An IS Research Agenda on Large Language Models: Development, Applications, and Impacts on Business and Management

MICHAEL CHAU, Faculty of Business and Economics, The University of Hong Kong, Hong Kong
JENNIFER XU, Computer Information Systems, Bentley University, Waltham, United States

Large language models have been advancing very rapidly and are making substantial impacts on all areas of business and management. We review the development of large language models and their applications in business and management, and identify the major issues and challenges faced by both practitioners and researchers. Based on our review, we propose an agenda for information systems researchers on large language models and discuss some of the potential directions for future research. Lastly, we present the articles in the special issue as exemplary research on large language models and discuss their implications.

CCS Concepts: • **Information systems → Information systems applications; • Computing methodologies → Natural language processing**;

Additional Key Words and Phrases: Large language models, artificial intelligence, information systems research, business and management

ACM Reference Format:
Michael Chau and Jennifer Xu. 2025. An IS Research Agenda on Large Language Models: Development, Applications, and Impacts on Business and Management. *ACM Trans. Manag. Inform. Syst.* 16, 1, Article 1 (February 2025), 11 pages. https://doi.org/10.1145/3713032

1 Introduction

We all are witnesses of the history. When the **Call for Papers** (**CFP**) for this special issue was announced at the beginning of 2023, OpenAI's ChatGPT had taken the world by storm since its launching in November 2022, gaining over 100 million users in merely two months. As a **large language model** (**LLM**), ChatGPT and its peers back then (e.g., Google's Bard, BigScience's BLOOM, and Baidu's ErnieBot) had attracted massive attention due to their unprecedented capabilities to understand natural languages and generate human-like responses for requests that go beyond traditional **natural language processing** (**NLP**) tasks.

The objective of this special issue is to curate a set of high-quality management **information systems** (**IS**) articles that focus on the design and application of LLMs in business and management and their impacts, as well as the ethical and social issues involved. As LLMs and generative AI are impacting nearly every aspect of our life and society [14], a plethora of research opportunities

Authors' Contact Information: Michael Chau, Faculty of Business and Economics, The University of Hong Kong, Hong Kong; e-mail: mchau@business.hku.hk; Jennifer Xu, Computer Information Systems, Bentley University, Waltham, Massachusetts, United States; e-mail: jxu@bentley.edu.

ACM 2158-656X/2025/02-ART1
https://doi.org/10.1145/3713032

are emerging for scholars in many disciplines including IS. In this editorial, we first provide a brief overview of the history of LLMs and their architectures, design, applications, and use cases. Section 3 summarizes LLMs' possible impacts on business, management, and society, and lays out major challenges in the research and development of LLMs. We then propose a new IS research agenda on LLMs and a set of research questions that may motivate IS researchers to investigate. The last section introduces eight articles in this special issue, showcasing promising applications of LLMs in various domains and valuable contributions made to the IS literature on LLMs.

2 LLMs

The advent of LLMs can be tracked back to 2018 when **Bidirectional Encoder Representations from Transformers** (**BERT**) was proposed [10]. BERT-base and BERT-large are encoder-only transformers and contain 110 million and 340 million parameters, respectively. Using self-supervised learning, BERT was trained on numerous text corpora such as books and Wikipedia entries and demonstrated outstanding performance in NLP tasks, including named entity recognition, document classification and summarization, speech recognition, and machine translation [10]. Since then, the research and development on LLMs have advanced rapidly. While we used BERT as an example of LLMs in our CFP two years ago, BERT is tiny compared with the state-of-the-art LLMs. At the time of this writing, the largest LLMs, particularly those GPTs (Generative Pretrained Transformers), have been on a scale of several orders of magnitude greater than the size of BERT. For example, GPT-4, which was released in early 2024, was estimated to have over 1.75 trillion parameters [46]. Although no public information can be found about the size of OpenAI's GPT-o3, which was recently announced in December 2024, it is reasonable to expect that its size will be no smaller than its predecessors. Other competitors such as Google's Gemini-1.5 and Anthropic's Claude 3.5 are on a similar scale [9]. Enabled by their massive sizes, GPTs have shown stunning capabilities to handle increasingly complex tasks, such as content generation, reasoning and planning, and decision-making [35].

LLMs usually are first pretrained using many text corpora and datasets, embedding into their large number of weights a comprehensive amount of language and world knowledge and a deep sense of language contexts. The general-purpose, pretrained models can then be fine-tuned using a small amount of data to fit into downstream tasks. The pretrained models can also serve as foundation models, which can be adapted using zero- or few-shot in-context learning without further fine-tuning to achieve specific tasks. To prevent LLMs from generating biased, false, or harmful content, they often need to go through alignment tuning to align their output and responses with human values [17]. This process takes the **reinforcement learning with human feedback** (**RLHF**) strategy [4] and accompanies the model with a reward mechanism that favors **helpful, harmless, and honest** (**HHH**) responses [2].

With properly engineered prompts, foundation models can fulfill various complex requests, such as helping people write e-mails, planning vacations, solving math problems, offering advice on stock markets, and coding or debugging software programs [35].

3 Applications of LLMs

There have been a growing number of applications of LLMs in various domains, such as marketing and **customer relationship management** (**CRM**), finance, law, healthcare, education, and personal assistance. In CRM, LLMs are being integrated into customer support applications to automatically handle queries, requests, and complaints [57]. For example, using the LangChain framework[1] and the **retrieval augmented generation** (**RAG**) technology [25], LLMs can be

[1] https://www.langchain.com

tailored to incorporate organizational knowledge and contextual information while interacting with customers, providing a seamless conversational user experience [31]. LLMs have also been used to analyze customer needs and offer product recommendations in unmanned retail stores [51]. In the broader marketing area, LLMs have been adopted to generate personalized advertising materials, new marketing insights and ideas, and creative contents [24]. Leveraging LLMs' dynamic messaging and analytical capabilities, companies can predict individual customers' personal preferences based on vast amounts of structured and unstructured data and "send the right message at the right time to customers" [24]. Marketers are also increasingly using foundation models and LLM-based tools (e.g., GPT-4, DALL-E, and Midjourney) to generate advertising materials and creative artwork to promote their products, services, and brands [13]. Moreover, due to elevated concerns with consumer privacy, LLMs have also been applied in understanding consumers' privacy preferences [52].

Financial companies are among the early adopters of LLMs, which are applied for financial reporting, market forecasting, risk management, and financial advising [64]. Specifically, LLMs can be utilized to identify sentiments in analyst reports, news articles, and financial statements to facilitate quantitative and algorithmic trading [11]; recognize patterns and relationships in financial reports, earnings calls, press releases, and other corporate communications to predict mergers and acquisitions and market trends [62]; detect anomalies and deviations to forecast company insolvency, financial distress, and even fraud [22, 29]. One of the successful applications of LLMs in finance is robo-advisors, which have reshaped the landscape of personal finance investing [64]. These LLMs-powered computer systems can provide individual customers with personalized, informed investment recommendations by analyzing market dynamics, while taking into consideration their investment and risk preferences [19]. Domain-specific LLMs have also been developed in the financial sector. Trained and fine-tuned on financial data and documents, FinBERT [18], FinGPT [62], and BloombergGPT [56] have demonstrated outstanding performance in processing financial text and extracting financial information from documents.

LLMs have also been increasingly employed to process and generate legal text [38]. For example, due to the complexity of legal documents and processes, case retrieval usually is a time-consuming, labor-intensive task. With LLMs, a great amount of time can be saved by leveraging LLMs' natural language understanding and reasoning abilities to identify and analyze relevant information presented in legal documents and make decision recommendations [27]. LLMs (e.g., GPT-4) have also been used to predict court judgment outcomes in labor dispute cases [49], draft contracts and reports [54], and annotate adjudicatory opinions and statutory provisions [45]. Augmented by RAG, GPT-4, and a few other LLMs can serve as a tax attorney to offer individual users tax advice and instructions [37].

The healthcare industry has long been using NLP technologies to process unstructured text data, including clinical notes, hospital discharge reports, and medical literature. With the advent of LLMs, there have been more NLP applications in information extraction, document summarization and classification, diagnose recommendation and prediction [28]. For example, BERT and its biomedical variants (e.g., BioBERT) are employed to identify from clinical documents and patient narratives medical terms, such as diseases, symptoms, drugs, and treatments [3]. Researchers have also used RoBERTa as the backbone to develop a deep learning framework to detect people with suicide ideation on social media texts [63]. Google's BART and T5 are used to summarize patient health conditions based on their medical histories, progression notes, lab reports, radiology imageries, and doctor-patient communications [15]. Automated disease classification based on medical coding schemas (e.g., ICD—International Classification of Diseases) can also been assisted using LLMs' abilities to process medical text. In addition, with zero- or few-short prompting, GPT-4 is found to be useful in predicting various diagnoses, health conditions, and in-hospital mortality [66]. A research team at Google has developed a conversational diagnostic AI system named AMIE that demonstrates greater diagnostic accuracy and superior performance in most of the test cases [50].

In the education domain, teachers have been using LLMs to provide students with personalized learning experiences, generate practice materials, motivate students with new ideas and thoughts, and assess student work [34]. However, concerns have also been raised regarding the misuse of LLMs and violations of academic integrity [7].

A significant trend in LLMs' development and application is agentic workflows, which are characterized by their abilities to plan, reason and reflect, use tools, and collaborate [48, 59]. These agentic systems often consist of multiple LLM-based agents that are responsible for different functions and tasks, including perception and sensing, planning and assessment, and action [1]. These agents collaborate and utilize various tools to achieve more complex tasks collectively, ranging from web navigation [1], systematic literature review [44], to trip planning [60]. Platforms like CrewAI[2] and Microsoft's AutoGen[3] offer frameworks and services for users to build agentic workflows to fulfill their specific needs. In the higher education context, takin.ai[4] provides a one-stop-shop for major GPTs in the market and enables educators to teach and students to learn GenAI and build multi-agent systems with minimal setup.

4 The Impacts and Challenges of LLMs

It is no doubt that LLMs will be a big game changer and can unleash tremendous technological, economic, and societal revolutions. Many enterprises and organizations are already preparing for the radical changes that may be brought by applications and adoptions of LLMs, such as automation of routine or mundane tasks and significant reduction in workforce. By adopting and applying LLMs in a timely, strategic manner, firms and organizations can enhance decision-making, improve productivity, reduce costs, and promote customer engagement [42]. For example, it has been reported that GenAI can save marketers over five hours of work per week [43]. A recent survey of professional users of an LLM-powered marketing chatbot reports that the top benefits of adopting LLM-based solutions are increased performance, creativity, and cost efficiency [8]. In the software engineering domain, Copilot has been broadly utilized to generate computer programs, review and debug code, and automate documentation, drastically enhancing productivity, code quality, and innovation [53].

Individuals can also benefit from applications of LLMs. Coupled with smart devices and **Internet of Things** (**IoT**), LLMs can serve as personal assistant or partner to provide individual users with various services and support, significantly improving efficiency, productivity, and user experience [26].

As LLMs are being adopted worldwide, they will also bring broader impacts on society. There have been serious concerns with the rapid advancement of AI and its possible negative impacts on employment, environments, and humanity [12, 32, 58]. These concerns and challenges include but are not limited to the following:

- **The blessing and curse of the scaling law.** It has been found that the performance of LLMs will increase as the model size (i.e., the number of parameters), amount of training data, and computational power increase [21]. Governed by this scaling law, LLMs have continued to grow in model size, taking more data, longer time, and more computational resources to train. It has become increasingly difficult to train LLMs due to their massive

[2] https://www.crewai.com/
[3] https://microsoft.github.io/autogen/0.2/
[4] https://takin.ai/

consumption of data, energy, and costs [35]. Unfortunately, the scaling law also predicts that the marginal improvement on performance will diminish. Other architectures, such as the **Mixture of Experts** (**MoE**) [47], which is adopted by DeepSeek [6], may offer a promising alternative to the traditional LLM architecture by using smaller, specialized expert models to provide better data and energy efficiency.

- **Misaligned responses.** LLMs may generate responses that do not align with human values, objectives, and expectations [17]. Hallucinations, for example, are responses that sound plausible but incorrect. This can cause serious risks and consequences in contexts such as healthcare [16]. Although alignment tuning has been widely employed to encourage LLMs to generate HHH (helpful, honest, and harmless) responses, it remains a challenge for LLMs to completely eliminate hallucinations and other misaligned responses.
- **Interpretability and explainability.** The transformer architecture is essentially a neural network which is a black box with invisible innerworkings. Such a lack of transparency can make it difficult to interpret and explain the rationale behind the decision-making process of a model. Although various approaches, such as **SHapley Additive exPlanations** (**SHAP**) [30], have been proposed to enhance the explainability by measuring the importance of features and their contributions to the decisions, users still are often puzzled by how LLMs generate responses and hesitant to trust the decisions completely.
- **Algorithmic bias.** The training data of LLMs may contain biases from various sources reflecting racial, gender, and other discriminant judgments in humans and society [23]. Trained on these data, LLMs may inherit and amplify such biases, causing the decisions to be unfair for some social groups, communities, or societies. For example, it has been reported that automated facial recognition systems can be discriminative toward certain minority groups [33].

5 An IS Research Agenda on LLMs

There has been a growing number of IS articles on LLMs in the past couple of years. We performed a simple search at the eLibrary of AIS[5] (The Association for Information Systems) using the keywords "large language model" and "LLMs." Figure 1 shows the number of LLM-related articles published at AIS journals and conferences since 2021.

As LLMs continue to advance and make impacts, IS researchers can investigate many interesting research questions, such as the development of LLM-based applications to solve business problems, the behavioral and technical aspects of human-AI interaction, and the ethical and safety issues in using LLMs. We propose a research agenda on LLMs that we believe IS researchers can work on and use to make significant contributions to the research, development, and practice of LLMs.

- **Applications of LLMs.** Due to the extensive demands for data and computational resources for model training, it is nearly impossible, and probably not necessary, for IS researchers to train foundation models from scratch. However, employing the transfer learning strategy [67], it is still feasible to fine-tune or adapt some LLMs to fit into a specific application domain. The FinGPT, for example, is a fine-tuned LLM suitable for financial applications [62]. Alternatively, foundation models can be tailored to specific tasks (e.g., travel planning) by developing properly engineered prompt templates and agentic workflows [48]. Following the computational design science approach [41], IS research along

[5] https://aisel.aisnet.org/. This eLibrary includes *MIS Quarterly* and *Journal of the AIS*, but does not include *ACM Transactions on Management Information Systems*, and other IS journals (e.g., *Information Systems Research, Journal of Management Information Systems*) named in the Senior Scholars' List of Premier IS Journals (https://aisnet.org/page/SeniorScholarListofPremierJournals).

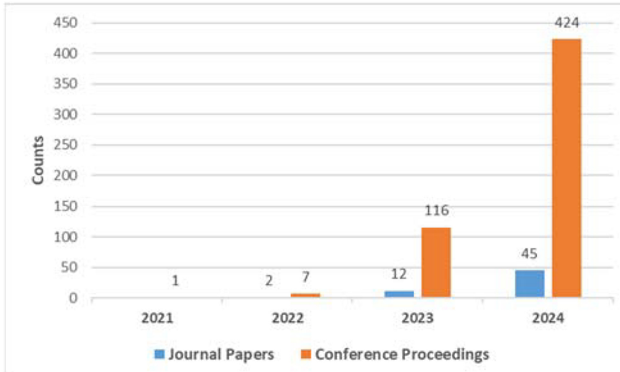

Fig. 1. The number of LLMs-related articles published at AIS journals and conferences since 2021.

this line has great potential to make contributions by developing innovative applications that can be used to address a wide range problems, such as disaster management [61], depression detection [40], and medical knowledge graph construction [55]. Another exciting area of research is the application of LLMs in improving existing research methods or even proposing novel ones. For example, LLMs have been widely used to extract variables from textual data in IS economics studies. In addition, ChatGPT has been used to conduct a case study using the grounded theory approach [65].

— **Evaluation and assessment.** While designing and developing applications of LLMs, it is essential to rigorously evaluate and assess LLM-based artifacts to demonstrate their novelty and practical values. However, there are several challenges in the evaluation of LLMs. First, traditional evaluation metrics for machine learning, such as accuracy, precision, recall, and AUC, may not be applicable for the evaluation of all types of LLMs. Some new metrics are needed for the evaluation of LLMs, especially for generative tasks, which are different from traditional classification and retrieval tasks. Besides technical metrics, other aspects, such as the HHH (helpful, harmless, and honest) criteria [2] and the security of LLMs, are also very important. IS researchers can develop and calibrate new metrics and design appropriate methodologies for the evaluation of LLMs based on these important aspects.
— **Human-AI interaction.** One of the core research areas of IS is the study of the interaction between humans and technology. There are many new research issues and challenges on the interaction between human users and AI applications [5], including LLM-based applications. For example, how much should we trust the output of LLMs, which may suffer from the problem of hallucinations and inaccuracies in training data? What are the factors that influence our trust in LLMs? It is interesting to study how and why people appreciate or avoid the use of LLM-based AI models [20]. Also, given that LLMs may inherit biases from their training data [36], IS researchers can investigate how to mitigate such biases through a design science approach. Another important area of interest is the explainability, interpretability, and transparency of LLMs. Design science researchers in IS can contribute to the design of new methods that can improve the explainability, interpretability, and transparency of LLMs. It would also be highly desirable to study how these factors affect the adoption and acceptance of LLMs and what can be done to enhance user trust.
— **Impacts, ethics, and social good.** AI in general, and LLMs in particular, may be one of the most disruptive technologies in human history. Undoubtedly, LLMs will have broad impacts on many aspects of our life. It would be interesting and important to examine

how LLMs are used by individuals, organizations, and institutions; how LLMs change the way we work, do business, and interact with each other; and what kind of benefits (and costs) LLMs may cause to our productivity and creativity. More importantly, IS researchers should pay special attention to the ethical implications of the use of LLMs, including the risks and consequences LLMs may pose on broader societal issues, such as humanity and social justice [39].

6 This Special Issue

This special issue is intended to showcase IS studies applying, experimenting with, and assessing LLMs. After several rounds of review and revision, eight articles are accepted. These articles provide varying perspectives on LLMs, demonstrating the great potential of IS research to contribute to the literature on LLMs.

Benjamin Ampel, Chi-Heng Yang, James Hu, and Hsinchun Chen propose a framework in the article, entitled "Large Language Models for Conducting Advanced Text Analytics Information Systems Research." This framework is intended to help the IS community embrace LLMs by leveraging LLMs' outstanding capabilities. The framework incorporates a set of guidelines regarding how to conduct meaningful text analytics studies in design science, behavioral, and econometric streams of research. Three business intelligence case studies using the framework were conducted, along with the discussion on the potential challenges and limitations of adopting LLMs for IS. This framework can serve as a useful roadmap for IS researchers who hope to utilize LLMs to analyze large volumes of text in their scientific inquiries.

The article, entitled "Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots" by Aijia Yuan, Edlin G. Colato, Bernice Pescosolido, Hyunju Song, and Sagar Samtani, provides a comprehensive review of chatbots in the mental health domain on their development, application, evaluation, ethical concerns, integration with traditional services, LLM-as-a-service, and various other business implications in organizational settings. Over 50 mental health-related chatbots, including 22 LLM-based models, targeting general mental health, depression, anxiety, stress, and suicide ideation, are identified and analyzed. In particular, this article focuses on mental health in the workplace, where these issues are increasingly prevalent. This article can help IS researchers, who develop LLM-based chatbots and other applications, identify interesting research topics and applications in the healthcare domain.

In "Unraveling the Impact of ChatGPT as a Knowledge Anchor in Business Education," Amrita George, Veda Storey, and Shuguang Hong examine the substitution and complementarity effects of using ChatGPT in business education. Based on the concept of technology anchor (e.g., computer playfulness), an empirical study was conducted to understand whether anchors impact students' goal orientation, learning outcomes, and well-being and how an anchor complements the effects of social support on learning outcomes. It was found that students' LLM use hindered their well-being and learning outcomes, substituting social support for simple and difficult tasks. This research offers useful insights into the adoption of LLMs in educational settings and into the strategies for incorporating LLMs into business curricula.

"Designing Heterogeneous LLM Agents for Financial Sentiment Analysis" is a single-authored article by Frank Xing. This article investigates the effectiveness of using LLMs and proposes a design framework with heterogeneous LLM agents for financial sentiment analysis. The framework consists of several types of agents, including knowledge agents, reasoning agents, and evaluative agents. Comprehensive evaluations using six datasets show that the framework performs better than benchmark multi-LLM agent settings, especially when the discussion contents are substantial. This study showcases how LLM-based multi-agent workflows and systems can be developed and employed in financial data analysis applications.

Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi also demonstrate how LLMs can be used for financial text classification. In their article, entitled "A Comparative Analysis of Instruction Fine-Tuning Large Language Models for Financial Text Classification," the authors investigate the use of instruction fine-tuning relatively small-scale LLMs and evaluate their performance across various settings. The results show that instruction fine-tuned models are more robust than baseline models on the financial text classification tasks.

In "On Leveraging Large Language Models for Multilingual Intent Discovery," Rudolf Chow, King Y. Suen, and Albert Lam show how they used LLMs to discover user intents during their interactions with conversational systems (e.g., chatbots), and present a method for multilingual intent discovery by leveraging multilingual LLMs. By performing joint extraction of intent and keyphrases, as well as a chain-of-thought styled reasoning, this method is shown to efficiently produce clustering results that are easy to interpret.

While LLMs have many viable applications, the safety and ethical issues involved are very important and need research attention. In many situations, it is useful to detect whether a piece of text is written by humans or AI (e.g., fake news detection). The research presented in "A Metric-Based Detection System for Large Language Model Texts" by Linh Le and Dung Tran focuses on the ethics of LLMs and proposes a metric-based method to perform such detection. In particular, the proposed detection method is based on learning similarities among LLM-generated texts and dissimilarities between LLM-generated and human-written texts. Five datasets have been created in this study and can be used in future research on LLM-generated text detection.

In the article entitled "Human-AI Synergy in Survey Development: Implications from Large Language Models in Business and Research," Ping Fan Ke and Ka Chung Ng propose to use LLMs to assist in the task of developing reliable measurement scales for survey in behavioral research. A human-in-the-loop approach was adopted to allow human-AI collaboration in the scale development process. A user study and a simulation test were conducted to evaluate the model and the results are encouraging. This study sheds light on how LLMs can assist in the academic research process.

Acknowledgments

We would like to thank Drs. Daniel Zeng and Heng Xu, the former and present editor-in-chief of TMIS, respectively, for their suggestions on the special issue. We also thank all the submitting authors and reviewers for their contributions, without which the special issue would not be possible. We are also grateful for the help provided by the editorial staff at *TMIS*.

References

[1] T. Abuelsaad, D. Akkil, P. Dey, A. Jagmohan, A. Vempaty, and R. Kokku. 2024. Agent-E: from autonomous web navigation to foundational design principles in agentic systems. arXiv:2407.13032. Retrieved from https://arxiv.org/abs/2407.13032

[2] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma et al. 2021. A general language assistant as a laboratory for alignment. arXiv:2112.00861. Retrieved from https://arxiv.org/abs/2112.00861

[3] T. Babaian and J. Xu. 2024. Entity recognition from colloquial text. *Decision Support Systems* 179, Article 114172.

[4] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback arXiv:2204.05862. Retrieved from https://arxiv.org/abs/2204.05862

[5] N. Berente, B. Gu, J. Recker, and R. Santhanam. 2021. Managing artificial intelligence. *MIS Quarterly* 45, 3 (2021), 1433–1450.

[6] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu et al. 2024. DeepSeek LLM: Scaling open-source language models with longtermism. arXiv:2401.02954. Retrieved from https://arxiv.org/abs/2401.02954

[7] S. A. Bin-Nashwan, M. Sadallah, and M. Bouteraa. 2023. Use of ChatGPT in academia: Academic integrity hangs in the balance. *Technology in Society* 75, Article 102370.

[8] Botco.ai. 2023. Nearly three-quarters of all marketing departments use generative AI tools, botco.ai survey finds. Retrieved January 14, 2025 from https://www.prweb.com/releases/nearly-three-quarters-of-all-marketing-departments-use-generative-ai-tools-botco-ai-survey-finds-863152119.html

[9] Cole. 2024. Most powerful LLMs. CodingScope. Retrieved January 14, 2025 from https://codingscape.com/blog/most-powerful-llms-large-language-models

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved January 14, 2025 from https://arxiv.org/abs/1810.04805

[11] S. Fatemi and Y. Hu. 2023. A comparative analysis of fine-tuned LLMs and few-shot learning of LLMs for financial sentiment analysis. arXiv:2312.08725. Retrieved from https://arxiv.org/abs/2312.08725

[12] E. Ferrara. 2024. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* 7, 1 (2024), 549–569. https://doi.org/10.1007/s42001-024-00250-1

[13] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech. 2024. Generative AI. *Business & Information Systems Engineering* 66, 1 (2024), 111–126. https://doi.org/10.1007/s12599-023-00834-7

[14] R. Gruetzemacher and J. Whittlestone. 2022. The transformative potential of artificial intelligence. *Futures* 135, Article 102884 (2022), 11 pages. https://doi.org/10.1016/j.futures.2021.102884

[15] V. Hartman, S. Bapat, M. Weiner, B. Navi, E. Sholle, and T. J. Campion. 2023. A method to automate the discharge summary hospital course for neurology patients. *Journal of American Medical Information Association* 30, 12 (2023), 1995–2003.

[16] R. Hatem, B. Simmons, and J. Thornton. 2023. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. *Cureus* 15, 9, Article e44720 (2023), 3 pages. https://doi.org/10.7759/cureus.44720

[17] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. 2023. Aligning AI with shared human values. arXiv:2008.02275. Retrieved from https://arxiv.org/abs/2008.02275

[18] A. Huang, H. Wang, and Y. Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research* 40, 2 (2023), 806–841.

[19] Z. Huang, C. Che, H. Zheng, and C. Li. 2024. Research on generative artificial intelligence for virtual financial robo-advisor. *Academic Journal of Science and Technology* 10, 1 (2024), 74–80.

[20] E. Jussupow, I. Benbasat, and A. Heinzl. 2024. An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly* 48, 4 (2024), 1575–1590.

[21] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. 2020. Scaling laws for neural language models. arXiv:2001.08361. Retrieved from https://arxiv.org/abs/2001.08361

[22] A. Kim and S. Yoon. 2021. Corporate bankruptcy prediction with domain-adapted BERT. In *Proceedings of the 3rd Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics, 26–36.

[23] N. Kordzadeh and M. Ghasemaghaei. 2022. Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems* 31, 3 (2022), 388–409.

[24] N. Kshetri, Y. K. Dwivedi, T. H. Davenport, and N. Panteli. 2024. Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda. *International Journal of Information Management* 75, Article 102716 (2024), 10 pages. https://doi.org/10.1016/j.ijinfomgt.2023.102716

[25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2021. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv:2005.11401. Retrieved from https://arxiv.org/abs/2005.11401

[26] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun et al. 2024. Personal LLM agents: Insights and survey about the capability, efficiency, and security. arXiv:2401.05459. Retrieved from https://arxiv.org/abs/2401.05459

[27] H. Li, W. Su, C. Wang, Y. Wu, Q. Ai, and Y. Liu. 2023. THUIR@COLIEE 2023: Incorporating structural knowledge into pre-trained language models for legal case retrieval. arXiv:2305.06812. Retrieved from https://arxiv.org/abs/2305.06812

[28] L. Li, Y. Zhou, Z. Gao, W. Hua, L. Fan, H. Yu, L. Hangen, Y. Zhang, T. L. Assimes, L. Hemphill, and S. Ma. 2024. A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs). arXiv:2405.03066. Retrieved from https://arxiv.org/abs/2405.03066

[29] D. Lin. 2024. Key considerations to be applied while leveraging machine learning for financial statement fraud detection: A review. *IEEE Access* 12 (2024), 168213–168228.

[30] S. Lundberg and S. I. Lee. 2017. A unified approach to interpreting model predictions. arXiv:1705.07874. Retrieved from https://arxiv.org/abs/1705.07874

[31] X. Ma, R. Zhao, Y. Liu, C. Deng, and D. Du. 2024. Design of a large language model for improving customer service in telecom operators. *Electronics Letters* 60, 10, Article e13218 (2024), 7 pages. https://doi.org/10.1049/ell2.13218

[32] J. Manyika, M. Chui, M. Miremadi, J. Bughin, K. George, P. Willmott, and M. Dewhurst. 2017. *A Future that Works: AI, Automation, Employment, and Productivity*. McKinsey Global Institute Research.

[33] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6, Article 115 (2021), 35 pages. https://doi.org/10.1145/3457607

[34] B. Memarian and T. Doleck. 2023. ChatGPT in education: Methods, potentials, and limitations. *Computers in Human Behavior: Artificial Humans* 1, 2, Article 100022 (2023), 11 pages. https://doi.org/10.1016/j.chbah.2023.100022

[35] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian. 2024. A comprehensive overview of large language models. arXiv:2307.06435. Retrieved from https://arxiv.org/abs/2307.06435

[36] R. Navigli, S. Conia, and B. Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality* 15, 2 (2023), 1–21.

[37] J. J. Nay, D. Karamardian, S. B. Lawsky, W. Tao, M. Bhat, R. Jain, A. T. Lee, J. H. Choi, and J. Kasai. 2024. Large language models as tax attorneys: A case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A* 382, 2270, Article 20230159 (2024), 15 pages. https://doi.org/10.1098/rsta.2023.0159

[38] B. Padiu, R. Iacob, T. Rebedea, and M. Dascalu. 2024. To what extent have LLMs reshaped the legal domain so far? A scoping literature review. *Information* 15, 11, Article 662 (2024), 25 pages. https://doi.org/10.3390/info15110662

[39] M. S. Pang, A. Kankanhalli, M. Aanestad, S. Ram, and L. M. Maruping. 2024. Digital technologies and the advancement of social justice: A framework and agenda. *MIS Quarterly* 48, 4 (2024), 1591–1610.

[40] R. Qin, R. Cook, K. Yang, A. Abbasi, D. Dobolyi, S. Seyedi, E. Griner, H. Kwon, R. Cotes, Z. Jiang, and G. Clifford. Forthcoming. Language models for online depression detection: A review and benchmark analysis on remote interviews. *ACM Transactions on Management Information Systems*.

[41] A. Rai. 2017. Editor's comments: Diversity of design science research. *MIS Quarterly* 41, 1 (2017), iii–xviii.

[42] R. Raj, A. Singh, V. Kumar, and P. Verma. 2023. Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 3, 3, Article 100140 (2023), 10 pages. https://doi.org/10.1016/j.tbench.2023.100140

[43] Salesforce. 2023. New research: 60% of marketers say generative AI will transform their role, but worry about accuracy. Retrieved January 14, 2025 from https://www.salesforce.com/news/stories/generative-ai-for-marketing-research/

[44] A. M. Sami, Z. Rasheed, K.-K. Kemell, M. Waseem, T. Kilamo, M. Saari, A. N. Duc, K. Systä, and P. Abrahamsson. 2024. System for systematic literature review using multiple AI agents: Concept and an empirical evaluation. arXiv:2403.08399. Retrieved from https://arxiv.org/abs/2403.08399

[45] J. Savelka and K. D. Ashley. 2023. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence* 6, Article 127979.

[46] M. Schreiner. 2023. GPT-4 architecture, datasets, costs and more leaked. The Decoder. Retrieved January 14, 2025 from https://thedecoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

[47] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv:1701.06538. Retrieved from https://arxiv.org/abs/1701.06538

[48] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei. 2024. Enhancing AI systems with agentic workflows patterns in large language model. In *Proceedings of the IEEE World AI IoT Congress (AIIoT)*. 527–532.

[49] D. H. Sleeman and K. Gilhooly. 2023. Groups of experts often differ in their decisions: What are the implications for AI and machine learning? A commentary on noise: A flaw in human judgment, by kahneman, sibony, and sunstein (2021). *AI Magazine* 44, 4 (2023), 555–567. https://doi.org/10.1002/aaai.12135

[50] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev et al. 2024. Towards conversational diagnostic AI. arXiv:2401.05654. Retrieved from https://arxiv.org/abs/2401.05654

[51] W. Wang, P. Zhang, C. Sun, and D. Feng. 2024. Smart customer service in unmanned retail store enhanced by large language model. *Scientific Reports* 14, 1, Article 19838 (2024), 15 pages. https://doi.org/10.1038/s41598-024-71089-9

[52] W. Wang and B. Li. Forthcoming. Learning personalized privacy preference from public data. *Information Systems Research*.

[53] T. Weber, M. Brandmaier, A. Schmidt, and S. Mayer. 2024. Significant productivity gains through programming with large language models. In *Proceedings of the Proceedings of the ACM on Human-Computer Interaction*. 1–29.

[54] O. Weller, B. Chang, S. MacAvaney, K. Lo, A. Cohan, B. V. Durme, D. Lawrie, and L. Soldaini. 2024. FollowIR: Evaluating and teaching information retrieval models to follow instructions. arXiv:2403.15246. Retrieved from https://arxiv.org/abs/2403.15246

[55] L. Wu, Y. Su, and G. Li. Forthcoming. Zero-shot construction of Chinese medical knowledge graph with GPT-3.5-turbo and GPT-4. *ACM Transactions on Management Information Systems*.

[56] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. 2023. BloombergGPT: A large language model for finance. arXiv:2303.17564. Retrieved from https://arxiv.org/abs/2303.17564

[57] J. Wulf and J. Meierhofer. 2024. Utilizing large language models for automating technical customer support. arXiv:2406.01407. Retrieved from https://arxiv.org/abs/2406.01407

[58] A. V. Wynsberghe. 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics* 1, 3 (2021), 213–218. https://doi.org/10.1007/s43681-021-00043-6

[59] J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li. 2024. Large multimodal agents: A survey. arXiv:2402.15116. Retrieved from https://arxiv.org/abs/2402.15116

[60] J. Xie, K. Zhang, J. Chen, T. Zhu, R. Lou, Y. Tian, Y. Xiao, and Y. Su. 2024. TravelPlanner: A benchmark for real-world planning with language agents. arXiv:2402.01622. Retrieved from https://arxiv.org/abs/2402.01622

[61] B. Yan, F. Mai, C. Wu, R. Chen, and X. Li. 2024. A computational framework for understanding firm communication during disasters. *Information Systems Research* 35, 2 (2024), 590–608.

[62] H. Yang, X.-Y. Liu, and C. D. Wang. 2023. FinGPT: Open-source financial large language models. arXiv:2306.06031. Retrieved from https://arxiv.org/abs/2306.06031

[63] D. Zhang, L. Zhou, J. Tao, T. Zhu, and G. Gao. Forthcoming. KETCH: A knowledge-enhanced transformer-based approach to suicidal ideation detection from social media content. *Information Systems Research.*

[64] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, H. Jiang et al. 2024. Revolutionizing finance with LLMs: An overview of applications and insights. arXiv:2401.11641. Retrieved from https://arxiv.org/abs/2401.11641

[65] Y. Zhou, Y. Yuan, K. Huang, and X. Hu. 2024. Can ChatGPT perform a grounded theory approach to do risk analysis? An empirical study. *Journal of Management Information Systems* 41, 4 (2024), 982–1015.

[66] Y. Zhu, Z. Wang, J. Gao, Y. Tong, J. An, W. Liao, E. M. Harrison, L. Ma, and C. Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. arXiv:2402.01713. Retrieved from https://arxiv.org/abs/2402.01713

[67] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, and H. Zhu. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109, 1 (2021), 43–76.

Received 14 January 2025; revised 14 January 2025; accepted 15 January 2025