No Size Fits All: The Perils and Pitfalls of Leveraging LLMs Vary with Company Size

Ashok Urlana[1] **Charaka Vinayak Kumar**[1] **Bala Mallikarjunarao Garlapati**[1]
Ajeet Kumar Singh[1] **Rahul Mishra**[2]
TCS Research, Hyderabad, India[1] IIIT Hyderabad[2]
ashok.urlana@tcs.com, charaka.v@tcs.com, balamallikarjuna.g@tcs.com
ajeetk.singh1@tcs.com, rahul.mishra@iiit.ac.in

Abstract

Large language models (LLMs) are playing a pivotal role in deploying strategic use cases across a range of organizations, from large pan-continental companies to emerging startups. The issues and challenges involved in the successful utilization of LLMs can vary significantly depending on the size of the organization. It is important to study and discuss these pertinent issues of LLM adaptation with a focus on the scale of the industrial concerns and brainstorm possible solutions and prospective directions. Such a study has not been prominently featured in the current research literature. In this study, we adopt a threefold strategy: first, we conduct a case study with industry practitioners to formulate the key research questions; second, we examine existing industrial publications to address these questions; and finally, we provide a practical guide for industries to utilize LLMs more efficiently. We release the GitHub[1] repository with the most recent papers in the field.

1 Introduction

Large language models (LLMs) have recently garnered significant attention due to their exceptional performance in various predictive and generative tasks (Hadi et al., 2023; Kar et al., 2023). Extensive research has been conducted to harness LLMs across diverse domains and tasks (Raiaan et al., 2024), including medicine (Thirunavukarasu et al., 2023), finance (Li et al., 2023b), and reasoning tasks (Huang and Chang, 2023; Qiao et al., 2023). Despite their unprecedented adaptation to numerous industrial applications, there is a notable lack of studies examining the potential challenges and risks associated with LLMs, which can vary depending on the size of the organization. Such studies would not only be valuable for industries seeking informed adaptation but also help shape research focus to address the key challenges and obstacles faced in real-world scenarios.

The challenges and bottlenecks faced by organizations of different sizes are not uniform. Factors such as funding availability, workforce size, skill and training deficits, ethical and regional considerations, and access to adequate hardware can all influence how these challenges manifest. Previous research has largely addressed general challenges (Raiaan et al., 2024) with LLMs, such as multilingual support, domain adaptation, and compute requirements. However, there is a lack of studies specifically focusing on the industrial perspective and the unique challenges of implementing LLMs in this context.

To this end, we conduct a study with a threefold strategy, firstly, we conduct a rigorous case study of real-world practitioners from the IT industry, who are trying to work on AI adaptation and formulate three guiding research questions. ***RQ1.*** How have industries adopted LLMs so far, and what challenges do they face? ***RQ2.*** What are the barriers hindering the full utilization of LLMs in industrial applications, and how can these barriers be addressed? ***RQ3.*** How can various industries advance to maximize the utility of LLMs in practical applications? Subsequently, with an aim to address guiding research questions, we perform a thorough scoping survey of existing research publications from industrial entities of all sizes. Finally, we discuss our takeaways and insights and present a practical pilot scenario-based guide for industries to adapt to LLMs in a more informed manner.

The key contributions of this work can be summarised as: this study identifies various categories of challenges associated with LLMs for industrial adoption and proposes potential solutions. These challenges broadly relate to data confidentiality, reliability of LLM responses, infrastructure bottlenecks across industries, domain-specific adoption,

[1] https://github.com/vinayakcse/IndustrialLLMsPapers

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 187–203
January 19–24, 2025.

synthetic data generation, and ethical concerns. Additionally, we offer a practical guide tailored for small, medium, and large industries to maximize the utilization of LLMs.

2 Related Work

In the literature, numerous studies focus on practical and ethical challenges associated with LLMs across diverse application domains includes education (Yan et al., 2024), finance (Li et al., 2023d), healthcare (Zhou et al., 2023) and security (Shao et al., 2024). Additionally, several studies address the task-specific challenges for LLMs' adoption in areas such as spoken dialog systems (Inoue, 2023), mathematical reasoning (Ahn et al., 2024), mining software repositories (Abedu et al., 2024). Moreover, studies explore the challenges based on LLMs capabilities with explanations generation (Kunz and Kuhlmann, 2024), data augmentation (Ding et al., 2024), support for multilingual context (Shen et al., 2024) and compliance with ethical challenges (Jiao et al., 2024).

Close to our work, Gallagher et al. (2024) addresses a few concerns on the adoption of LLMs for specific high-stake applications, particularly intelligence reporting workflows. In contrast to existing studies, our work specifically concentrates on the utilization of LLMs for industrial applications. Moreover, this study provides a comprehensive overview of several roadblocks to LLMs adoption for industrial use cases and corresponding potential solutions. Additionally, our study offers a suggestive guide to maximize the utilization of LLMs for various industries.

3 Methodology

This section aims to explore how industries have adopted LLMs and the challenges they face (**RQ1**).

3.1 Industrial Case Study on LLMs

We conduct an industrial case study to understand, how the LLMs are shaping industry practices, identify the underlying challenges and benefits. Through a meticulous process of expert consultation and iterative refinement, the questionnaire was designed to capture insightful data and serve as a tool for understanding the evolving role of LLMs in the industry. This case study covers a multitude of aspects related to LLM usage for specific application domains, corresponding risks, trust attributes, and challenges. In crafting a succinct questionnaire, our objective was to gauge the adoption and impact of LLMs in various industries. These questions can be found in Appendix B Table 4. We receive 26 responses in total from real-world practitioners of the IT industry. We did a case study on 26 companies which are leveraging LLMs for their use-cases. This exercise is non-trivial as most companies have not made their LLM-related use cases public.

3.2 Quantitative Analysis

Based on the responses obtained from the industrial case study, we make the following observations.

Participants of the case study. We shared the questionnaire with the IT professionals, who are either working on LLMs or have developed some solutions. The participants are industry professionals and practitioners with expertise ranging from beginner to expert level.

Widely adapted applications by leveraging LLMs. Even though LLMs are being utilized for various applications, we observe that the majority of these industrial applications are related to financial, retail, security, and healthcare domains.

Modality of the datasets. More than 60% of the industry practitioners prefer to use either textual or tabular data as shown in Figure 1.a.

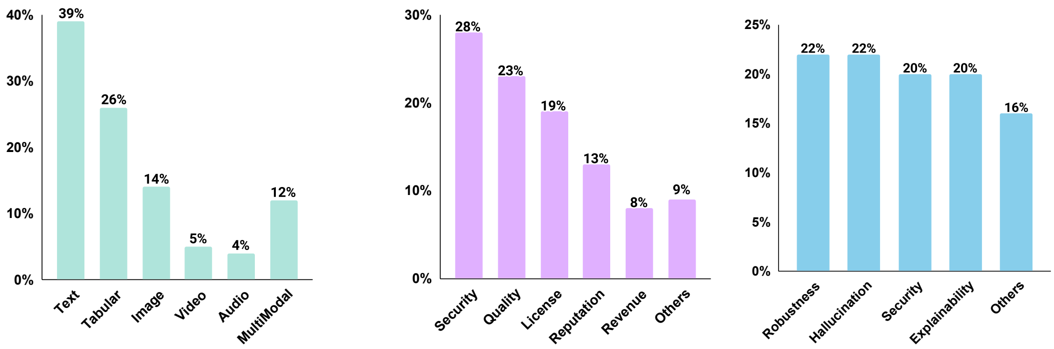
Widely used LLMs. Our case study indicates more than 50% of the applications utilize the GPT-3.5 and GPT-4 models. Recently, researchers have been assessing the capabilities of LLaMA-2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023a).

Prompting strategy. We observe that zero-shot and in-context learning prompting strategies are widely adapted compared to fine-tuning.

Risks associated with LLMs. Based on our case study, LLMs pose risks associated with security and safety, quality of service, and license-related challenges as depicted in Figure 1.b.

Trust attributes to be considered. We observe that robustness, security, and hallucination are the major challenges that need to be considered to utilize the LLMs as shown in Figure 1.c.

Moreover, to gain a better understanding of the barriers to leverage the LLMs for industrial use cases, we also survey 68 research papers specifically from the industry. In this study, we compile several prominent challenges and present potential solutions to address them. The selection criteria for the papers can be found in the Appendix A.



(a) Data modalities considered.

(b) Foreseen risks associated with LLMs.

(c) Opted trust attributes.

Figure 1: Industrial case study statistical overview of various aspects

4 Challenges and Potential Solutions

In this section, we explore several barriers to leveraging LLMs for industrial applications and discuss potential solutions **(RQ2)**.

4.1 Data Confidentiality

4.1.1 Pre-training data issues

Potential privacy risks. To deploy large language models (LLMs) on cloud platforms, robust data privacy protocols are required to handle extensive sensitive datasets while pre-training. Key challenges include mitigating data breaches and preventing unauthorized extraction of sensitive information. Despite the adoption of LLMs in applications like disaster response management (Goecks and Waytowich, 2023), public health intervention (Jo et al., 2023), and assisting Augmentative and Alternative Communication (AAC) users (Valencia et al., 2023), there is noticeable lack of focus on privacy and security aspects. Moreover, it is imperative that potential risks associated with deploying LLMs in high-stakes scenarios are addressed.

Regulations. GDPR in Europe and CCPA in California introduce stringent guidelines for deploying LLMs by enforcing strict data handling and intellectual property rules to ensure transparency and fairness. As highlighted by Mesko and Topol (2023), adhering to these laws in sensitive domains like healthcare is crucial to avoid harm and protect privacy.

Potential solution. Developing a comprehensive framework that aids in LLM compliance is essential for responsible use and interaction with users.

4.1.2 Usage of APIs

To access the closed-source LLMs, passing the commercial data through third-party APIs raises potential privacy concerns (Laskar et al., 2023).

Potential solutions. 1. Robust security and privacy techniques like federated learning are essential to safeguard user data while maintaining the functionality of LLMs, 2. A strategic way of crafting prompts is essential to avoid Personally Identifiable Information (PII) leakage (Kim et al., 2024).

4.2 Reliability of LLMs' Responses

Control the level of AI proactivity. LLMs should minimize social awkwardness, enhance expressiveness, and adapt to different scenarios (Liu et al., 2023b; Urlana et al., 2024). The open-ended generation of LLMs makes it challenging to customize dialog systems for public health intervention applications (Jo et al., 2023).

Outdated knowledge. The open-endedness of LLMs often leads to hallucinations due to a lack of an updated knowledge base (Faizullah et al., 2024). Additionally, the training data might contain errors and become outdated over time.

Potential solutions. Techniques such as Retrieval-Augmented Generation are effective in reducing hallucinations. However, such systems struggle with complex questions that require additional information often generating out-of-context content. Moreover, techniques such as diverse beam search (Vijayakumar et al., 2018), confident decoding (Tian et al., 2019) are promising in mitigating hallucinations. Additionally, model editing techniques (Hoelscher-Obermaier et al., 2023) can address the unintended associations, enhancing the practical usage of LLMs.

4.3 Infrastructure Accessibility

Carbon emissions. Infrastructure is crucial for deploying LLMs, influencing factors like processing speed, latency, cost, and training needs. High-performance hardware is necessary to boost speed and reduce latency, enhancing user experience but it requires careful budgeting due to associated high costs. Achieving an optimal balance between cost and performance is crucial for the efficient training and scalability of LLM applications.
Potential Solution. Implementing robust small language models lead to reduced carbon emissions.
Compute requirements. Despite the state-of-the-art performance of the large language models, utilizing them for small-scale industries is not feasible due to high compute requirements.
API costs. While LLMs like GPT-3.5 and GPT-4 (Achiam et al., 2023) demonstrate superior performance over open-source models, their high cost of API access is prohibitively expensive to perform comprehensive studies (Laskar et al., 2023).
Potential Solution. Balancing the trade-off between performance and cost is necessary for the practical usage of LLMs (Laskar et al., 2023).
High inference latency. APIs can be slow when demand is high. For instance, tasks like business meeting summarization can take GPT-4 around 40 seconds to generate a single response (Laskar et al., 2023). Additionally, longer prompts increase computational demand (Jiang et al., 2023b).
Potential Solution. Open-source models like LLaMA-2 (Touvron et al., 2023) are more favorable for industrial deployment. Further studies on efficient model optimization techniques such as quantization, pruning, and distillation are required (Laskar et al., 2023). Moreover, closed-source models that can utilize prompt compression techniques such as LLMLingua (Jiang et al., 2023b).

4.4 Domain Adaption

Lack of domain-specific datasets. The ability of LLMs in the finance and medical domains is lacking due to insufficient domain-specific training data in the foundation models (Liu et al., 2024; Li et al., 2023c). Consequently, the current versions of GPT-4 and ChatGPT do not meet the industrial requirements to build financial analyst agents (Li et al., 2023c). While LLMs can generate relevant reasoning, they fall short of the desired standard, indicating significant room for improvement.

Diversity. LLMs fail to mitigate social bias due to a lack of diverse demographic data (Lee et al., 2023). Foundation models must equally consider factors like ethnicity, nationality, gender, and religion, as most currently reflect western perspectives.
In-context learning (ICL). The scope of in-context learning is limited by its pretraining data (Han et al., 2023). It is unlikely that any model will perform well when using ICL with data significantly different from its pretraining data.
Potential Solution. 1. Pre-training data should consist of various domain mixtures; however, finding the right mixture is still an open challenge. 2. LLMs should be carefully tested to ensure they treat marginalized individuals and communities equally (Kotek et al., 2023). 3. Continuous pretraining can help overcome the drawbacks of the in-context learning strategy.

4.5 Data Creation Using LLMs

Few works attempt to generate synthetic datasets by utilizing LLMs. However, three major concerns exist with using LLMs for synthetic data creation/annotation; 1). **Lack of diversity.** Synthetic datasets may lack diversity due to the limited knowledge base (Ramakrishna et al., 2023) of LLMs, 2). **Quality and compute.** The quality of the annotated data might improve with the size of the LLM used for the annotation (Sun et al., 2023). However, leveraging large LLMs requires higher computational resources, 3). **In-context learning (ICL) challenges.** ICL is a widely adopted approach for textual task data annotation tasks (Li et al., 2023c). However, the main challenge lies in responsibly incorporating the model's output is to deliver value to users without misleading them or inadvertently amplifying malicious behavior (Deng et al., 2023).
Potential solution. Currently, tools like FABRICATOR (Golde et al., 2023), support tasks like classification, sentence similarity and QA for data labeling and other tasks should be explored.

4.6 Sub-standard Performance of LLMs

Code generation. LLMs' coding ability is limited to generate general-purpose coding tasks. However, the generation of high-quality code for complex network management tasks remains challenging (Mani et al., 2023). Moreover, LLMs have limited capabilities in repository-level coding tasks except in C and Python languages (Bairi et al., 2024) and fail to complete code with potential bugs (Dinh et al., 2024). Most of the code-LLMs struggle with code completion tasks, with undefined names and unused variables (Ding et al., 2023) being the most prominent static error cases.

Conversational applications. LLMs face challenges in providing emotional support and maintaining long-term memory, impacting their effectiveness in conversational applications (Jo et al., 2023). Future research on a longitudinal deployment of LLM-driven chatbots for public health interventions would help understand how users' engagement changes over time.

Multilingual and Multi-Modal: Most of the LLMs are being limited to English, there is significant room for creating robust multilingual models. Only a few studies have focused on utilizing LLMs for such multi-modal industrial applications (Feng et al., 2024; Lu et al., 2023). More efforts are needed to integrate LLMs with voice assistants and Robotics (Yamazaki et al., 2023).

4.7 Explainability and Interpretability

The robust performance of LLMs across various tasks underscores the importance of explainability and interpretability to foster trust in their predictions. However, several challenges impede the development of explainable models.

Black Box Nature: Many popular LLMs, such as ChatGPT and Gemini (Team et al., 2023), are accessible only through APIs, limiting users' understanding of their internal workings.

Scale and Complexity of Models: The large-scale training on vast data leads to complex models, making it hard to identify which parameters influence specific decisions (Brown et al., 2020).

Performance Trade-Off: Balancing model performance with the ability to provide meaningful explanations is a significant challenge; many models struggle to maintain this equilibrium.

Language Ambiguity: The inherent ambiguity of language complicates the generation of clear explanations, as words and sentences can have multiple meanings depending on context (Wang, 2023).

Potential Solutions. **Model Simplification:** Developing simpler models can enhance interpretability, provides a clear understanding of the decision-making processes of LLMs (Che et al., 2016).

Training Data Transparency: Sharing details about training datasets and their sources can illuminate knowledge gaps and potential biases in the models (Bender and Friedman, 2018).

Interactive Exploration Tools: Creating interactive platforms that allow users to manipulate inputs, visualize attention patterns, and observe changes in outputs can provide valuable insights into model behavior (Olah et al., 2018).

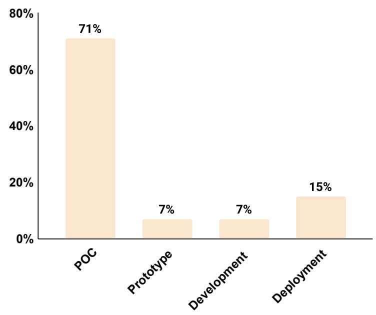


Figure 2: Current state of the industrial applications utilizing the LLMs; POC stands for proof of concept.

4.8 Evaluation of LLMs

In sectors like legal, finance, and healthcare, blending LLMs with human feedback is crucial to lowering false positives, underscoring the importance of human oversight in safety-critical applications (Liu et al., 2023a). Moreover, our analysis (see Appendix C) reveals that less than 15% of studies conduct human evaluations to assess LLM outputs, indicating a need for more rigorous validation methods. Evaluating long-form question answering is challenging for LLMs (Zhao et al., 2023), as additional contextual information may not always be available in practical QA scenarios. Current metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) primarily evaluate the similarity, but are insufficient for assessing the reasonableness of LLM responses.

4.9 Ethical Concerns

The most common ethical challenges with LLMs are violation of the model license, model theft, copyright infringement, producing harmful content, and trustworthiness (Foley et al., 2023) and following are the **potential solutions**.

Protecting LLMs. Watermarking techniques (Peng et al., 2023) are essential for copy-right protection of industrial LLMs, aim to minimize the adverse impact on the original LLM.

Enhancing creativity. AI models should enhance, not replace, human creativity by generating new ideas and insights (Shen et al., 2023).

Fairness in data visualization. Interactive data visualization can help detect and address hidden biases (Kwon and Mihindukulasooriya, 2023).

	Small Scale Industries	Medium Scale Industries	Large Scale Industries
Mode of LLMs usage	API Integration, Pre-trained Models, Low-code or No code platforms, Zero-shot, Few-shot	Domain-specific Fine-tuning, Chain of thought	In-house deployment, Continuous pre-training, Collaborative tools and frameworks, Pre-training from scratch, Re-pretraining
Challenges	Cost, Technical expertise, Data privacy, Performance	Scalability, Domain Adoption	Ethical concerns, Regulations, Data governance
Data modalities	Uni-modal	Multi-modal (Max two)	Multi-modal (2 or more)
Training time	Few hours to days	Few days to weeks	Few weeks to months
Dataset size	100 to 10k samples	10k to 100k samples	More than 100k samples
Compute resources	Cloud	Cloud and Moderate GPUs	In-house high-end GPUs and TPUs
Optimization	Quantization	PEFT techniques, Distillation, Pruning	Prompt compression techniques
Languages	Monolingual	Monolingual	Multi-lingual/Cross-lingual
Ethical complexity	Low	Moderate to high	High to very high
Type and size	Open-source <= 3B	Open-source $\sim$7B	Any open-source model

Table 1: A suggestive guide to various industries to maximize the utilization of LLMs for NLG applications.

Linking models. Techniques such as LLM Attribution (Foley et al., 2023) link fine-tuned models to their pre-trained versions.

Protecting integrity. Guardrails such as NeMo (Rebedea et al., 2023), LangKit[2], and TrustLLM (Sun et al., 2024) help to maintain LLM integrity by preventing biased or inaccurate outputs.

Addressing these challenges requires a combination of technical expertise, ethical considerations, and further research efforts. In Figure 2, we categorize each paper (total of 68) based on its application life cycle and observed that, due to the above-mentioned pitfalls, more than 70% of LLM-based studies are still in the conceptual phase.

5 Maximizing LLM Utilization Across Industries

This section offers a suggestive guide to various industries to maximize the utilization of LLMs for Natural Language Generation (NLG) applications (**RQ3**). As shown in Table 1, our suggestions are tailored to various industries, considering their distinct goals, resources, and workforce capabilities. The recommendations for small and medium-sized industries equally apply to large-scale industries.

1) Small-scale industries such as startups with less than 100 employees need to optimize the use of LLMs within constraints of limited computational resources and workforce. These industries should emphasize prompt engineering and transfer learning techniques to utilize robust small LLMs with up to 3 billion parameters with permissive licenses. Further, these industries should focus on monolingual tasks and actively perform the inference on a few hundred samples. To reduce the inference duration, these industries should opt for optimization techniques such as quantization. Moreover, these industries encounter challenges such as potential reductions in model accuracy, costs, and need for technical expertise. Some of these can be addressed by partnering with AI consulting firms.

2) Medium-scale industries up to 1000 employees should focus on utilizing the RAG-based pipelines and domain-specific parameter efficient fine-tuning and distillation techniques for LLMs up to 7B parameters. Additionally, these industries can develop domain-specific adapters to enhance LLMs' performance on specific tasks. These industries can explore moderate multi-modal (text + vision) tasks. Additionally, the key challenges for medium-scale industries are scalability and domain adoption.

3) Large scale industries such as MNCs should focus on continuous pre-training of LLMs while ensuring compliance with regulatory requirements. These industries can leverage LLMs effectively across multi-lingual, cross-lingual, and multi-modal generation tasks. Training such models can take from a few weeks to months, which requires high-quality data and huge compute as well. These industries should focus on establishing several collaborative tools and frameworks to maximize LLM utilization. For all industries, we recommend using open-source models with appropriate licenses to address ethical concerns and comply with LLM regulatory guidelines. Additionally, robust testing and validation protocols are essential to meet industry standards. Fostering strong collaborations and

[2] https://docs.whylabs.ai/docs/langkit-api/

knowledge sharing between industry and academia is crucial for advancing responsible LLM development and deployment.

6 Conclusions

This study delves into the utilization of Large Language Models (LLMs) through an industrial lens, with a specific focus on identifying roadblocks to their adoption. It meticulously examines various pitfalls and provides potential solutions. Moreover, this study offers a guide to organizations of all sizes to maximize the utilization of LLMs for industrial use cases. By identifying pitfalls and suggesting potential directions, the study offers a strategic road-map for optimizing LLM effectiveness in industrial operations.

7 Limitations

Our study has the following limitations.

Scope. To provide a practical guide to various industries, we restrict our scope to only Natural Language Generation (NLG) applications. Prospective works should focus on providing an extensive guide to various other tasks as well.

Coverage. With the rapid development of LLMs and the voluminous research in this field, it's not feasible to comprehensively cover all the papers. Recognizing this, our survey has focused specifically on industry-related papers. This allowed us to delve deeper and gain an understanding of the unique requirements and challenges faced within industrial applications of LLMs.

Confidentiality. Due to the confidential nature of the industrial applications not many details were available for specific scenarios or challenges. Hence, we only focused on providing recommendations/insights that can be applicable to a broad range of industrial applications.

8 Ethics Statement

To our knowledge, this study presents minimal ethical concerns. However, to maintain transparency, we provide a detailed analysis of all 68 papers present in the survey in Appendix Section C. Each paper is reviewed by at least three individuals to validate its claims and findings. We conduct the industrial case study, following the guidelines outlined by the ACL ethics review policy [3], thereby Ethics Review Boards (ERB) approval is not necessary. It's important to note that our research involving human subjects does not entail the collection of any medical or sensitive information from the users.

[3]https://aclrollingreview.org/ethicsreviewertutorial

References

Samuel Abedu, Ahmad Abdellatif, and Emad Shihab. 2024. Llm-based chatbots for mining software repositories: Challenges and opportunities. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, pages 201–210.

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237.

Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can generative llms create query variants for test collections? an exploratory study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1869–1873.

Amit Alfassy, Assaf Arbelle, Oshri Halimi, Sivan Harary, Roei Herzig, Eli Schwartz, Rameswar Panda, Michele Dolfi, Christoph Auer, Peter Staar, et al. 2022. Feta: Towards specializing foundational models for expert task applications. *Advances in Neural Information Processing Systems*, 35:29873–29888.

Ben Athiwaratkun, Sanjay Krishna Gouda, Zijian Wang, Xiaopeng LI, Yuchen Tian, Ming Tan, Wasi Ahmad, Shiqi Wang, Qing Sun, Mingyue Shang, Sujan Gonugondla, Hantian Ding, Varun Kumar, Nathan Fulton, Arash Farahani, Siddhartha Jain, Robert Giaquinto, Haifeng Qian, Murali Krishna Ramanathan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Sudipta Sengupta, Dan Roth, and Bing Xiang. 2023. Multilingual evaluation of code generation models. In *ICLR 2023*.

Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Talukdar.

2023. Bootstrapping multilingual semantic parsers using large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2455–2467, Dubrovnik, Croatia. Association for Computational Linguistics.

Ramakrishna Bairi, Atharv Sonwane, Aditya Kanade, Arun Iyer, Suresh Parthasarathy, Sriram Rajamani, B Ashok, and Shashank Shet. 2024. Codeplan: Repository-level coding using llms and planning. *Proceedings of the ACM on Software Engineering*, 1(FSE):675–698.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Arno Candel, Jon McKinney, Philipp Singer, Pascal Pfeiffer, Maximilian Jeblick, Chun Ming Lee, and Marcos Conde. 2023. H2O open ecosystem for state-of-the-art large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 82–89, Singapore. Association for Computational Linguistics.

Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2016. Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings*, volume 2016, page 371. American Medical Informatics Association.

Yinfang Chen, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, et al. 2024. Automatic root cause analysis via large language models for cloud incidents. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pages 674–688.

Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. Llmr: Real-time prompting of interactive worlds using large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. What do llms know about financial markets? a case study on reddit market sentiment analysis. In *Companion Proceedings of the ACM Web Conference 2023*, pages 107–110.

Victor Dibia. 2023. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 113–126, Toronto, Canada. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Hantian Ding, Varun Kumar, Yuchen Tian, Zijian Wang, Rob Kwiatkowski, Xiaopeng Li, Murali Krishna Ramanathan, Baishakhi Ray, Parminder Bhatia, and Sudipta Sengupta. 2023. A static evaluation of code completion by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 347–360, Toronto, Canada. Association for Computational Linguistics.

Tuan Dinh, Jinman Zhao, Samson Tan, Renato Negrinho, Leonard Lausen, Sheng Zha, and George Karypis. 2024. Large language models of code fail at completing code with potential bugs. *Advances in Neural Information Processing Systems*, 36.

Abdur Rahman Bin Mohammed Faizullah, Ashok Urlana, and Rahul Mishra. 2024. Limgen: Probing the llms for generating suggestive limitations of research papers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 106–124. Springer.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36.

Besnik Fetahu, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. InstructPTS: Instruction-tuning LLMs for product title summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 663–674, Singapore. Association for Computational Linguistics.

Myles Foley, Ambrish Rawat, Taesung Lee, Yufang Hou, Gabriele Picco, and Giulio Zizzo. 2023. Matching pairs: Attributing fine-tuned models to their pre-trained large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7423–7442, Toronto, Canada. Association for Computational Linguistics.

Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. " i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 205–216.

Shannon K Gallagher, Jasmine Ratchford, Tyler Brooks, Bryan P Brown, Eric Heim, William R Nichols, Scott Mcmillan, Swati Rallapalli, Carol J Smith, Nathan VanHoudnos, et al. 2024. Assessing llms for high stakes applications. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, pages 103–105.

Vinicius G Goecks and Nicholas R Waytowich. 2023. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. In *ICML 2023 DeployableGenerative AI Workshop*.

Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch, and Alan Akbik. 2023. Fabricator: An open source toolkit for generating labeled training data with teacher LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–11, Singapore. Association for Computational Linguistics.

Priyanshu Gupta, Avishree Khare, Yasharth Bajpai, Saikat Chakraborty, Sumit Gulwani, Aditya Kanade, Arjun Radhakrishna, Gustavo Soares, and Ashish Tiwari. 2023. Grace: Language models meet code edits. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1483–1495.

Muhammad Usman Hadi, R Qureshi, A Shah, M Irfan, A Zafar, MB Shaikh, N Akhtar, J Wu, and S Mirjalili. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673, Toronto, Canada. Association for Computational Linguistics.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. 2023. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11548–11559, Toronto, Canada. Association for Computational Linguistics.

Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. 2023. Ralle: A framework for developing and evaluating retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 52–69.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2963–2975.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Koji Inoue. 2023. Challenges and approaches in designing social sds in the llm era. In *Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, pages 24–25.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. 2024. Navigating llm ethics: Advancements, challenges, and future directions. *arXiv preprint arXiv:2406.18841*.

Pengxiang Jin, Shenglin Zhang, Minghua Ma, Haozhe Li, Yu Kang, Liqun Li, Yudong Liu, Bo Qiao, Chaoyun Zhang, Pu Zhao, et al. 2023. Assess and summarize: Improve outage understanding with large language models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1657–1668.

Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Arpan Kumar Kar, PS Varsha, and Shivakami Rajan. 2023. Unravelling the impact of generative artificial intelligence (gai) in industrial applications: A review of scientific and grey literature. *Global Journal of Flexible Systems Management*, 24(4):659–689.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of LLM-generated explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.

Bum Chul Kwon and Nandana Mihindukulasooriya. 2023. Finspector: A human-centered visual inspection tool for exploring and comparing biases among foundation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 42–50, Toronto, Canada. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Amanda Li, Jason Wu, and Jeffrey P Bigham. 2023a. Using llms to customize the ui of webpages. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3.

Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023b. On the steerability of large language models toward data-driven personas. In *CIKM 2023*.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023c. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023d. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew Lungren, Ozan Oktay, and Javier Alvarez-Valle. 2023a. Exploring the boundaries of GPT-4 in radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14414–14445, Singapore. Association for Computational Linguistics.

Xingyu" Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang" Anthony" Chen, and Ruofei Du. 2023b. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36.

Yuzhe Lu, Sungmin Hong, Yash Shah, and Panpan Xu. 2023. Effectively fine-tune to improve large multimodal models for radiology report generation. In *Deep Generative Models for Health Workshop NeurIPS 2023*.

Sathiya Kumaran Mani, Yajie Zhou, Kevin Hsieh, Santiago Segarra, Trevor Eberl, Eliran Azulai, Ido Frizler, Ranveer Chandra, and Srikanth Kandula. 2023. Enhancing network management using code generated by large language models. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*, pages 196–204.

Reza Yousefi Maragheh, Lalitesh Morishetti, Ramin Giahi, Kaushiki Nag, Jianpeng Xu, Jason Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Llm-based aspect augmentations for recommendation systems. *Openreview*.

Bertalan Mesko and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *npj Digit. Med.*

Nandana Mihindukulasooriya, Sarthak Dash, Sugato Bagchi, Ariel Farkash, Michael Glass, Igor Gokhman, Oktie Hassanzadeh, Nhan Pham, et al. 2023. Unleashing the potential of data lakes with semantic enrichment using foundation models. In *ISWC 2023*.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill*, 3(3):e10.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Youngja Park and Weiqiu You. 2023. A pretrained language model for cyber threat intelligence. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 113–122, Singapore. Association for Computational Linguistics.

Nick Pawlowski, James Vaughan, Joel Jennings, and Cheng Zhang. 2023. Answering causal questions with augmented llms. In *ICML 2023 Deployable-Generative AI Workshop*.

Wenjun Peng, Jingwei Yi, Fangzhao Wu, Shangxi Wu, Bin Bin Zhu, Lingjuan Lyu, Binxing Jiao, Tong Xu, Guangzhong Sun, and Xing Xie. 2023. Are you copying my model? protecting the copyright of large language models for EaaS via backdoor watermark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7653–7668, Toronto, Canada. Association for Computational Linguistics.

Savvas Petridis, Michael Terry, and Carrie Jun Cai. 2023. Promptinfuser: Bringing user interface mock-ups to life with large language models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Tung Phung, Victor-Alexandru Pădurean, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. 2023. Generative ai for programming education: Benchmarking chatgpt, gpt-4, and human tutors. *International Journal of Management*, 21(2):100790.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Anil Ramakrishna, Rahul Gupta, Jens Lehmann, and Morteza Ziyadi. 2023. INVITE: a testbed of automatically generated invalid questions to evaluate large language models for hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5422–5429, Singapore. Association for Computational Linguistics.

Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 431–445, Singapore. Association for Computational Linguistics.

Hadeel Saadany and Constantin Orasan. 2023. Automatic linking of judgements to UK Supreme Court hearings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 492–500, Singapore. Association for Computational Linguistics.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *Proceedings of the 17th ACM conference on recommender systems*, pages 890–896.

Minghao Shao, Boyuan Chen, Sofija Jancheska, Brendan Dolan-Gavitt, Siddharth Garg, Ramesh Karri, and Muhammad Shafique. 2024. An empirical evaluation of llms for solving offensive security challenges. *arXiv preprint arXiv:2402.11814*.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing ai support for real-world expository writing tasks. *arXiv preprint arXiv:2304.02623*.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in llms. In *NeurIPS 2023 Second Table Representation Learning Workshop*.

David Sun, Artem Abzaliev, Hadas Kotek, Christopher Klein, Zidi Xiu, and Jason Williams. 2023. DELPHI: Data for evaluating LLMs' performance in handling controversial issues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 820–827, Singapore. Association for Computational Linguistics.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2024. Controllable text summarization: Unraveling challenges, approaches, and prospects - a survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1603–1623, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K Kane. 2023. "the less i type, the better": How ai language models can enhance or impede communication for aac users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *AAAI*, 32.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Chatgpt empowered long-step robot control in various environments: A case application. *IEEE Access*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Bryan Wang, Gang Li, and Yang Li. 2023a. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Liang Wang, Nan Yang, and Furu Wei. 2023b. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Yifei Wang. 2023. Deciphering the enigma: A deep dive into understanding and interpreting llm outputs. *Authorea Preprints*.

Takato Yamazaki, Katsumasa Yoshikawa, Toshiki Kawamoto, Tomoya Mizumoto, Masaya Ohagi, and Toshinori Sato. 2023. Building a hospitable and reliable dialogue system for android robots: a scenario-based approach with large language models. *Advanced Robotics*, 37(21):1364–1381.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–312, Singapore. Association for Computational Linguistics.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 174–184.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023b. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753, Singapore. Association for Computational Linguistics.

Zhen Zeng, William Watson, Nicole Cho, Saba Rahimi, Shayleen Reynolds, Tucker Balch, and Manuela Veloso. 2023. Flowmind: Automatic workflow generation with llms. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 73–81.

Jesse Zhang, Jiahui Zhang, Karl Pertsch, Ziyi Liu, Xiang Ren, Minsuk Chang, Shao-Hua Sun, and Joseph J Lim. 2023. Bootstrap your own skills: Learning to solve new tasks with large language model guidance. In *Conference on Robot Learning*, pages 302–325. PMLR.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.

Chunfang Zhou, Qingyue Gong, Jinyang Zhu, and Huidan Luan. 2023. Research and application of large language models in healthcarecurrent development of large language models in the healthcare fielda framework for applying large language models and the opportunities and challenges of large language models in healthcare: A framework for applying large language models and the opportunities and challenges of large language models in healthcare. In *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, pages 664–670.

A Survey Papers Selection Criteria

We used keywords such as "large language models", "LLM" and "LLMs for industrial applications" for selecting the relevant papers. We selected the majority of papers from the reputed databases including the ACL Anthology[4], ACM Digital library[5], Google Scholar[6], which are known for hosting peer-reviewed articles that meet high academic standards. Subsequently, we finalize suitable research papers for the survey based on the following criteria.

Criteria	Number of papers
arXiv version	37
Non organizational papers	10
Not related to application	6
Relevant	**68**
Total	121

Table 2: Survey papers filtration criteria.

- The paper should be a peer-reviewed and published version.
- At least one of the paper's authors should be from the industry.
- Paper should use at least one or more LLM.
- The paper should report at least one real-world application using LLM(s).

Necessary Concessions: We believe that having at least one author from the industry brought the following advantages.

- We found that considering papers with only researchers from industry led to very few research papers. Also, in recent times, collaboration between academia and industry has rightfully expanded resulting in more practical and applicable research works.
- Also, they brought practical perspectives that were grounded in real-world applications and challenges.

In total, we have collected 121 research papers, and out of them, we have discarded 53 that do not

[4]https://aclanthology.org/
[5]https://dl.acm.org/
[6]https://scholar.google.com/

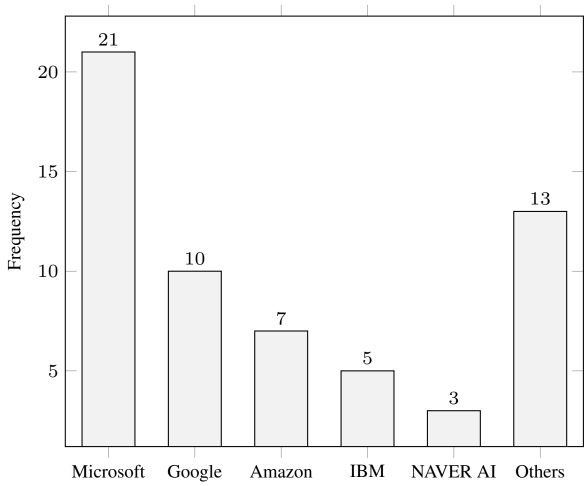


Figure 3: Distribution of research papers from industrial organizations. Others include Apple, Sony, Alibaba, Allen Inst for AI, JP Morgan, Nvidia, Adobe.

fall under one or more above-mentioned criteria as mentioned in Table 2. We have omitted 40 papers because they are not peer-reviewed and 10 more papers came from the non-organizations typically submitted by academic labs/universities. Moreover, we have discarded six papers, which did not discuss any industrial application. After applying the filtering criteria we left with 68 relevant papers. This distribution of the list of papers from various industrial organizations is mentioned in Figure 3.

B Industrial Case Study Details

We have created a questionnaire to conduct the industrial case study as shown in Fig 3.

C Survey Papers Checklist

This paper provides a review of 68 papers and for each paper, we reported 22 features as mentioned in Table 5. We briefly describe each feature in the master table for better understanding.

- *Paper:* Citation of the paper.
- *Venue:* The venue where the paper was published.
- *Year:* Year of paper publication.
- *LLM name:* Names of the LLMs used in the paper.
- *Organization:* Name of the industrial organization involved in the work.
- *Domain:* Domain information of the application in the paper.
- *Application:* The type of application under which the work was categorized into.
- *Use case:* The information of how the paper leverages an LLM in a specific scenario or a task.
- *Dataset Name:* Datasets used by the paper for modeling and evaluation.
- *Prompting Strategy:* Prompting strategies used in the paper.
- *Evaluation metrics:* Details of the evaluation metrics used in the paper.
- *Application life cycle:* Information of application's life cycle stage.
- *GitHub:* Link to the GitHub repository, if any, that was published in the paper.
- *License:* This field indicates if the paper contains license-related information.
- *Privacy:*This field indicates if the paper contains privacy-related information.
- *Use cases:* This field indicates if the paper mentions a use case or not.
- *Limitations:* Major limitations of the paper, if any.

1. Participant level of expertise in LLMs?
   - ☐ Beginner
   - ☐ Intermediate
   - ☐ Proficient
   - ☐ Expert
   - ☐ NA
2. Application Domain
   - ☐ Healthcare
   - ☐ Banking
   - ☐ Financial
   - ☐ Retail
   - ☐ Security
   - ☐ Privacy
   - ☐ Legal
   - ☐ Marketing & Advertising
   - ☐ Education
   - ☐ Media and entertainment
   - ☐ Human Resources(HR)
   - ☐ eCommerce
   - ☐ Other: ___________
3. What is the name of the task that LLM(s) performs in your project?
4. Type of data used?
   - ☐ Tabular
   - ☐ Image
   - ☐ Video
   - ☐ Audio
   - ☐ Text
   - ☐ More than one modality
   - ☐ Other: ___________
5. How are the LLMs used?
   - ☐ Fine-tuning
   - ☐ Zero-shot
   - ☐ In-context learning
   - ☐ Other: ___________

Table 3: Questionnaire for industrial case study: Part 1

6. Did you consider any of the following Trust attributes or guard rails while designing/implementing the LLM-based solution?
   - ☐ Security
   - ☐ Robustness
   - ☐ Privacy
   - ☐ Bias & Fairness
   - ☐ Interpretability or Explainability
   - ☐ Toxicity
   - ☐ Hallucination
   - ☐ None
   - ☐ Other: ____________
7. Name of the LLMs being used?
   - ☐ LLaMA
   - ☐ LLaMA-2
   - ☐ Falcon
   - ☐ Mistral
   - ☐ GPT3.5 (ChatGPT)
   - ☐ GPT4
   - ☐ MPT
   - ☐ Meta OPT
   - ☐ Bard
   - ☐ PaLM
   - ☐ Pythia
   - ☐ Cerebras-GPT
   - ☐ NA
   - ☐ Other: ____________
8. What are the risks associated with the LLMs being used in your project?
   - ☐ Security and Safety
   - ☐ Reputation
   - ☐ Quality of service
   - ☐ Revenue
   - ☐ License
   - ☐ NA
   - ☐ Other: ____________

Table 4: Questionnaire for industrial case study: Part 2

	Paper	Venue	Year	LLMs used	Organization	Domain	Application	Use case	Dataset Name	Prompting strategy	Evaluation metrics	Application life cycle	Github	License	Privacy	Use cases	Limitations
1	Li et al. (2023c)	EMNLP Industry Track	2023	ChatGPT, GPT-4, BloombergGPT, GPT-NeoX, OPT66B, BLOOM176B, FinBERT	J.P. Morgan AI Research	Financial	Analytics	Financial text analysis	FPB/FiQA/TweetFinSent, Headline, NER, REFinD, FinQA/ConvFinQA	Zero-shot, Few-shot and CoT	Accuracy, F1 Score	Conceptualization/PoC	NA	Yes	NA	Yes	NA
2	Chen et al. (2024)	EuroSys	2024	GPT-3.5, GPT-4	Microsoft	Fault diagnosis	Cloud management	Cloud incident root cause analysis	653 incidents from Microsoft's transport service	Zero-shot	Micro and Macro F1-score	Conceptualization/PoC	NA	NA	NA	NA	Effectivness of the method depends on incident monitors/alerts.
3	Bairi et al. (2024)	FMDM@NeurIPS	2023	GPT-4-32k	Microsoft research	Software	Code generation	Automate repository level code planning tasks	Proprietary	Zero-shot	Block metrics, Edit metrics	Conceptualization/PoC	NA	NA	NA	NA	Dynamic languages may not be ideal for a coded plan approach
4	Mani et al. (2023)	HotNet's	2023	GPT-4, GPT-3, Text-davinci-003, Bard	Microsoft research	Communication	Code generation	Code generation for graph manipulation tasks	Public code repositories	Zero-shot	Accuracy	Conceptualization/PoC	Link	Yes	NA	NA	High quality domain specific code synthesis is still an open challenge
5	Ding et al. (2023)	ACL	2023	CodeGen-350M,CodeGen-2B, CodeGen-6B,CodeGen-16B, GPT-3.5	AWS AI Labs	Software	Code generation	Static code analysis for completion	Function completion dataset	None	Percentages of AST errors, Undefine variable, unused variables etc.	Conceptualization/PoC	NA	NA	No	Yes	Cross-file context based broader categorization of errors was not employed
6	Li et al. (2023a)	UIST	2023	Legacy (text-davinci-003), Legacy (code-davinci-002), Legacy (text-davinci-edit-001)	NA	Software	Code generation	Webpage customization using LLMs	NA	Fewshot (In-context learning)	NA	Prototype	NA	No	No	Yes	Poor performance on complex website customization
7	Phung et al. (2023)	ICER	2023	GPT-3.5, GPT-4	Microsoft	Software	Code generation	LLM tutor for programming education	NA	Zero-shot	Match	Conceptualization/PoC	NA	NA	NA	Yes	Limited to python language and introductory educational content
8	Gupta et al. (2023)	ESEC/FSE	2023	CODEX, CODEt5	Microsoft	Software	Code generation	Prediction of code edits using LLMs	C3PO, Overwatch	Fewshot (In-context learning)	Exact match of the code segment	Conceptualization/PoC	NA	NA	NA	Yes	Proposed approach may fail due to LLMs hallucination and context length requirements
9	Dinh et al. (2024)	NeurIPS	2023	CODEGEN, INCODER	AWS	Software	Code generation	Context based code completion	Buggy-HumanEval, Buggy-FixEval	Few-shot and CoT	pass@k	Conceptualization/PoC	Link	Yes	No	Yes	Proposed method may not be aligned to general software development setting as buggy datasets are based on programming contest submissions
10	Athiwaratkun et al. (2023)	ICLR	2023	Decoder-only transformer models	AWS AI Labs	Software	Code generation	Evaluation LLMs on multilingual programming datasets	MBXP, Multilingual HumanEval, MathQA-X	Zero-shot and Few-shot	pass@k scores	Conceptualization/PoC	Link	Yes	NA	Yes	Lack of language specific evaluations
11	Jo et al. (2023)	CHI	2023	HyperCLOVA	NAVER AI Lab, NAVER CLOUD	Healthcare	Conversational	Voice assistant	NA	Zero-shot	NA	Deployment	Link	NA	No	Yes	Skewed age distribution in pilot subjects and pilot deployment time
12	Valencia et al. (2023)	CHI	2023	NA	Google Research	Accessibility	Conversational	Evaluation of LLMs as a tool for AAC users	Proprietary	None	NA	Conceptualization/PoC	NA	NA	NA	Yes	NA
13	Gadiraju et al. (2023)	FAccT	2023	LaMDA	Google Research	Accessibility	Conversational	Chatbot	NA	None	Human evaluation	NA	NA	Yes	No	Yes	Limited diversity of identity within individual focus groups
14	Feng et al. (2024)	NeurIPS	2023	Codex, GPT-3.5, GPT-3.5-chat, GPT-4	Google	Generic	Data Generation	Visual planning for text-to-image generation	NSR-1K, 3D-FRONT	Fewshot (In-context learning)	CLIP cosine similarity, GLIP accuracy, Attribute binding Accuracy, KL divergence	Conceptualization/PoC	Link	Yes	NA	Yes	Generation of overly dense layouts and unusual sized bounding boxes
15	Golde et al. (2023)	ACL	2023	Used existing LLMs from Hugginhface, openAI, Azure, Anthropic, Cohere	Deepset GMBH	Generic	Data Generation	Generation of labeled training data	IMDB, MRPC, SNLI, TREC-6, SQUAD	Zero-shot and Few-shot	F1 score	Conceptualization/PoC	Link	No	No	Yes	Evaluation cover only subset of commonly encountered tasks
16	Yu et al. (2023b)	EMNLP Industry Track	2023	GPT-4, LLaMA	Amazon	Financial	Forecasting	Explainable financial time series forecasting	Stock price data, Company profile data, Finance/Economy News Data	Zero-shot and Few-shot	Binary Precision, Bin Precision, MSE, ROUGE-1,2	Conceptualization/PoC	NA	Yes	Yes	NA	Generalization to other types of financial temporal data unexplored
17	De La Torre et al. (2024)	NeurIPS	2023	Language model for mixed reality (LLMR) Dall.E-2 GPT-4	Microsoft, Microsoft Research	Generic	Frameworks	Generation of interactive 3D objects	NA	NA	Error rate, Average generation time.	Development	NA	NA	NA	Yes	For complex tasks manual code editing might be still necessary
18	Wang et al. (2023a)	CHI	2023	PaLM	Google Research	NLP	Frameworks	Conversational Interaction with Mobile UI	PixelHelp , AndroidHowTo, Rico, Screen2Words,	Zero-shot and Few-shot	Grammar Correctness, UI Relevance, Question Coverage, BLEU, CIDEr, ROUGE-L, and METEOR, Micro-F1	Conceptualization/PoC	Link	Yes	NA	Yes	Fails to handle generation of incorrect or irrelevant information
19	Hoshi et al. (2023)	EMNLP System demonstrations	2023	Llama-2 Chat (13B, 70B), WizardVicunaLM-13B, Vicuna	Kioxia Corporation	Generic	Frameworks	Framework for knowledge intensive tasks	KILT Benchmark	NA	Exact Match, F1, Accuracy, RL, R-precision	Deployment	Link	Yes	NA	Yes	Falls behind specialized RAG models on KILT tasks
20	Zhang et al. (2023)	CoRL	2023	LLaMA-13b	Google AI	Robotics	Frameworks	LLM guided skill chaining	ALFRED	Zero-shot	NA	Development	NA	NA	No	Yes	Greedy skill chaining may not be optimal for consistent behaviour generation
21	Liu et al. (2023a)	EMNLP	2023	GPT-3.5-turbo, text-davinci-003, GPT-4-32k	Microsoft health futures	Healthcare	Frameworks	Evaluation of GPT-4 on understanding and generation of radiology tasks	MS-CXR-T, RadNLI, Chest ImaGenome, MIMIC, Open-i	Zero-shot, Few-shot, One-shot, Many-shot, CoT	macro F1, micro F1, RougeL, CheXbert score	Deployment	NA	NA	Yes	Yes	Qualitative evaluation of the findings on summarization task is limited to a single radiologist
22	Jiang et al. (2023b)	EMNLP	2023	GPT-3.5-Turbo-0301 and Claude-v1.3	Microsoft	NA	None	Prompt compression for higher inference speed	GSM8K, BBH, ShareGPT, Arxiv-March23	Zero-shot	BLEU, ROUGE, BERTScore	Conceptualization/PoC	Link	Yes	NA	Yes	Performance drops when compression of prompts go beyond 25%
23	Yang et al. (2023)	EMNLP Industry Track	2023	GPT-4, GPT3.5, LLaMA-2	Microsoft	NLP	Question-answering	Domain specific industrial QA	MSQA	Zero-shot	BLEU, ROUGE, METEOR, BERTScore, F1, Key-word/Span-Hit-Rate (KHR), Can-Answer-Rate (CAR), LLM-based Metrics	Conceptualization/PoC	NA	NA	NA	NA	Only works with English data
24	Zeng et al. (2023)	ICAIF	2023	GPT-3.5-turbo	J. P. Morgan AI Research	Finance	Question-answering	Dynamic workflow generation	NCEN-QA, NCEN-QA-Easy, NCEN-QA-Intermediate, NCEN-QA-Hard	Zero-shot	NA	Prototype	NA	NA	Yes	Yes	NA
25	Hu et al. (2023)	ICCV	2023	GPT-3	Microsoft, Allen Institute of AI	Vision	Question-answering	Natural language based question-aware caption model	COCO, OK-VQA, A-OKVQA, WebQA	Fewshot (In-context learning)	Accuracy	Conceptualization/PoC	Link	NA	No	Mix	Focuses only on knowledge-based VQA tasks
26	Liu et al. (2024)	NeurIPS	2023	GPT-3.5 turbo, GPT-4, ChatGLM, LLaMA, Vicuna, Alpaca	Alibaba Group, Ant Group	Medical	Question-answering	Medical domain QA	CMExam	Fewshot (In-context learning)	accuracy, weighted F1, BLEU, ROUGE	Conceptualization/PoC	Link	Yes	Yes	Yes	Excluding non-textual questions might introduce unexpected bias
27	Imani et al. (2023)	EMNLP Industry Track	2023	text-davinci-002, PaLM	Microsoft	NLP	Reasoning	Mathematical Reasoning	MultiArith dataset	Zero-shot, Few-shot and CoT	Accuracy	Conceptualization/PoC	NA	Yes	NA	NA	Non trivial probability of producing incorrect results using algebraic and pythonic expressions
28	Lu et al. (2024)	NeurIPS	2023	GPT-3.5-turbo , GPT-4	Microsoft research	Generic	Reasoning	Multi-modal knowledge intensive reasoning tasks	ScienceQA, TabMWP	Zero-shot and CoT	Accuracy	Conceptualization/PoC	Link	Yes	NA	NA	Computationally expensive for complex tasks
29	Li et al. (2023b)	CIKM	2023	GPT-Neo-1.3B, GPT-Neo-2.7B, GPT-J-6B, Falcon-7B-Instruct	Amazon Alexa AI	Political, Education	Reasoning	Steerability of LLM based on persona	OpinionQA	None	User study	Conceptualization/PoC	NA	NA	No	Yes	Complex personas may not be possible
30	Ye et al. (2023)	SIGIR	2023	CODEX	Alibaba Group	Generic	Reasoning	Reasoning on large tables based on textual prompts	TabFact, WikiTableQuestion, FetaQA	Fewshot (In-context learning)	binary classification accuracy, denotation accuracy, BLEU, ROUGE-1, ROUGE-2 and ROUGE-L	Conceptualization/PoC	NA	Yes	Yes	Yes	NA
31	Pawlowski et al. (2023)	ICML Workshop	2023	GPT-3.5, GPT-4	Microsoft Research	Generic	Reasoning	LLM based causal QA system	NA	Text completion	NA	Conceptualization/PoC	NA	NA	NA	Yes	Performance decreases with increased context length
32	Sanner et al. (2023)	RecSys	2023	PaLM	Google	Retail	Recommended systems	Visually augmented real-time conversations	Proprietary	Completion, zero-shot and Few-shot	Mean NDCG	Conceptualization/PoC	NA	No	Yes	Yes	NA
33	Yamazaki et al. (2023)	Advanced robotics	2023	Hyperclova	LINE corporation	Generic	Recommender systems	Voice Chatbot	Proprietary	Fewshot (In-context learning)	Technical score on Informativeness, Naturalness, Likability, Satisfaction with dialog	Testing	NA	NA	No	Yes	For low frequency words gives long responses which floods user with information and hallucination
34	Maragheh et al. (2023)	ICML Workshop	2023	PaLM2	Walmart Global Tech	Retail	Recommender systems	Enhance the capabilities of recommendation systems	Proprietary	Zero-shot	MRR, NDCG	Conceptualization/PoC	NA	NA	NA	NA	NA
35	Alfassy et al. (2022)	NeurIPS	2022	CLIP	IBM Research, MIT-IBM AI-Watson Lab	Vision	Retrieval	Evaluation of LLMs on expert tasks for \ image-to-text and text-to-image retrieval	FETA	Zero-shot, Few-shot, One-shot, Many-shot	Accuracy	Conceptualization/PoC	NA	NA	NA	NA	FETA contains only a small subset of available technical documents for different expert V&L data domains
36	Yu et al. (2023a)	ICLR	2023	InstructGPT	Microsoft Cognitive Service Research	Generic	Retrieval	LLM based retrieval for knowledge-intensive tasks	TriviaQA, WebQ	Zero-shot	Accuracy, F1, ROUGE-L	Conceptualization/PoC	Link	NA	Yes	Yes	Limited ability to update knowledge to new domains
37	Wang et al. (2023b)	EMNLP	2023	Text-davinci-001, Text-davinci-003, GPT-4, Babbage, curie	Microsoft Research	Generic	Retrieval	Query expansion based retrieval systems	MS-MARCO, TREC DL 2019	Fewshot	MRR, nDCG	Conceptualization/PoC		NA	NA	NA	Efficiency of retrieval system
38	Liu et al. (2023b)	CHI	2023	GPT3	Google Research	Generic	Retrieval	Augmenting video conferencing with visual captions	VC 1.5K	Zero-shot	User study	Deployment	Link	Yes	Yes	Yes	Visual captions in conversations should have a threshold to filter out potentially distracting or inappropriate content
39	Lu et al. (2023)	NeurIPS	2023	GPT2-S (117M), GPT2-L (774M) [29], OpenLLaMA-7B (7B)	AWS GAIIC, AWS AI	Healthcare	Retrieval	Writing radiology reports from medical images	MIMIC-CXR	None	Factual completeness and correctness F1-CXB-14 score, F1-CXB-5, BLEU4, ROUGE-L	Conceptualization/PoC	Link	NA	No	Yes	soft visual prompt doesn't receive consistent attention, especially when using LLMs.
40	Alaofi et al. (2023)	SIGIR	2023	text-davinci-003	Microsoft	Generic	Retrieval	Generation of query variants for building test collections and document pool	UQV100	One-shot	Jaccard Index, RBP, RBO	Conceptualization/PoC	NA	NA	NA	Yes	Small size human-generated data is not sufficient for few-shot prompting
41	Park and You (2023)	EMNLP Industry Track	2023	CTI-BERT	IBM T. J. Watson Research Center	Security	Security	Cyber threat intelligence	Attack description, Security Textbook, Academic Paper, Security Wiki, Threat reports, Vulnerability	NA	Micro and Macro F1 Score	Conceptualization/PoC	NA	Yes	Yes	Yes	Pretrained only on English data.
42	Foley et al. (2023)	ACL	2023	BERT, GPT, BLOOM, codegen-350M, DialoGPT, DistilGPT2, OPT, GPT-Neo, xlnet-base-cased, multilingual-miniLM-L12-v2	IBM Research	Generic	Security	Tracing back to the origin of fine-tuned models to alleviates the problem of accountability of LLMs	GitHub, The BigScience ROOTS Corpus, CC-100, Reddit, and THEPILE	NA	F1, ROC	Conceptualization/PoC	NA	Yes	Yes	Yes	Considered only a limited number of LLMs for the study.
43	Peng et al. (2023)	ACL	2023	text-embedding-ada-002, BERT	Microsoft Research Asia, Sony AI, Micorsoft STC Asia	Security	Security	Copy right protection of EaaS (Embeddings as a Service) LLMs	SST2, Mind, Enron Spam, AG news	NA	Accuracy, Detection performance	Conceptualization/PoC	NA	NA	NA	NA	NA
44	Deng et al. (2023)	WWW	2023	GPT-3, PaLM	Google Research	Finance	Sentiment Analysis	Labels generation for financial data	FiQA-News	Fewshot (In-context learning)	Accuracy	Conceptualization/PoC	NA	NA	No	Yes	-
45	Lee et al. (2023)	EMNLP Industry Track	2023	HyperCLOVA (30B and 82B), and GPT-3	NAVER AI Lab	Generic	Societal Impact	Social bias risk mitigation	KoSBi	NA	F1 Score	Conceptualization/PoC	Link	Yes	Yes	Yes	The performance of the filter models are not very competitive
46	Sun et al. (2023)	EMNLP Industry Track	2023	GPT-3.5-turbo-0301, Falcon 40B-instruct, Falcon 7B-instruct, Dolly-v2-12b	Apple	Security	Societal Impact	Comprehensive handling of controversial issues	DELPHI	Zero-shot	Controversy Acknowledgement Rate, Comprehensiveness Answer Rate	Conceptualization/PoC	Link	Yes	Yes	Yes	Dataset may not cover all the controversial questions. and may contain expired ground truth controversy labels
47	Athiwaratkun et al. (2023)	ICLR	2023	Decoder-only transformer model	AWS AI Labs	Software	Code generation	Evaluation of LLMs on multilingual programming datasets	MBXP, Multilingual HumanEval, MathQA-X	Zero-shot and Few-shot	pass@k scores	Conceptualization/PoC	Link	Yes	NA	NA	Does not support language-specific functionalities
48	Laskar et al. (2023)	EMNLP Industry Track	2023	GPT-4, GPT3.5, PaLM-2, LLaMA-2 13b, 7b	Dialpad Canada Inc	NLP	Summarization	Business meeting summarization	QMSUM, AMI, ICSI	Zero-shot	ROUGE, BERTScore	Conceptualization/PoC	NA	Yes	Yes	Yes	Generalizability to domain-specific datasets is in question because only academic datasets were used for testing
49	Fetahu et al. (2023)	EMNLP Industry Track	2023	FLAT-T5	Amazon	NLP	Summarization	Summarization of length product titles	NA	NA	ROUGE, BLEU	Conceptualization/PoC	NA	NA	NA	NA	Do not guarantee inclusion of salient words in the summary
50	Acharya et al. (2023)	RecSys	2023	Alpaca-LoRa	Sony Research India	Retail	Summarization	Generation of product descriptions sans webscrapping	MovieLens, Goodreads Book graph	Fewshot (In-context learning)	Hit Rate, Normalized Discount Cumulative Gain (NDCG), Mean Reciprocal Rank (MRR)	Conceptualization/PoC	NA	NA	NA	NA	Generates factually incorrect descriptions
51	Jin et al. (2023)	ESEC/FSE	2023	GPT-3.X	Microsoft	Infrastructure	Summarization	Cloud outage management	Proprietary	NA	BLEU-4, ROUGE-L and METEOR	Deployment	NA	NA	NA	Yes	Evaluation metrics not fully reflect readability and usefulness of outage summary
52	Shen et al. (2023)	CHI In2Writing Workshop	2023	NA	Allen Institute of AI	NLP	Summarization	Evidence based knowledge generation	NA	Fewshot (In-context learning)	NA	NA	NA	NA	NA	Yes	NA
53	Zhao et al. (2023)	EMNLP Industry Track	2023	GPT4, TULU, Pythia, Alpaca, Vicuna, LLaMA-2, GPT-3.5	Allen Institute for AI	NLP	Table-to-text-generation	Evaluation of LLMs on table-to-text generation	LOTNLG, F2WTQ	Zero-shot and Few-shot	BLEU, ROUGE, SP-Acc, NLI-Acc, TAPAS-Acc, TAPEX-Acc, Exact-match, F1 Score, Accuracy	Conceptualization/PoC	NA	Yes	NA	Yes	NA
54	Mihindukulasooriya et al. (2023)	ISWC	2023	GPT4, Llama2, FLAN-T5	IBM Research	NLP	Table-to-text-generation	Automation of meta data generation and enrichment	NA	None	NA	Conceptualization/PoC	NA	NA	NA	NA	NA
55	Singha et al. (2023)	NeurIPS	2023	GPT-3.5 (text-davinci-003 endpoint)	Microsoft	Generic	Table-to-text-generation	Noise induction to better understanding of table structures	AirQuality, HousingData, Diabetes, Wine Testing, Iris, Titanic, and ENB2012_data	Zero-shot	F1 score	Conceptualization/PoC	Link	Yes	NA	NA	Performance of structural tasks with downstream task such as table question answering remains an open challenge.
56	Wake et al. (2023)	IEEE Access	2023	ChatGPT	Microsoft	Robotics	Task Planning	Translating natural language instructions to executable robot actions	NA	Fewshot	Executability, Correctness	Conceptualization/PoC	Link	Yes	Yes	Yes	Only static environment is considered
57	Saadany and Orasan (2023)	EMNLP Industry Track	2023	GPT- text-embedding-ada-002	Kingfisher Labs Ltd , Just Access	Legal	Tool	Automatic linkage of judgements to bookmarks in court hearing videos	UK National Archive	Zero-shot	Mean Average Precision (MAP), Recall	Conceptualization/PoC	NA	Yes	Yes	NA	NA
58	Petridis et al. (2023)	CHI Extended Abstract	2023	-	Google Research	Generic (HCI)	Tool	input-output interaction , Frame change	NA	Zero-shot	Questionnaire	Conceptualization/PoC	NA	NA	NA	Yes	Needs formal evaluation and in-depth analysis on how functional prompts effect prototyping process
59	Dibia (2023)	ACL	2023	NA	Microsoft Research	Generic	Tool	Automatic generation of grammar-agnostic visualizations and infographics	Proprietary	Zero-shot and Few-shot	Visualization Error Rate (VER), Self-Evaluated Visualization Quality (SEVQ), code accuracy, data transformation, goal compliance, visualization type, data encoding, and aesthetics	Prototype	Link	Yes	No	Yes	Code execution step increases computational complexity.
60	Singh et al. (2023)	ICRA	2023	text-davinci-*, Codex, GPT3	Nvidia corporation	Robotics	Tool	Generate programmatic robot instructions using LLMs	NA	Fewshot (In-context learning)	Success rate(SR), Goal condition recall(GCR), Executability(Exec)	Development	Link	Yes	No	Yes	Robotic action success feedback is not shared with agent leading to failure scenarios
61	Awasthi et al. (2023)	EACL	2023	mT5-Large, PaLM	Google research India	NLP	Translation	Translating English datasets to several other languages	MTOP, MASSIVE	Zero-shot and Few-shot	corrected Exact Match	Prototype	NA	NA	No	Yes	Computationally expensive
62	Kwon and Mihindukulasooriya (2023)	ACL	2023	BERT, ALBERT, RoBERTa	IBM Research	Generic	Trustworthy AI	Inspect the fairness and Bias of foundation models	CrowS-Pairs	NA	NA	Conceptualization/PoC	Link	NA	NA	Yes	The tool's effectiveness not tested for decoder-only models.
63	Ramakrishna et al. (2023)	EMNLP	2023	GPTNeo-2.7B, GPTJ-6B, Open-LLaMA-7B, RedPajama-7B,GPT3.5-Turbo, GPT4	Amazon Alexa AI	Security	Trustworthy AI	Evaluating the LLMs for Hallucinations	DBpedia, TriviaQA	Zero-shot	BLEU, ROUGE, METEOR, BERTScore, AlignScore	Conceptualization/PoC	Link	NA	NA	NA	Lack of diversity in test set.
64	Kotek et al. (2023)	CI	2023	Not disclosed	Apple	Generic	Trustworthy AI	Identify the presence of gender bias in LLMs	Proprietary	Zero-shot	NA	Conceptualization/PoC	NA	NA	NA	Yes	Results may not reflect the real gender bias
65	Wan et al. (2023)	EMNLP	2023	ChatGPT, Alpaca	Adobe Research	Generic	Trustworthy AI	Identify the presence of gender bias in LLMs	WikiBias-Aug	Zero-shot	WEAT	Conceptualization/PoC	Link	NA	Yes	Yes	Only consider binary gender when analyzing biases
66	Kim et al. (2024)	NeurIPS	2023	OPT-350M OPT-1.3B OPT- 2.7B	NAVER AI Lab, Parameter Lab	Generic	Trustworthy AI	Probing for PII in a given LLM	Pile	Fewshot (In-context learning)	Sentence matching, Likelihood ratio	Conceptualization/PoC	NA	NA	Yes/	Yes	Evaluation dataset uses private information sourced from open-source datasets
67	Rebedea et al. (2023)	EMNLP System demonstrations	2023	text-davinci-003, GPT-3.5-turbo, falcon-7b-instruct, llama2-13b-chat	NVIDIA	Generic	Trustworthy AI	Tool kit for adding programmable guardrails for conversational LLMs	Anthropic Red-Teaming and Helpful datasets	Fewshot (In-context learning)	Accuracy	Development	Link	Yes	NA	Yes	Toolkit not suitable as standalone solution
68	Candel et al. (2023)	EMNLP System demonstrations	2023	Generic	H2O.ai	Generic	Trustworthy AI	Deploy and test efficiency of wide variety of LLMs on private databases and documents	NA	NA	NA	Deployment	Link	Yes	Yes	Yes	Datasets, Biases and Offensiveness, Usage, Carbon footprint, Hallucinations of LLMs

Table 5: Master table of the survey with 68 research papers.