Proceedings of CONF-SEML 2026 Symposium: Multimodal Data Acquisition: Applications in Physiological and Behavioral Research
DOI: 10.54254/2755-2721/2026.AD32321

Challenges and Future Development of Generative AI in Data Analysis

Tianyi Yu

School of Earth Sciences and Engineering, Nanjing University, Nanjing, China
191830195@smail.nju.edu.cn

Abstract. The rapid advancement of generative AI technologies has transformed numerous industries, with data analysis being one of the most significantly affected fields. Large language models and their generative counterparts have opened doors that seemed firmly shut just a few years ago, offering unprecedented avenues for intelligent transformation in how we handle data. This paper takes stock of where we currently stand: mapping existing applications across diverse sectors, confronting the messy realities of data quality and model opacity, and wrestling with privacy and ethical dilemmas that have no simple solutions. Our findings suggest that while the technology promises substantial gains in efficiency (some organizations report productivity jumps of 40% in data preparation alone), realizing sustainable value requires more than technical prowess. It demands careful attention to governance, interpretability, and responsible deployment. It is expected that this work offers practitioners and researchers alike a grounded perspective on navigating this rapidly shifting landscape.

Keywords: Generative AI, Data Analysis, Large Language Models, Artificial Intelligence, Future Development

1. Introduction

Something fundamental has changed in how organizations approach data-driven decision-making—and generative AI sits at the heart of this shift. Brown et al. [1] demonstrated back in 2020 that large language models could understand and generate human-like text with uncanny proficiency, capabilities that immediately addressed with data interpretation challenges. The transformer architecture introduced by Vaswani et al. [2] provided the technical scaffold upon which today's generative models stand, while Bommasani et al. [3] have more recently offered a sober accounting of both the promise and perils inherent in these foundation models.

This paper addresses three core research areas. First, what does the current landscape of generative AI in data analysis actually look like on the ground? Second, where do these technologies stumble—what barriers persist despite the hype? And third, where we might sensibly steer this ship in the coming years? Through a comprehensive literature review and case study analysis, we aim to present a nuanced perspective on this field, which we hope will bridge the profound gap between the theoretical capabilities of AI and the practical realities of its application.

DOI: 10.54254/2755-2721/2026.AD32321

2. Application status and challenge analysis

2.1. Current application status

Generative AI has carved out surprisingly diverse niches across the data analysis workflow, often in ways that fundamentally alter how organizations extract insights from their digital assets. In natural language processing for data interpretation, models like GPT-4 and its successors [4] have proven remarkably adept at bridging the gap between raw data and human understanding. They automatically generate descriptive statistics, identify patterns that might elude rushed analysts, and craft narrative summaries that render complex findings accessible to non-technical stakeholders. Perhaps more practically, its code generation capabilities now handle much of the routine drudgery—data cleaning, transformation, formatting—including that previously consumed analyst hours. Organizations deploying these tools report productivity improvements reaching 40% in data preparation tasks, freeing human analysts for more sophisticated, judgment-intensive work.

The predictive analytics domain presents a different but equally compelling use case. Synthetic data generation—using models to create training datasets that mirror real data statistically while preserving privacy—has solved what once seemed an intractable tension between data scarcity and confidentiality [5]. Financial institutions, for instance, increasingly rely on GANs to simulate transaction histories that maintain statistical fidelity without exposing actual customer records. Meanwhile, anomaly detection applications have gained traction in manufacturing, where models learn "normal" operational patterns and flag deviations predictive of equipment failure. Early adopters in this space have seen an average reduction in downtime of 25%.

Visualization and reporting represent another frontier. Instead of manually selecting chart types and drafting explanatory text, analysts now leverage generative AI to automatically match visualizations to data characteristics and generate comprehensive reports that synthesize multiple analytical angles [6]. The time compression is dramatic: what once required several hours now often takes minutes, enabling decision-making rhythms that approach real-time—a capability that seemed implausible just years ago.

Healthcare offers perhaps the most striking examples. Researchers apply these technologies to parse vast clinical datasets, identifying subtle patterns invisible to conventional statistical methods. Pharmaceutical firms use generative models to predict molecular properties and screen drug candidates, significantly accelerating discovery timelines. These applications underscore a crucial point: generative AI excels precisely in contexts where datasets grow too complex and multidimensional for human cognition alone.

Customer analytics has embraced the technology with equal enthusiasm. E-commerce platforms analyze browsing trajectories, purchase histories, and demographic signals to generate personalized recommendations and marketing copy. The capacity to process unstructured data—social media sentiment, product reviews, chat logs—provides insights into customer preferences that traditional structured-data approaches miss. Retailers implementing these systems report conversion rate improvements ranging from 15% to 30%.

Supply chain management constitutes an emerging application domain worth watching. By integrating historical demand patterns with weather data and market indicators, generative AI can optimize inventory levels and anticipates supply disruptions. Logistics firms have optimized routing to reduce fuel consumption, with some realizing cost savings approaching 20%. The ability to synthesize disparate data streams creates supply chains that are simultaneously more responsive and resilient.

DOI: 10.54254/2755-2721/2026.AD32321

2.2. Key challenges

Despite these promising applications, several significant challenges persist that must be addressed for widespread adoption. Data quality remains stubbornly problematic. Generative AI models are, in essence, mirrors of their training data; feed them biased or incomplete datasets, and they will reflect those distortions back with alarming confidence [7]. Organizations must build robust governance frameworks—not merely as compliance checkboxes, but as essential infrastructure ensuring representative, high-quality inputs. The stakes are particularly high in sectors where historical data encodes past inequities, raising the specter of AI systems that perpetuate or amplify existing biases under the pretense of algorithmic neutrality

Interpretability presents equally thorny difficulties. The "black box" nature of many generative models makes it genuinely difficult—even for skilled analysts—to trace how specific conclusions emerge [8]. This opacity erodes trust and complicates regulatory compliance, especially in sectors such as financial services and healthcare, where explainability is not just preferred but also legally mandated. Developing interpretable models and effective explanation mechanisms remains an active, urgent research frontier.

Ethical and privacy concerns demand sustained attention. The deployment of generative AI in data analysis inevitably surfaces uncomfortable questions about consent, surveillance, and algorithmic fairness [9]. Navigating the GDPR and similar regulatory frameworks while maintaining analytical utility requires careful architectural decisions—trade-offs that organizations cannot afford to treat as afterthoughts.

Technical limitations, too, constrain the field's expansion. Training and inference demand substantial computational resources, creating barriers for smaller organizations and raising legitimate sustainability concerns about the carbon footprint of large-scale model training. Then there is the "hallucination" problem: these models can generate outputs that sound authoritative yet contain factual errors. In high-stakes decision-making contexts, such errors pose serious risks, necessitating ongoing research into more efficient architectures and rigorous validation mechanisms.

Human-AI collaboration remains under-theorized in practice. While generative AI clearly enhances human analytical capabilities, integrating these tools into existing workflows requires—more often than not—profound organizational change. Analysts need new skills: not just prompt engineering, but also the critical evaluation of AI-generated outputs, the ability to spot subtle errors masquerading as coherence. Successful implementations have consistently invested in training and change management, designing workflows that leverage distinctively human judgment alongside computational power.

Finally, integration with legacy infrastructure presents practical headaches. Many existing data systems were not designed for AI consumption; modernizing them requires significant capital investment. Data silos within organizations—departmental hoarding of information—further limit the effectiveness of generative AI, which thrives on comprehensive, cross-functional data access. Breaking down these barriers demands both technical solutions and cultural shifts toward collaborative data sharing.

3. Future development directions and recommendations

Looking ahead, several trajectories appear particularly consequential for the field's evolution. Multimodal integration likely represents the next significant leap. Future systems will process text, imagery, audio, and structured data simultaneously, enabling analyses that capture the full complexity of customer interactions and operational realities [10]. Organizations will stitch together

DOI: 10.54254/2755-2721/2026.AD32321

voice recordings, chat transcripts, and behavioral telemetry into coherent, holistic pictures of user experience—analytical syntheses previously impossible.

Domain-specific models will increasingly displace one-size-fits-all approaches. Rather than deploying general-purpose LLMs across every use case, organizations will fine-tune specialized models trained on industry-specific corpora. This specialization promises not only improved accuracy but also reduced computational requirements. We anticipate bespoke models for financial risk analysis, clinical diagnostics, supply chain optimization—each embedding tacit domain knowledge that generic models cannot replicate.

Efficiency gains and edge deployment will democratize access. Smaller, distilled models capable of running on standard hardware will lower barriers to entry, allowing smaller organizations to participate. Federated learning architectures, meanwhile, will enable collaborative model improvement without centralizing sensitive data—addressing privacy concerns while preserving the benefits of shared learning.

For organizations positioning themselves in this landscape, we suggest several pragmatic steps. Organizations should invest seriously in data infrastructure and governance—not as overhead, but as a strategic foundation. Cultivate AI literacy among data professionals, emphasizing critical evaluation skills alongside technical proficiency. Foster genuine collaboration between technical teams and domain experts; the best implementations we have observed emerge from such partnerships rather than siloed development process.

Adopt iterative implementation strategies: pilot projects that test assumptions before scaling, allowing organizational learning to keep pace with technological capabilities. Consider establishing centers of excellence that can consolidate expertise and disseminate best practices.

Engage proactively with evolving regulatory landscapes. Standards for transparency, accountability, and fairness are crystallizing; early alignment with emerging best practices will yield compliance advantages. More importantly, investment in ethical AI practices builds the trust capital necessary for sustainable operations.

Education and workforce development will prove decisive. Universities must update curricula to include practical AI literacy; organizations must commit to continuous learning for existing staff. The ability to collaborate effectively with AI systems—not merely use them—will increasingly differentiate successful organizations. Public-private partnerships that accelerate this workforce evolution deserve serious attention.

4. Conclusion

Generative AI in data analysis stands at an inflection point—poised between transformative potential and substantial unresolved challenges. Our analysis confirms that these technologies offer genuine capacities for automating insight generation, enhancing accessibility through natural language interfaces, and streamlining routine analytical workflows. Applications spanning healthcare, finance, retail, and manufacturing demonstrate the technology's versatility across organizational contexts.

Yet potential is not destiny. Realizing sustainable value requires systematic attention to data quality assurance, model interpretability, ethical governance, and regulatory compliance. The organizations that thrive will be those addressing these challenges proactively rather than reactively.

Future development likely lies in multimodal systems, domain-specific models, and more sophisticated human-AI collaboration frameworks. The transition to AI-augmented analysis represents not merely a technological upgrade but a fundamental reorientation in how organizations generate knowledge and make decisions. Success will depend less on the sophistication of the

DOI: 10.54254/2755-2721/2026.AD32321

models themselves than on the thoughtfulness with which human expertise is integrated with machine capabilities.

We acknowledge several limitations in this work. The field's rapid evolution means some observations may date quickly; our reliance on published literature may miss cutting-edge industrial innovations not yet documented in academic sources. Future research would benefit from longitudinal empirical studies that track how these applications evolve over time, as well as comparative analyses across industries that might illuminate which implementation strategies prove most robust.

As the field matures, continued dialogue between researchers, practitioners, and policymakers will be essential. Generative AI in data analysis is not merely a technical phenomenon but a socio-technical one; understanding its long-term impacts on organizational decision-making will require sustained and critical inquiry.

References

[1] Brown, T. (2020). Language models are few-shot learners. Advances in Neural Information. In 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. Retrived from https: //proceedings. neurips. cc/pa.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[3] Bommasani, R. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv: 2108.07258.

[4] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv: 2303.08774.

[5] Yoon, J., Jordon, J., & Schaar, M. (2018, July). Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. In International Conference on Machine Learning (pp. 5699-5707). PMLR.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). A robustly optimized bert pretraining approach. arXiv preprint arXiv: 1907.11692.

[7] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.

[8] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.

[9] Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, 535-545.

[10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.