claims that are plausible but erroneous or poorly grounded, necessitating human verification. Gaps in uniform regulatory control increase ethical and assurance problems as these models proliferate and provide wide-ranging benefits across numerous use cases (Chenhui Shen, 2023) . This emphasizes the necessity of audit methods that address both known unknowns—identifiable risks like bias and privacy exposure—and unknown unknowns—unexpected failure modes that can only be discovered by methodical investigation, stress testing, and performance assessment in practical settings.

Conclusion

In conclusion, because they speed up drafting, analysis, and communication, LLMs like ChatGPT, Gemini, and Claude are now integrated into education and knowledge work, including audit and assurance tasks. But from the standpoint of an IT audit, their worth is inextricably linked to a number of unique and significant hazards that conventional compliance-only inspections fall short of capturing. Sensitive prompts, prompt/output recording, vendor retention procedures, and the potential for memorization or extraction of proprietary data are all potential sources of confidentiality exposures that auditors should assess. Additionally, when LLM outputs have the potential to cause downstream actions, they must handle "insecure output handling" as a vital control point and manage integrity issues such as hallucinations, inaccurate citations, and dangerous code recommendations. Security concerns in LLM-enabled systems increase due to tool misuse, direct and indirect prompt injection, and instruction/data misunderstanding in RAG and agentic processes. Operational risks to availability and cost, such as abuse-driven usage spikes, denial-of-service patterns, and runaway token consumption, can cause budgetary and service delivery disruptions. Lastly, because the majority of businesses rely on outside suppliers, supply-chain risk is introduced by model updates, plugins, hosting agreements, and training data dependencies, necessitating more stringent vendor governance and assurance.

Practically speaking, LLMs can be used safely if controls are built for how people actually use them: enforce data classification and acceptable-use rules; establish robust identity and access management; secure the LLM SDLC with threat modeling and adversarial testing; require human review for high-impact outputs; maintain robust logging and incident response; manage vendors through explicit contractual and assurance requirements; and continuously monitor performance and risk post-deployment rather than treating the audit as a one-time event.

References

Bommasani, R. (. (n.d.). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Chenhui Shen, L. C.-P. (2023). Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. *https://arxiv.org/abs/2305.13091*.

Deloitte. (2024). How we harness Generative AI for a smarter, digital audit". Available at: https://www2.deloitte.com/content/dam/Deloitte/se/Documents/audit/AI%20presentation%20for%20the%20Audit%20Committee%20-%20Second%20draft.pdf (accessed 5 April 2024).

Eloundou, T. M. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130, 10.

Engler, A. (2023). Outside auditors are struggling to hold AI companies accountable. FastCompany. https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue (2021).

EY. (2023). https://www.ey.com/en_gl/news/2023/08/ey-releases-more-than-20-new-assurance-technology-capabilities-supported-by-microsoft-alliance-in-first-year-of-us1b-investment-program (accessed 3 April 2024). *EY releases more than 20 new Assurance technology capabilities supported by Microsoft alliance in first year of US$1b investment program*.

Floridi, L. H. (2022). CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. Available at SSRN 4064091.

Ganguli, D. e. (2023). Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. ArXiv, 2022. [Online]. https://github.com/anthropics/hh-rlhf.

IESBA. (2024). Handbook of the International Code of Ethics for Professional Accountants. New York: International Federation of Accountants, Professional Code.

ISO.: ISO/IEC 23894, S. -I.—A.—G. (2023). *https://www.iso.org/standard/77304.html, (2023). Accessed 20 Jan 2023*.

KPMG. (2023). KPMG LLP and Microsoft Establish Industry-Leading Initiative to Scale Generative AI Across Audit, Tax and Advisory. . *Available at: https://info.kpmg.us/news-perspectives/technology-innovation/kpmg-microsoft-genai-2023.html (accessed 5 April 2023)*.

Lazarus Elad Fotoh, T. M. (2025). Exploring Large Language Models in external audits: Implications and ethical considerations. *https://doi.org/10.1016/j.accinf.2025.100748*.

NIST AI, 1.-1. (2023). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf?utm_source=chatgpt.com.

NIST: AI Risk Management, F. (2022). Second Draft Notes for Reviewers: Call for comments and contributions. National Institute of Standards and Technology.

OECD. (2022). OECD Framework for the Classification of AI systems. Paris. [Online]. https://doi.org/10.1787/cb6d9eca-en.

Perez, E. (2022). Red teaming language models with language models. ArXiv (2022). https://doi.org/10.48550/arxiv.2202.03286.

Raji, I. X. (2022). Outsider oversight: designing a third party audit ecosystem for AI governance. In: AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 557–571, Jul. 2022, https://doi.org/10.1.

Steed, R. P. (2022). Upstream mitigation is not all you need: testing the bias transfer hypothesis in pre-trained language models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 3.