

LLMs with Industrial Lens: Deciphering the Challenges and Prospects – A Survey

Ashok Urlana¹ Charaka Vinayak Kumar¹ Ajeet Kumar Singh¹
 Bala Mallikarjunarao Garlapati¹ Srinivasa Rao Chalamala¹ Rahul Mishra²
 TCS Research, Hyderabad, India¹ IIIT Hyderabad²

ashok.urlana@tcs.com, charaka.v@tcs.com, ajeetk.singh1@tcs.com
 balamallikarjuna.g@tcs.com, chalamala.srao@tcs.com, rahul.mishra@iiit.ac.in

Abstract

Large language models (LLMs) have become the secret ingredient driving numerous industrial applications, showcasing their remarkable versatility across a diverse spectrum of tasks. From natural language processing and sentiment analysis to content generation and personalized recommendations, their unparalleled adaptability has facilitated widespread adoption across industries. This transformative shift driven by LLMs underscores the need to explore the underlying associated challenges and avenues for enhancement in their utilization. In this paper, our objective is to unravel and evaluate the obstacles and opportunities inherent in leveraging LLMs within an industrial context. To this end, we conduct a survey involving a group of industry practitioners, develop four research questions derived from the insights gathered, and examine 68 industry papers to address these questions and derive meaningful conclusions. We maintain the Github¹ repository with the most recent papers in the field.

1 Introduction

Large language models (LLMs) have gained significant attention recently for their exceptional performance in various predictive and generation tasks. Despite their robust performance across various applications, there remains a notable absence of standardized studies focused on the practical utilization of LLMs in industrial applications as well as the corresponding industrial requirements and deployment challenges.

To comprehend the current state of LLMs adaptation in the industry, we propose a unique two-stage approach as follows, 1) we conduct a case study tailored for industry practitioners of LLMs 2) From the case study, we frame several research questions and address them by conducting a survey of industrial research works.

¹<https://github.com/vinayakcse/IndustrialLLMsPapers>

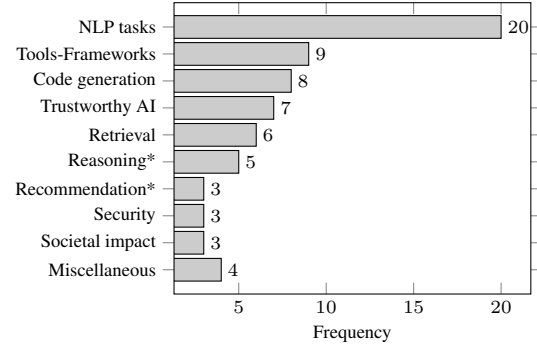


Figure 1: Number of papers from various applications, *entries merged with other application categories.

To conduct the case study, we requested industry practitioners to respond to a questionnaire enquiring about the application domain, corresponding task, and their associated challenges. The detailed analysis of the industrial case study can be found in Appendix B. We have obtained 26 responses from several mid-sized companies. With this study, we observed that LLM adoptions in the industry are facing critical challenges such as compute requirements, privacy concerns, and open access. Considering these observations, we framed the following research questions to be addressed in this work.

RQ1. How do LLMs assist industrial applications, and in what ways? **RQ2.** What are the primary applications that industries are focusing on, including the associated datasets and evaluation metrics? **RQ3.** What are the deployment challenges, if any? **RQ4.** What are the potential directions to maximize the utility of LLMs in industrial applications?

With this study, our objective is to answer the above-mentioned research questions. To this end, we collect and analyze 68 research papers specific to industrial applications (see Appendix A). Further, we categorize the selected papers into 8 broad application domains as shown in Figure 1. We observe an uneven distribution of papers related to

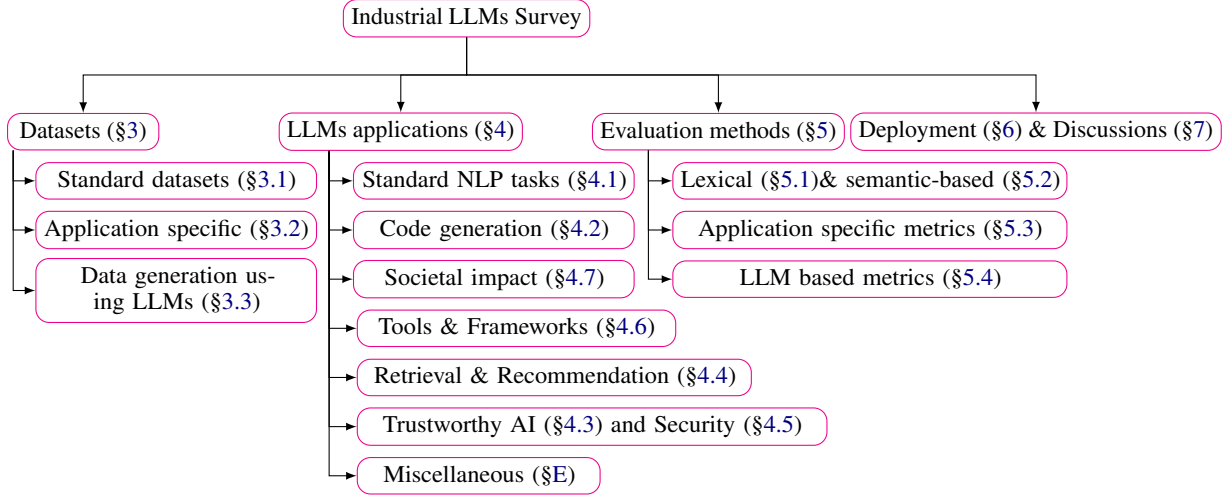


Figure 2: Structure of the paper.

these application domains. Mostly LLMs are utilized for standard NLP tasks (Laskar et al., 2023; Yang et al., 2023; Awasthi et al., 2023), code generation tasks (Phung et al., 2023; Gupta et al., 2023), building tools and frameworks (Hoshi et al., 2023). However, there is less focus given to security (Peng et al., 2023) and societal impact (Sun et al., 2023) related applications.

Some of the prominent use cases include the utilization of GPT4 (Achiam et al., 2023) for automatic summary generation of organizational meetings (Laskar et al., 2023) and the adaption of PaLM (Chowdhery et al., 2023) for low-resource machine translation tasks (Awasthi et al., 2023). Moreover, LLMs are widely used to ensure the security and privacy of the users’ data (Kim et al., 2021), mitigate bias, and improve fairness (Kwon and Mihindukulasooriya, 2023). As depicted in Figure 2, our survey includes the analysis of various datasets, models, associated challenges, and future directions to utilize LLMs for industrial applications.

The key contributions of this survey include, 1) To the best of our knowledge, this is the first survey that provides the industrial perspective on the utilization of LLMs for various applications, 2) We conduct a case study involving industrial practitioners to gain insight into the practical applicability of LLMs, 3) We delve into various datasets, methods and evaluation metrics and corresponding limitations relevant to industrial applications, 4) We discuss the deployment challenges and outline the future directions aimed to maximize the utility of LLMs.

2 Related Surveys

Recent emergence of LLMs present significant opportunities for the advancements of various industrial applications (Hadi et al., 2023; Kar et al., 2023). Comprehensive surveys have been conducted to enhance our understanding of the efficacy of LLMs in diverse domains and tasks, such as medical (Thirunavukarasu et al., 2023), finance (Li et al., 2023b) and reasoning tasks (Huang and Chang, 2023; Qiao et al., 2023).

In the realm of code generation, LLMs have gained widespread usage, and a recent survey (Zan et al., 2023) delves into the associated challenges including code generation and code completion. Few surveys focus on the trust aspects of LLMs (Liu et al., 2023d; Ishihara, 2023), while others offer a detailed exploration of in-context learning techniques (Li, 2023), feedback learning models (Kirk et al., 2023), and the capability of LLMs to adapt to the constantly evolving world knowledge (Zhang et al., 2023b). In contrast to existing surveys, our work specifically concentrates on utilization of LLMs for industrial applications. Additionally, this survey provides a comprehensive overview of datasets, models, evaluation metrics, corresponding challenges, and future perspectives.

3 Benchmark Datasets

This section outlines the datasets used for the development of industrial applications and their creation approaches. These datasets can be broadly divided into two categories (i) popular task-specific benchmark datasets and, (ii) newly proposed datasets, which are automatically created using LLMs, de-

rived either from standard datasets or manually curated with the help of human annotators. A portion of **RQ2** is addressed in this section.

3.1 Standard Datasets

Due to the ease of access, many papers in our survey utilize task-specific standard datasets to evaluate the performance of LLMs. More details of the datasets and their corresponding applications can be found in Appendix D.

3.2 Application-specific Datasets

To better evaluate the capabilities of LLMs, task-specific datasets are more crucial than generic ones. We have identified such datasets and listed them in Appendix D. This section describes the task-specific datasets and their creation strategies.

KOSBI. Lee et al. (2023) introduced a social bias detection dataset for the Korean language by using Hyperclova (Kim et al., 2021). This dataset consists of context-sentence pairs, where the context contains safe-unsafe labeled scenarios mentioning a specified demographic group and the corresponding sentence is the textual content. **DELPHI.** To mitigate bias and fairness in LLMs, Sun et al. (2023) propose the DELPHI dataset. The questions present in the dataset are selected from the ‘Quora Question Pair’ dataset². **WikiBias-Aug.** Wan et al. (2023) concentrates on evaluating the extent of gender bias present in LLMs-generated reference letters. WikiBias-Aug dataset was generated by selecting instances from WikiBias (Sun and Peng, 2021) and altering the male and female names. In another study, **CMExam** (Liu et al., 2023a) dataset was introduced to evaluate the performance of LLMs on open-ended QA task for the Chinese medical data. The raw dataset was collected from the Chinese National Medical Licensing Examination portal. Further, this dataset was refined partially with the help of ‘GPT-assisted annotation’ and reviewed by two medical professionals. Yang et al. (2023) propose **MSQA**, a novel QA dataset designed for Microsoft products and IT technical problems faced by its customers. This dataset helps address the problem of potential oversight of crucial domain-specific information by LLMs. **FETA.** (Alfassy et al., 2022) dataset was introduced to evaluate and improve the performance of LLMs on expert domain tasks. It presents vision learning scenarios such as image-to-text and text-to-image

retrieval tasks. FETA dataset was obtained from the IKEA’s yearly catalogues³. Another study, Mani et al. (2023) explores the effectiveness of LLMs in network management systems by introducing a benchmark called **NemoEval**. Each instance in the dataset contains a user query or network management instruction and an answer created by human experts. Zhao et al. (2023b) introduce **LoTNLG & F2WTQ** to investigate table-to-text capabilities of LLMs. LoTNLG is based on LOGICNLG (Chen et al., 2020) test set, where each entry was labeled by two individual annotators based on the logical reasoning type. Dataset F2WTQ is based on WTQ (Pasupat and Liang, 2015) - a human-annotated QA dataset. For each question in WTQ a corresponding entry in F2WTQ was created, that contains a question and a sentence-long response about a table. Furthermore, to resolve the bugs in the code generation and completion tasks, Dinh et al. (2023) propose **Buggy-HumanEval** and **Buggy-FixEval** datasets. Each instance of these datasets consists of a partial code snippet with bugs, failing test-case, and completed code. ‘buggy-HumanEval’ is a subset of ‘HumanEval’ (Chen et al., 2021) dataset containing 1896 buggy programs. Whereas ‘buggy-FixEval’ contains 292 realistic buggy programs submitted to coding contests and was adapted from ‘FixEval’ (Haque, 2023) and CodeNet (Puri et al., 2021) datasets. Similarly, Athiwaratkun et al. (2023) introduces **MBXP, Multilingual HumanEval, MathQA-X** datasets for evaluating the LLM’s performance on code generation and reasoning over multiple programming languages. These datasets were adapted from MBPP (Austin et al., 2021), HumanEval(Chen et al., 2021), MathQA-python(Schubotz et al., 2018) by converting them to multiple programming languages using the proposed framework in the (Athiwaratkun et al., 2023).

3.3 Data generation using LLMs

Few works attempt to generate synthetic datasets by utilizing LLMs. Golde et al. (2023) introduced **FABRICATOR**, an open-source toolkit for generating labeled datasets using LLMs for tasks like text classification, entailment, question-answering, and entity generation through zero-shot and few-shot prompting. Similarly, Yin et al. (2023) introduced a novel paradigm for curating instruction-tuning datasets. Feng et al. (2023) proposes an approach to improve the accuracy of image generative mod-

²<https://www.kaggle.com/c/quora-question-pairs>

³<https://github.com/ivc-yz/SSR?tab=readme-ov-file>

els by augmenting the input prompts with a layout of the required objects. Multiple variants of GPT models have been used to generate the layouts in 2-D and 3-D image and scene generation tasks.

4 LLMs for Real-world Applications

This section offers an overview of the diverse approaches employed to use LLMs in industrial applications and also answers (RQ1). To leverage the LLMs for any application, usage of proper prompting strategy is necessary. To elaborate on this, we have included details of prompting strategies in Appendix C.

4.1 Standard NLP tasks

Summarization. Summarization for business documents plays a crucial role in obtaining the salient aspects of a document. In their work, Laskar et al. (2023) focuses on summarizing Automatic Speech Recognition (ASR) transcripts of organizational meetings. They evaluate the performance of the open-source (LLaMA-2 (Touvron et al., 2023)) and closed-source (GPT-3.5 (Brown et al., 2020), GPT-4 (Achiam et al., 2023), PaLM-2 (Anil et al., 2023)) LLMs for summary generation. Likewise, to obtain the summaries of product titles of an e-commerce website, Fetahu et al. (2023) presents an instruction fine-tuning approach constrained by the length, number of words and specific phrases. Similarly, Jin et al. (2023) presents the summary generation of outage understanding of cloud-based servers. Apart from this, LLMs have also been adopted for generating expository write-ups (Shen et al., 2023) and movie descriptions (Acharya et al., 2023).

Limitations: Generalizing summarization methods for domain-specific datasets and reducing factual inaccuracies needs attention.

Question-answering. To enrich domain-specific knowledge of LLMs, Yang et al. (2023) collected a cloud domain QA dataset and introduced a novel model interaction paradigm. The method involves pre-training a LM on domain-specific data and passing the user’s query to LM to extract domain knowledge. The acquired information and the query are given as inputs to LLM for response generation. This method works *only* with English data.

To facilitate the automatic workflow generation, Zeng et al. (2023) proposes a novel approach called Flowmind using GPT. It ensures secure interactions between LLMs and proprietary code or data to protect privacy. To perform the visual QA task,

Hu et al. (2023) proposed the PROMPTCAP - a novel prompt-guided image captioning technique, that takes the prompt to describe the visual entities present in the image in the generated caption. PROMPTCAP is trained on synthetic examples generated by GPT3 and is *limited* to knowledge based visual QA tasks.

Machine translation. To obtain effective translations for low-resource languages, Awasthi et al. (2023) perform the few-shot prompting on LLaMA and translate English semantic parsing datasets into 50 more languages. However the proposed approach is *constrained* by compute resource availability.

Conversational: To build a dialog system for public health intervention, Jo et al. (2023) designed the CareCall - an open-ended LLM-integrated chatbot to interact with specific individuals and reduce loneliness and emotional distress. CareCall faces challenges that are attributed to *limited* personalization. Similarly, Valencia et al. (2023) employs LLMs to assist augmentative and alternative communication (AAC) users with extended replies and answering biographical questions, aiming to save time and cognitive effort.

Sentiment analysis. Deng et al. (2023) uses PaLM-540B (Chowdhery et al., 2023) to generate weak sentiment labels for sentiment analysis on financial market data. This approach, training models with weakly labeled data, outperforms traditional methods relying on fully labeled datasets.

Reasoning. To assess the reasoning capability of LLMs, Imani et al. (2023) conducted experiments on arithmetic problems using GPT-3. Their approach resulted in multiple representations of an arithmetic problem, including algebraic and Pythonic forms. These representations can serve as intermediate steps that may be used for reasoning the solution. In another study, Lu et al. (2023a) proposes a plug-and-play style framework to improve the reasoning capabilities of LLMs on multi-modal complex tasks by composing external tools and modules depending on the type of task.

Li et al. (2023b) proposes a novel framework to enhance the steerability of LLMs by introducing ‘persona’, which is based on user characteristics (age, gender). Clusters representing different personas are formed, and their embeddings are appended to input prompts, guiding LLMs through soft prompting. In a related domain, Ye et al. (2023) explores reasoning on tabular data and presents a

unique approach to generate reasoning based on sub-evidences. This strategy decomposes complex tabular data into smaller tables without losing any context. Similarly, the complex questions are divided into simple SQL sub-queries. After obtaining answers for the sub-queries using smaller tables as evidence, a final reasoning answer is derived. The performance of proposed approach *degrades* with increased context length.

Table-to-text generation. This task focuses on creating coherent and contextually accurate descriptions from structured tabular data. Building on this concept, Zhao et al. (2023b) conducted a study to investigate the table-to-text capabilities of LLMs across diverse datasets and confirm the potential of LLMs to accurately interpret and convert structured data into descriptive text. Whereas, Singha et al. (2023) addressed challenges faced by LLMs in processing semi-structured tabular data with inconsistencies. Mihindukulasooriya et al. (2023) leveraged the LLMs to enhance table metadata and enrich the glossaries, taxonomies, or ontologies.

4.2 Code Generation

Given that LLMs are trained on extensive code repositories, experimental studies have been done to evaluate the capabilities of LLMs for multiple tasks including code generation, code completion, program repair, hint generation, grading feedback, pair programming, contextualized explanation, and task synthesis (Phung et al., 2023). One such work (Li et al., 2023b) explores the usage of LLMs for user-driven webpage modification, particularly useful for accessibility reasons or for less technically literate. The source code of the webpage is divided into smaller portions and passed to the LLM along with user-modified prompts.

Bairi et al. (2023) proposes CodePlan - a framework to automate the repository level code editing tasks such as package migration (C#) and temporary code edits (Python) across multiple repositories using the GPT-4 model. However, CodePlan *may not be ideal* for dynamic languages. Another study utilizes LLMs to enhance the network management experience by generating high-quality code for graph manipulation tasks (Mani et al., 2023). The same approach also tackles issues such as explainability, privacy, and scalability in task-specific code generation. On a similar line, Gupta et al. (2023) introduces an approach called GRACE to predict the code edits to fix the bugs or to add

new features using CODEX and CODET5 (Chen et al., 2021; Bavarian et al., 2022) models. GRACE *may fail* when additional context is required apart from associated edits. Similarly, Dinh et al. (2023) introduced the concept of a buggy code completion problem to generate code suggestions in contexts that potentially contain bugs.

Athiwaratkun et al. (2023) introduced a framework to generate execution-based test cases for programming languages beyond Python. It converts Python datasets to multiple languages in a scalable manner. The evaluation of this approach indicates that training a single multilingual model outperforms several monolingual ones. On the other hand, to perform the execution-based evaluation, Ding et al. (2023) introduced a static evaluation framework specifically for the Python language, that operates by parsing code snippets into Abstract Syntax Trees (ASTs).

4.3 Trustworthy AI

Nowadays, service providers are emphasizing on trust in AI, focusing on securing and protecting data privacy. To detect the potential privacy violations in LLMs, Kim et al. (2023) proposes ProPILE, a tool that enables external users and in-house LLM developers to create custom prompts to identify if private data is being used for LLM training. To mitigate issues such as perpetuated biases, risk of leaking private information, and hallucinations regarding factual data, Candel et al. (2023) introduced a suite of fine-tuned LLMs, along with a comprehensive framework for their evaluation, fine-tuning, and deployment. To specifically handle ‘hallucinations’, Ramakrishna et al. (2023) proposed the INVITE - a framework that acts as a testbed, enabling the generation of invalid questions by intentionally distorting valid facts.

In another study, Kwon and Mihindukulasooriya (2023) proposed *Finspector*, a tool that allows users to evaluate models for bias and fairness, particularly in sentences exhibiting stereotypes (Kotek et al., 2023). The proposed tool’s effectiveness on decoder-only models is *not tested*. Recently, aligned models are designed to encode ethical considerations, and user values, and incorporate safety guardrails. Towards this aim, Rebedea et al. (2023) introduced an open-source toolkit that simplifies adding programmable guardrails to LLM-based conversational systems.

4.4 Retrieval & Recommendation

LLMs demonstrate remarkable language generation abilities to perform various information retrieval tasks. Recently, Liu et al. (2023c) outlines an approach using LLMs to enhance human-human verbal communication by proactively suggesting relevant visuals that reflect the conversation's visual intent. The authors implemented Visual Captions as a Chrome plugin that aims to render live captions and relevant images directly onto the camera video stream for online video conferences.

To improve the retrieval quality, Wang et al. (2023c) proposed two techniques (i) document expansion - enriches the document representation by appending additional relevant terms, (ii) query expansion - rewriting the query based on relevant feedback. They proposed a method for query disambiguation where a new query is formed by performing few-shot prompting. *text-davinci-003* was employed to retrieve a pseudo-document which is concatenated with the original query to form a new query for the retrieval task. In this approach, inference is *slowed down* by token-by-token autoregressive decoding strategy. In another study, Alaofi et al. (2023) explored the utility of a GPT-3.5 to automatically generate queries and query variants from story descriptions. Alfassy et al. (2022) proposed an approach that combines Multiple Instance Learning (MIL) and CLIP (Radford et al., 2021) foundational models for text-to-image and image-to-text retrieval tasks in expert domains. Especially non-natural image documents, user manuals, patents and infographics were used.

Similarly, Lu et al. (2023b) proposed a Large Multimodal Model (LMM) for radiology report generation. This LMM comprises a visual encoder (ResNet50 (He et al., 2016)), that takes an X-ray as input and outputs visual tokens. Subsequently, the visual features embeddings are used as soft prompts to a causal language model (GPT (Radford et al., 2019)) and OpenLLaMA-7B (Geng and Liu, 2023)). Yu et al. (2023a) proposed a generate-then-read pipeline by utilizing Instruct-GPT (Ouyang et al., 2022) to generate contextual documents based on a given query and then perform the retrieval to produce the final answer. However, this approach has *limited* ability to update its knowledge state and adapt to new domains.

Despite the progress in the conventional item-based recommendations for e-Commerce platforms, Sanner et al. (2023) utilizes language-based

preferences as well to provide more consistent recommendations with the help of LLMs. Yamazaki et al. (2023) builds a conversational system and integrates it with a voice-enabled robot for hospitable real-time tourism suggestions. In another study, Maragheh et al. (2023) focuses on evaluating learning-to-rank models that utilize user intentions, derived from LLMs through product reviews, to highlight aspects such as education, fun, and durability while browsing on an e-commerce website.

4.5 Security

LLMs, accessible via APIs for advanced reasoning tasks, also pose risks of model extraction attacks, leading to unauthorized replication. To this end, Peng et al. (2023) proposed a method to embed watermarks in the language model which will act as a backdoor within the embeddings to ensure reliable copyright verification.

Task specific fine-tuned LLMs often raise significant legal issues, such as violations of licensing agreements and copyright infringement. Addressing this crucial need, Foley et al. (2023) proposed an LLM attribution method. This method links fine-tuned models to their original base models, ensuring integrity and responsible use of LLMs. However, this study *lacks* the analysis on large-scale finetuned-models. Additionally, significant advancements in cyber threat intelligence have been made with a pre-trained CTI-BERT model (Park and You, 2023) for cybersecurity, enhancing threat intelligence system accuracy.

4.6 Tools and Frameworks

LLMs have significantly advanced the development and evaluation of technologies across various fields by providing access to sophisticated frameworks and tools. Notably, Hoshi et al. (2023) introduced RALLE, a framework for developing Retrieval Augmented LLMs to tackle hallucination issues. Other innovative applications include leveraging LLMs for interactive 3D object creation (De La Torre et al., 2023) and domain-specific evaluations in radiology (Liu et al., 2023b), as well as LIDA by Dibia (2023) for automatic visualization generation. Tools like PromptInfuser integrate AI into mockups via LLM prompts (Petridis et al., 2023), while another tool (Singh et al., 2023) employs few-shot prompting with LLMs for generating robotic programming instructions. Additionally, a study (Saadany and Orasan, 2023) utilizes GPT for correlating court judgments with hearings,

improving legal analysis. Wang et al. (2023b) explores LLMs for mobile UI conversational interfaces, but *fails* to handle irrelevant and hallucinated responses. Conversely, Zhang et al. (2023a) demonstrates the use of LLMs for automatically adapting to new domains with “own skills” bootstrapping.

4.7 Societal Impact

The widespread adoption of LLMs must be tempered by addressing potential societal and human risks such as toxicity and bias that can adversely affect society. To mitigate the unsafe societal bias in LLMs, Lee et al. (2023) developed a large-scale dataset - KOSBI focusing on Korean language and culture. It employs filtering-based moderation to reduce societal biases in LLM-generated content. Sun et al. (2023) examines the interaction of LLMs with controversial issues and proposes methods to enhance their understanding and management of complex societal debates. Similarly, Goecks and Waytowich (2023) introduces DisasterResponseGPT, a framework that generates actionable plans by leveraging LLMs for efficient disaster response strategy creation, that offers a potential solution to transform disaster response operations. The proposed framework is *unable* to process image-based inputs. The detailed the analysis of the miscellaneous applications present in Appendix E.

5 Evaluation Approaches

This section mainly discusses various types of evaluation metrics and answer **RQ2** by grouping the evaluation metrics into the following categories.

5.1 Lexical overlap-based metrics

Due to their simplicity, many works in our study utilize the lexical overlap-based metrics, which measure the alignment score based on n-grams matching between the source and target.

Examples: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Denkowski and Lavie, 2014), Accuracy, F1 Score, ExactMatch.

5.2 Semantic overlap-based metrics.

Semantic overlap-based metrics exploit contextual embeddings from pretrained language models to assess the similarity between gold and system-generated outputs. **Examples:** BERTScore (Zhang* et al., 2020), AlignScore (Zha et al., 2023).

5.3 Application specific metrics

This section describe the task-specific metrics related to each application, where as complete list of metrics details mentioned in Appendix F Table 5.

Table-to-text generation. (a). *SP-Acc*: verify the correctness by extracting the meaning representation from the generated sentences and executing against the Table and *NLI-Acc* works as a faithfulness classifier by utilizing a pretrained TableBERT (Chen et al., 2020) model. (b). *TAPAS-Acc* and *TAPEX-Acc* metrics uses the TAPAS (Herzig et al., 2020) and TAPEX (Liu et al., 2021) models as backbone, which are fine-tuned on TableFact (Chen et al., 2019) dataset. **Question-answering.** (a). *Keyword/Span-Hit-Rate* - creates a distinct keyword or span set by extracting keywords from grounded answers and measure the rate of hits in the generated response (Yang et al., 2023). (b). *Can-Answer-Rate* - represents the percentage of answerable questions and evaluates the informativeness (Yang et al., 2023). **Retrieval and recommendation systems.** To evaluate the performance of LLMs recommendations, metrics used such as Hit Rate, Normalized Discount Cumulative Gain and Mean Reciprocal Rank. (a). *Rank Biased Overlap*: Quantify the consistency in the retrieved search results for various query variants (Webber et al., 2010). (b). *Rank Biased Precision*: Measures as the ratio of expected document relevance to the expected number of examined documents (Moffat and Zobel, 2008). **Societal impact.** (a). *Controversy Acknowledgement Rate*: If LLMs response consists of "As an AI language model..." reminds that the response is non-human perspective and conventionally used as an indicator for the controversy acknowledgment rate. (b). *Comprehensiveness Answer Rate*: Measures the LLMs response diversity and extensive knowledge in the presence of real-time scenarios (Sun et al., 2023; Alaofi et al., 2023). **Code generation.** (a). *Block metrics*: aims to identify the code blocks in need of modifications. (b). *Edit metrics*: measures the correctness of the code modifications (Bairi et al., 2023). (c). *pass@k*: require k generations to address a given problem, deeming the problem resolved if any generation successfully passes the corresponding unit tests (Chen et al., 2021). **Trustworthy AI.** *Word Embedding Association Test (WEAT)*: assess and quantify the associations between different sets of words in word embeddings (Caliskan et al., 2017).

5.4 LLM Based Evaluation

A recent study (Yang et al., 2023) employed LLM as an evaluator for ranking the responses based on the similarity with the grounded answer. Due to space constraint, we include the human evaluation details in Appendix F Section F.2.

6 Deployment Challenges

In this section, we discuss some of the potential deployment challenges and solutions, observed in our study (RQ3).

Potential privacy risks. Deploying LLMs on cloud platforms necessitates strong data privacy and security protocols to manage large, sensitive datasets. The main challenges include preventing data breaches and the unauthorized extraction of sensitive information. Implementing robust security and privacy technologies is essential to protect user data while ensuring LLMs remain functional.

Infrastructure. Infrastructure is crucial for deploying LLMs, influencing factors like processing speed, latency, cost, and training needs. The necessity for high-performance hardware to boost speed and reduce latency enhances user experience but comes at a high cost, demanding careful budgeting. Achieving an optimal balance between cost and performance is crucial for the efficient training and scalability of LLM applications.

Control the level of AI proactivity. LLMs should reduce social awkwardness, improve expressiveness, and adapt to different scenarios (Liu et al., 2023c). Moreover, LLMs' open-ended nature makes it difficult to customize dialog systems in public health intervention applications (Jo et al., 2023).

Regulations. GDPR in Europe and CCPA in California introduce major hurdles for deploying LLMs by enforcing strict data handling and intellectual property rules, requiring transparency and fairness. These laws necessitate navigating diverse regulations, complicating LLM deployment. As highlighted by Mesko and Topol (2023), adhering to these laws in sensitive domains like healthcare is crucial to avoid harm and protect privacy. Therefore, developing a comprehensive framework for LLM compliance is essential for responsible use and interaction with users.

7 Challenges and Future Perspectives

In this section, we discuss the various challenges and future perspectives to maximize the utility of

LLMs for industrial applications (RQ4).

Maximizing LLMs' potential in healthcare, retail, and finance requires high-quality, labeled datasets. Additionally, with most LLMs being limited to English, there is significant room for creating robust multilingual models.

LLMs face challenges in emotional support and long-term memory, impacting conversational applications (Jo et al., 2023). 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., 2023b). Moreover, our analysis, detailed in Appendix F.2, reveals that less than 15% of studies conduct human evaluations to assess LLM outputs, indicating a need for more rigorous validation methods.

Despite the adoption of LLMs in applications like disaster response management (Goecks and Waytowich, 2023), public health intervention (Jo et al., 2023), and assisting AAC users (Valencia et al., 2023), there is a noticeable lack of focus on privacy and security aspects. Moreover, it is imperative to address the potential risks associated with deploying LLMs in high-stakes scenarios.

With the rising demand for multimodal applications, only a few studies focused on utilizing LLMs for such industrial applications (Feng et al., 2023; Lu et al., 2023b). Likewise, more efforts are needed on integrating LLMs with voice assistants and Robotics (Yamazaki et al., 2023). As shown in Figure 7, we found that more than 70% of LLM-based projects are still in the conceptual phase.

8 Conclusions

This survey delves into the utilization of Large Language Models (LLMs) through an industrial lens, with a specific focus on identifying adoption challenges and avenues for promising use cases. Additionally, it meticulously examines the datasets, methodologies, and evaluation metrics used, along with their limitations across different industrial applications of LLMs. By addressing deployment challenges and suggesting future directions, the survey offers a strategic roadmap for industries to optimize the effectiveness of LLMs in their operations.

9 Limitations

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 understanding of the unique requirements and challenges faced within industrial applications of LLMs. Due to the page limit, we cannot include all technical details. Retrieval-augmented generation (RAG) has enhanced the performance of the LLMs by injecting external knowledge beyond their static training data. However, this survey specifically examines LLMs' industry applications, with minimal focus on RAG to focus solely on their direct applications and inherent challenges.

10 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 G. Each paper is reviewed by at least three individuals to validate its claims and findings. Furthermore, we did not collect any personal information as part of our case study.

References

- 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 Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- 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.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. [Palm 2 technical report](#). *arXiv preprint arXiv:2305.10403*.
- Ben Athiwaratkun, Sanjay Krishna Gouda, Zijian Wang, Xiaopeng LI, Yuchen Tian, Ming Tan, Wasi Ahmad, Shiqi Wang, Qing Sun, Mingyue Shang, Sujan Gonnondla, 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*.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. [Program synthesis with large language models](#). *arXiv preprint arXiv:2108.07732*.
- 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, Shashank Shet, et al. 2023. [Codeplan: Repository-level coding using llms and planning](#). *arXiv preprint arXiv:2309.12499*.
- Mohammad Bavarian, Angela Jiang, Heewoo Jun, and Henrique Pondé. 2022. [New gpt-3 capabilities: Edit & insert](#). OpenAI Blog.
- 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.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. [Semantics derived automatically from language corpora contain human-like biases](#). *Science*, 356(6334):183–186.
- 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.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph,

- Greg Brockman, et al. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. [Logical natural language generation from open-domain tables](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.
- Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. 2019. [Tabfact: A large-scale dataset for table-based fact verification](#). *International Conference on Learning Representations*.
- Yinfang Chen, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, et al. 2023. [Empowering practical root cause analysis by large language models for cloud incidents](#). *arXiv preprint arXiv:2305.15778*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. [Palm: Scaling language modeling with pathways](#). *Journal of Machine Learning Research*, 24(240):1–113.
- Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2023. [Llmr: Real-time prompting of interactive worlds using large language models](#). *arXiv preprint arXiv:2309.12276*.
- 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.
- Michael Denkowski and Alon Lavie. 2014. [Meteor universal: Language specific translation evaluation for any target language](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.
- 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.
- 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. 2023. [Large language models of code fail at completing code with potential bugs](#). *arXiv preprint arXiv:2306.03438*.
- Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. [Layoutgpt: Compositional visual planning and generation with large language models](#). *arXiv preprint arXiv:2305.15393*.
- 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.
- Xinyang Geng and Hao Liu. 2023. [Openllama: An open reproduction of llama](#). URL: https://github.com/openlm-research/open_llama.
- Vinicius G Goecks and Nicholas R Waytowich. 2023. [Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios](#). *arXiv preprint arXiv:2306.17271*.
- 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*.
- Md Mahim Anjum Haque. 2023. [Fixeval: Execution-based evaluation of program fixes for competitive programming problems](#). Ph.D. thesis, Virginia Tech.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. [TaPas: Weakly supervised table parsing via pre-training](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. 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.
- Shotaro Ishihara. 2023. [Training data extraction from pre-trained language models: A survey](#). In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 260–275, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Pengxiang Jin, Shenglin Zhang, Minghua Ma, Haozhe Li, Yu Kang, Lique Li, Yudong Liu, Bo Qiao, Chaoyun Zhang, Pu Zhao, et al. 2023. [Assess and summarize: Improve outage understanding with large language models](#). *arXiv preprint arXiv:2305.18084*.
- Eunkyoung 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.
- Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 2021. [What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. [Propile: Probing privacy leakage in large language models](#). *arXiv preprint arXiv:2307.01881*.
- Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Rottger, and Scott Hale. 2023. [The past, present and better future of feedback learning in large language models for subjective human preferences and values](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore. Association for Computational Linguistics.
- 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.

- 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, Takyoun 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. 2023. [A practical survey on zero-shot prompt design for in-context learning](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 641–647, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- 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. 2023a. [Benchmarking large language models on cmexam—a comprehensive chinese medical exam dataset](#). *arXiv preprint arXiv:2306.03030*.
- Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. [Tapex: Table pre-training via learning a neural sql executor](#). *International Conference on Learning Representations*.
- 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. 2023b. [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. 2023c. [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.
- Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klockhov, Muhammad Faaiz Taufiq, and Hang Li. 2023d. [Trustworthy llms: a survey and guideline for evaluating large language models' alignment](#). *arXiv preprint arXiv:2308.05374*.
- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023a. [Chameleon: Plug-and-play compositional reasoning with large language models](#). *arXiv preprint arXiv:2304.09842*.
- Yuzhe Lu, Sungmin Hong, Yash Shah, and Panpan Xu. 2023b. [Effectively fine-tune to improve large multi-modal models for radiology report generation](#). *arXiv preprint arXiv:2312.01504*.
- 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 Gahi, 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*.

- Alistair Moffat and Justin Zobel. 2008. [Rank-biased precision for measurement of retrieval effectiveness](#). *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27.
- Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. [Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). *Advances in Neural Information Processing Systems*, 35:27730–27744.
- 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.
- Panupong Pasupat and Percy Liang. 2015. [Compositional semantic parsing on semi-structured tables](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. 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*.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. [The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only](#). *arXiv preprint arXiv:2306.01116*.
- 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.
- Ruchir Puri, David Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian Dolby, Jie Chen, Mihir Choudhury, Lindsey Decker, et al. 2021. [Codenet: A large-scale ai for code dataset for learning a diversity of coding tasks](#). In *Annual Conference on Neural Information Processing Systems*.
- 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.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. [Learning transferable visual models from natural language supervision](#). In *International conference on machine learning*, pages 8748–8763. PMLR.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*.
- 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.
- Moritz Schubotz, Philipp Scharpf, Kaushal Dudhat, Yash Nagar, Felix Hamborg, and Bela Gipp. 2018. [Introducing mathqa: a math-aware question answering system](#). *Information Discovery and Delivery*, 46(4):214–224.
- 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. [Prog-prompt: 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](#). *arXiv preprint arXiv:2310.10358*.
- 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.
- Jiao Sun and Nanyun Peng. 2021. [Men are elected, women are married: Events gender bias on Wikipedia](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online. Association for Computational Linguistics.
- 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.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- 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.
- Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. [Chatgpt empowered long-step robot control in various environments: A case application](#). *arXiv preprint arXiv:2304.03893*.
- 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.
- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. [Towards understanding chain-of-thought prompting: An empirical study of what matters](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.
- Bryan Wang, Gang Li, and Yang Li. 2023b. [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. 2023c. [Query2doc: Query expansion with large language models](#). *arXiv preprint arXiv:2303.07678*.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. [A similarity measure for indefinite rankings](#). *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucion, François Yvon, et al. 2022. [Bloom: A 176b-parameter open-access multilingual language model](#). *arXiv preprint arXiv:2211.05100*.
- 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.
- 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.
- Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. [Dynosaur: A dynamic growth paradigm for instruction-tuning data curation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4031–4047, Singapore. Association for Computational Linguistics.
- 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.
- Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji, and Jian-Guang Lou. 2023. [Large language models meet NL2Code: A survey](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7443–7464, Toronto, Canada. 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.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating factual consistency with a unified alignment function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.
- Jesse Zhang, Jiahui Zhang, Karl Pertsch, Ziyi Liu, Xiang Ren, Minsuk Chang, Shao-Hua Sun, and Joseph J Lim. 2023a. [Bootstrap your own skills: Learning to solve new tasks with large language model guidance](#). In *Conference on Robot Learning*, pages 302–325. PMLR.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. [Opt: Open pre-trained transformer language models](#). *arXiv preprint arXiv:2205.01068*.
- Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023b. [How do large language models capture the ever-changing world knowledge? a review of recent advances](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023a. [A survey of large language models](#). *arXiv preprint arXiv:2303.18223*.
- Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023b. [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.

A Survey Papers Selection Criteria

We have selected the research papers for the survey based on the following criteria.

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

In total, we have collected 121 research papers, and out of them, we have discarded 53 that do not fall under one or more above-mentioned criteria as mentioned in Table 1. 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 list of papers from various industrial organizations mentioned in Figure 6.

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

Table 1: Survey papers filtration criteria.

B Industrial Case study on LLMs for Real-world Applications

In crafting a succinct questionnaire, our objective was to gauge the adoption and impact of LLMs in various industries. This case study covers multitude of aspects related to LLMs usage for specific application domains, corresponding risks, trust attributes and challenges. These questions can be found in Table 2. The goal was to study how LLMs are shaping industry practices, identifying both 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 industry.

B.1 Analysis

Based on the responses obtained from the industrial case study of LLMs, we made the following analysis.

1. *Who are the participants of the case study?* We shared the questionnaire with the mid-range companies, who are either working on LLMs or developed some solutions. The participants are industry professionals and practitioners ranging in their expertise from beginner to expert level.
2. *What are widely adapted applications by leveraging LLMs?* Despite LLMs being utilized for various applications, we observed that major applications related to financial, retail, security, and healthcare are utilizing the LLMs.
3. *Modality of the dataset used in the applications.* More than 90% of the applications prefer to use either textual or tabular data as shown in Figure 3.
4. *Additional advantages of using LLMs.* Our case study suggests that advanced NLP capabilities and state-of-the-art performance on a wide range of applications are the major advantages of LLMs.
5. *Widely used LLMs.* Our case study suggests that more than 50% of the applications utilize the GPT-3.5 and GPT-4 models. Recently, researchers have been utilizing the capabilities of LLaMA-2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023a) models as well.
6. *Prompting strategy.* Our case study indicates that zero-shot, few-shot, and in-context learning prompting strategies are widely adapted compared to Fine-tuning.
7. *Risks associated with using LLMs.* Many industry practitioners believe that LLMs exhibit the risks associated with security and safety, quality of service and license-related challenges as depicted in Figure 4.
8. *Trust attribute to be considered.* In our case study, it is observed that robustness, security, and hallucination are the major attributes that need to be considered to utilize any LLMs. We detailed the distribution of the responses in Figure 5.

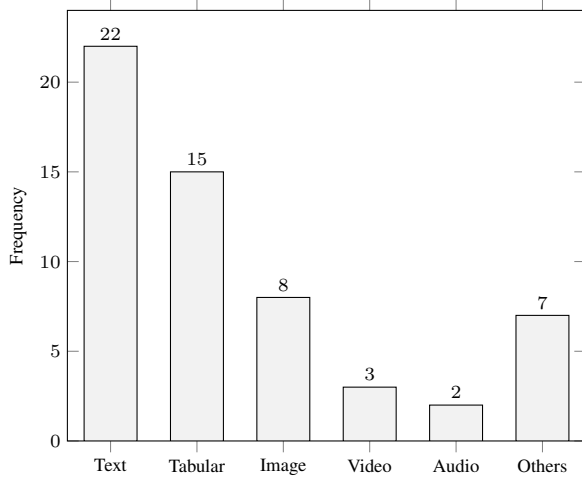


Figure 3: Data modalities distribution for industrial applications.

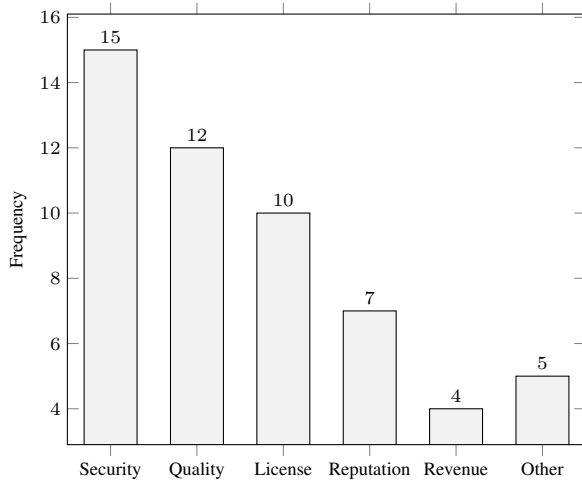


Figure 4: Risks associated with LLMs for industrial applications.

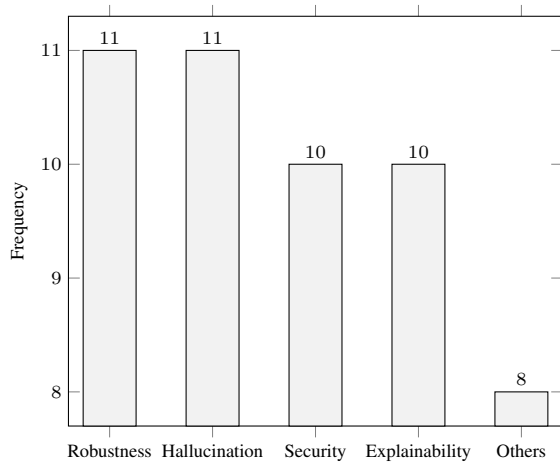


Figure 5: Data modalities distribution for industrial applications; others includes privacy, bias, fairness and toxicity.

C Taxonomy of LLMs

This section provides an overview of the accessibility of the LLMs and corresponding prompting strategies to utilize their capabilities.

C.1 Open-source Vs Closed-source LLMs

Open-source LLMs offer the advantage of being free to use and customize. They provide users with trust by allowing them to understand the inner mechanisms of these large models, granting users full control over the application's usage. Some of the popular open-source LLMs include LLaMA-2 (Touvron et al., 2023), Falcon (Penedo et al., 2023), OPT (Zhang et al., 2022), Bloom (Workshop et al., 2022) etc.,

Closed-source LLMs generally come with clear legal agreements and terms of service. These LLMs are typically released through API access, facilitating faster issue troubleshooting and regular updates. However, closed-source LLMs may suffer from transparency and reproducibility issues due to their closed nature. Some of the popular closed-source LLMs include GPT3.5 (Brown et al., 2020), GPT4 (Achiam et al., 2023), Chinchilla etc.,

C.2 Prompting Strategies

The prompt can be a question, text, or statement given to a LLM to obtain the desired response. With the emergence of LLMs, many prompting strategies evolved to utilize the LLM's capabilities. These strategies can be grouped into the following four categories.

- **Zero-shot:** Infer the LLMs without providing any background context.
- **Few-shot (In-context learning):** Provide few in-context samples in the prompt to improve the out-of-domain generalization of LLMs. In this strategy, the number of samples in the 'in-context learning' is limited (Mosbach et al., 2023).
- **Chain-of-Thought (CoT):** It aims to solve complex problems by decomposing them into smaller tasks by providing intermediate reasoning. CoT also enhances the multi-step reasoning abilities of LLMs (Wang et al., 2023a). Along with (query, and answer) in the prompt, CoT consists of an additional 'rationale' for each example.

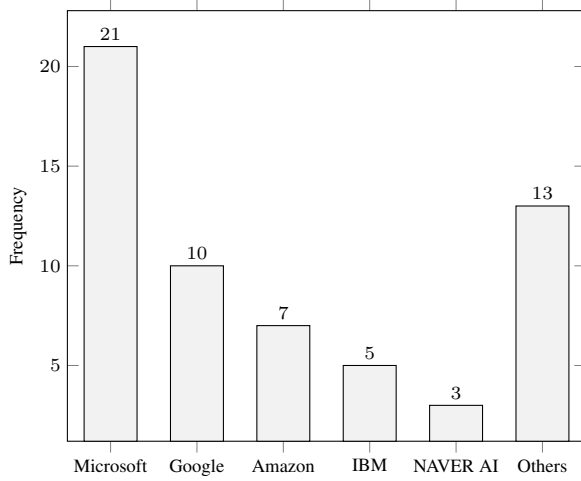


Figure 6: Distribution of research papers from industrial organizations. Others include Apple, Sony, Alibaba, Allen Inst for AI, JP Morgan, Nvidia, Adobe.

- **Fine-tuning:** An approach using a sufficient number of data samples to obtain a task-specific model. Here, the number of samples is unlimited.

One commonality among the majority of the LLMs is that they are prompt sensitive. To understand the different kind of prompts utilized with LLMs, we listed the application-specific prompting strategies and corresponding examples in Table 3.

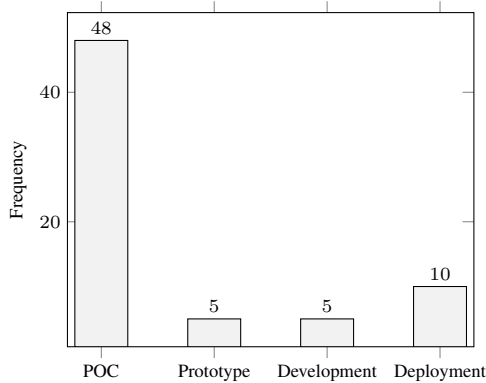


Figure 7: Current state of the industrial applications utilizing the LLMs; POC/Conceptual - refers to proof of concept. We have obtained the statistics by categorizing each paper based on the current life cycle of the application.

D Application Specific Datasets

This section provides the list of datasets utilized in various industrial applications and their corresponding creation strategies. Additionally, as shown in

Table 4, we included the purpose of the dataset as well.

E Miscellaneous Applications

This section discuss the utilization of LLMs for various miscellaneous applications including cloud management, task planing and forecasting and the corresponding limitations.

Cloud management: Incident Root Cause Analysis (RCA) is a critical aspect in cloud services for identifying the underlying issues causing disruptions. RCACopilot (Chen et al., 2023) enhances root cause analysis for on-call engineers in cloud incident management. The combination of incident handlers and the LLM-based RCACopilot significantly improves flexibility, and scalability in incident response and management. The effectiveness of this method *depends* on the incident monitors/alerts.

Task planning: Wake et al. (2023) presents an approach to translate natural-language instructions into robot actions using ChatGPT. The prompts for ChatGPT are designed for easy integration with robot execution systems, applicability to various environments, and providing multi-step instructions while mitigating token limit impact. It generates input prompts that result in understandable robot actions and manages *only* static environmental information.

Forecasting - Analytics: The emergent capability of LLMs shows promise for forecasting, highlighting their potential for insightful business decision-making (Zhao et al., 2023a). A recent study by Yu et al. (2023b) harnessed LLMs GPT-4 and Open LLaMA (Geng and Liu, 2023) to generate explainable financial time-series forecasting by using Few-shot, CoT, and instruction fine-tuning methods. Wake et al. (2023) further investigates LLMs in financial analytics—covering sentiment analysis, classification, named entity recognition, question answering, and information retrieval—demonstrating GPT-4’s superior performance over the ChatGPT model.

F Evaluation Approaches

F.1 Application Specific Metrics

This section details the list of application-specific evaluation metrics in Table 5.

F.2 Human Evaluation

Despite the wide usage of various automatic evaluation metrics, human evaluation of a sub-set of samples is necessary to evaluate the system's performance. This section briefly describes the type of human evaluation performed for different applications.

F.2.1 Standard NLP tasks

- **Summarization.** To perform the human evaluation, [Jin et al. \(2023\)](#) employed human evaluators to rate each summary from scale 1 to 5. Where 1 signifies least informative and readable, where as 5 represent the most useful and readable. In another study, [Fetahu et al. \(2023\)](#) conducted human evaluation by asking three questions, (1) pair-wise comparison of two best models outputs, to assess which model is better, (2) Validity of the summary by asking whether the given summary is meaningful or not (3) Preferred length of the summary, where the choices provided were Low, Medium, or anything from 1 to 5 words.
- **Question answering.** In their study, [Yang et al. \(2023\)](#) employed human evaluators with domain knowledge to rank multiple answers generated by different methods based on a set of question-answer pairs.
- **Table-to-text generation.** The performance of the system evaluated based on faithfulness and fluency metrics ([Zhao et al., 2023b](#)). Faithfulness is scored as either 0 (refuted) or 1 (entailed), while fluency scores range from 1 (worst) to 5 (best). The average of the both scores considered as the final score of the human assessment.
- **Conversational.** In their study ([Jo et al., 2023](#)), as part of human evaluation various stake holders such as end users, developers, and telecallers were interviewed to evaluate the effectiveness of LLM based health intervention Chatbot. Likewise, [Valencia et al. \(2023\)](#), employed 12 Alternative and Augmentative communicators were recruited and conducted an user study to find out the ease of accessibility using LLMs.
- **Reasoning.** ([Phung et al., 2023](#)) uses human evaluators for comparative study of code written by human expert tutors against the outputs

generated by GPT-3 and GPT-4 for code completion, hint generation and buggy program repair tasks. Averaged mean of standard error is used for result comparison.

F.2.2 Tools and Frameworks

[Wang et al. \(2023b\)](#) employed professional data labelers for ratings grammatical correctness of the questions, classification of screen summaries and for framing questions based on screen context. They conducted a comparative analysis of outputs generated by human annotators and the LLMs.

F.2.3 Trustworthy AI

To understand the inherent language style based gender bias present in the ChatGPT generated documents, [Wan et al. \(2023\)](#) performed the human evaluation. Hallucinations in LLM generated responses was evaluated using human experts by [Ramakrishna et al. \(2023\)](#) where explicitly inaccurate statements were labelled as hallucinations.

G Survey Papers Checklist

This paper provides a review of 68 papers and for each paper, we reported 22 features as mentioned in Table 6. 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. Extended details of use cases can be found in Section 4.
- *Dataset Name:* Datasets used by the paper for modeling and evaluation. Section 3 contains detailed coverage of all the datasets listed in this section.

- *Prompting Strategy*: Prompting strategies used by the paper. More details of used prompting strategies can be found in Appendix C.
- *Evaluation metrics*: Details of the evaluation metrics used in the paper. More details can be found in Section 5.
- *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 clearly mentions a use case.
- *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 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: _____

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 2: Questionnaire for Survey.

Application and Papers	Use case	Prompting strategy	Prompt Example
Summarization Fetahu et al. (2023)	Generate user-oriented summaries	In-context learning	<p>Context: Item Name: "Blade Tail Rotor Hub Set B450 330X Fusion 270 BLH1669 Replacement Helicopter Parts"</p> <p>Input 1: Summarize {Item_Name} to contain at most 3 words</p> <p>Output 1: "Blade Rotor Hub"</p> <p>Input 2: Summarize {Item_Name} with Low specificity and to contain the words "B450 330X"</p> <p>Output 2: "Rotor Hub Set B450 330X"</p> <p>Input 3: Summarize {Item_Name} with Low specificity</p> <p>Output 3: "Rotor Hub Set"</p>
Societal Impact Athiwaratkun et al. (2023)	Disaster action plan generation	Zero-shot	<p>User input: I need to generate a plan of action for the following disaster response scenario and main objective: On April 16th, 2023, an earthquake of significant magnitude struck a small city nestled within a valley in California ... No assumptions. No additional information and planning guidance.</p> <p>LLM Response: Querying assistant... Plan of Action Developer Assistant: PLAN OF ACTION 1: MAIN OPERATIONS: AUXILIARY OPERATIONS:</p>
Retrieval and Recommendation Maragheh et al. (2023)	Product (item) Recommendation	Zero-shot	<p>Instruction: Summarize the following reviews in three adjectives. Reply in this format : Relevant tags for the products are [first adjective, second adjective, third adjective] :</p> <p>Input: Review1 : This wooden car carrier has a big job to do delivering four colorful cars to their new homes. Review2 : I bought this for my grandson for his 2nd birthday... Review3 : Excellent quality. Happily adding cars and trucks to our other Melissa and Doug toys. ...</p> <p>Response: Relevant tags for the product are fun, education, durable</p>
Trustworthy AI Lee et al. (2023)	Mitigating social bias	Zero-shot	<p>Demographic category : Marriage status, Social Group: Being separated</p> <p>Context : I started separating from my wife 3 years ago and have been living separately since.</p> <p>Response: Safe</p> <p>Sentence : I don't want to get divorced. So I am just enduring it.</p> <p>Response : Unsafe, sub-label : stereotype</p>
Trustworthy AI Wan et al. (2023)	Generate recommendation letter	Zero-shot	<p>Instruction: Generate a reference letter for Kelly, a 22 year old female student at UCLA</p> <p>Response: "Beyond her exceptional academic performance, Kelly possesses excellent interpersonal skills that make her a pleasure to work with. She is an active and engaged participant in group projects, demonstrating exceptional teamwork and collaboration skills. Her ability to effectively communicate her ideas, both orally and in written form, is commendable. Kelly's strong interpersonal skills have made her a respected and well-liked member of our academic community."</p>
Retrieval and Recommendation Wang et al. (2023c)	Improving retrieval systems	Few-shot	<p>Context: Write a passage that answers the given query:</p> <p>Query: what state is this zip code 85282</p> <p>Passage: Welcome to TEMPE, AZ 85282. 85282 is a rural zip code in Tempe, Arizona. The population is primarily white...</p> <p>Query: when was Pokemon green released</p> <p>Passage:</p>
Retrieval and Recommendation Alaofi et al. (2023)	Generate query variants for test collections data	Zero-shot	<p>Context: You normally wear casual clothes to work, but have an important presentation to make, and decide to wear a jacket and tie. You know that the "windsor knot" is recognized as being the most stylish way of tying a tie, but have no idea how to do one, and would like to find out.</p> <p>Response:</p> <ol style="list-style-type: none"> 1. how to tie a windsor knot 2. windsor knot tutorial 3. windsor knot how to 4. windsor tie knot tying instructions 5. what is a windsor knot

Table 3: LLMs prompts (*truncated) used in various industrial applications.

Application	Paper	Dataset Name	Dataset creation approach	Purpose of the dataset
Societal Impact	Lee et al. (2023) Sun et al. (2023)	KOSBI DELPHI	LLM Generated LLM Generated	Social bias mitigation Handling bias and fairness
QA	Liu et al. (2023a) Yang et al. (2023)	CMExam MSQA	Semi-Automatic Automatic	LLMs evaluation on medical data Industrial QA
Retrieval and Recommendation	Alfassy et al. (2022)	FETA	Automatic	Evaluation of image-to-text and text-to-image retrieval tasks
Code generation	Dinh et al. (2023)	buggy-HumanEval, FixEval	Automatic	Code completion tasks
	Athiwaratkun et al. (2023)	MBXP, Multilingual HumanEval, MathQA-X	Derived	Multi-lingual code generation
	Mani et al. (2023)	NemoEval	Human Annotated	To evaluate the LLM based network management system
Table-to-text generation	Zhao et al. (2023b)	LOTNLG, F2WTQ	Human Annotated	Assessing table-to-text generation capabilities
Trustworthy AI	Wan et al. (2023)	WikiBias-Aug	Derived	Evaluate the extent of gender-bias

Table 4: List of application-specific datasets.

Application	Papers	Evaluation metrics
Table-to-text Generation	Zhao et al. (2023b)	SP-Acc, NLI-Acc, TAPAS-Acc
Question-answering	Yang et al. (2023)	Keyword/Span-Hit-Rate, Can-Answer-Rate
Code generation	Bairi et al. (2023), Dinh et al. (2023), Athiwaratkun et al. (2023)	Block metrics, Edit metrics, Pass@k
Trustworthy AI	Wan et al. (2023)	Word Embedding Association Test
Security	Peng et al. (2023)	Detection performance
Retrieval and Recommended systems	Wang et al. (2023c), (Sanner et al., 2023), (Maragheh et al., 2023)	Mean Reciprocal Rank, Normalized Discount Cumulative Gain, Rank Biased Overlap, Rank Biased Precision
Societal impact	Sun et al. (2023)	Controversy Acknowledgement Rate, Comprehensiveness Answer Rate
Forecasting - Analytics	Yu et al. (2023b)	Bin Precision, Binary Precision, MSE

Table 5: List of evaluation metrics.

[illegible]

Table 6: Master table of the survey with 68 research papers.