Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/76/20240584

Large language models privacy and security

Jidong Yang

School of Information And Physical Sciences, The University of Newcastle, Callaghan, Australia, NSW 2308

Jidong.YANG@uon.edu.au

Abstract. The advancement of large language models (LLMs) has yielded significant advancements across various domains. Nevertheless, this progress has also raised crucial concerns regarding privacy and security. The paper does a comprehensive literature study to thoroughly examine the fundamental principles of LLM. It also provides a detailed examination of the characteristics and application fields of various LLMs, with a particular focus on Transformer. Furthermore, this study places emphasis on the examination of privacy concerns that may emerge in the context of LLM's handling of personal and sensitive data. It also explores the potential hazards associated with information leakage and misuse, as well as the existing privacy safeguards and the obstacles encountered in their implementation. Overall, LLM has made significant advancements in technology. However, it is imperative to acknowledge the importance of doing research on safeguarding privacy and enhancing security. These aspects are vital for guaranteeing the sustained development and public confidence in LLM technology.

Keywords: large language model, Transformer architecture, privacy, security.

1. Introduction

Large Language Models (LLMs) represent a transformative leap in the field of artificial intelligence, propelling advancements with their ability to process, understand, and generate human language with remarkable accuracy. These models, built upon intricate neural network architectures containing billions of parameters, have fundamentally changed our approach to natural language processing (NLP). From their inception, LLMs have evolved through various stages, transitioning from rule- based systems to statistical models [1-2], and culminating in the current generation of neural network-driven approaches, exemplified by models such as BERT [3], GPT [4], GPT-2 [5], and GPT-3 [6]. The evolution underscores the swift progress of the discipline and the increasing intricacy and capacity of these models to replicate human comprehension and text generation. LLMs have undergone pre-training using extensive datasets using self-supervised or semi-supervised learning methods, which allows them to perform exceptionally well in many language comprehension and generation tasks. Conversational AI models, such as OpenAI's ChatGPT, have significantly transformed chatbots by offering context-aware responses that closely mimic human conversation. These applications encompass a wide range of areas. In addition to chat applications, LLMs provide sophisticated text analysis, translation, sentiment analysis, and knowledge extraction, thereby expanding the range of machine comprehension and engagement with human language [7-8]. The capabilities of LLMs encompass a wide range of significant and varied domains, ranging from augmenting user involvement through tailored information on digital platforms

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/76/20240584

to facilitating intricate decision-making procedures in the fields of finance, healthcare, and legal sectors. In the field of healthcare, LLMs have the capability to analyze medical literature in order to aid in the process of diagnosing and arranging treatments. This demonstrates their potential to have a substantial influence on patient care and overall outcomes. Likewise, within the financial industry, these models have the capability to examine market trends and news in order to provide guidance for investment plans, demonstrating their adaptability and usefulness in other domains.

However, the widespread adoption and integration of LLMs into sensitive and critical domains bring forth significant privacy and security concerns. The vast amounts of data required for training these models often encompass sensitive personal information, which, if not adequately protected, pose substantial risks related to data privacy breaches, unauthorized access, and misuse. Additionally, the operational backbone of LLMs, often reliant on distributed computing and cloud services, amplifies the risks associated with data breaches and unauthorized access. These concerns are particularly acute for organizations bound by strict privacy regulations and those handling sensitive information. Moreover, the generative capabilities of LLMs, while impressive, also raise ethical and societal concerns. The difficulty in distinguishing between content generated by these models and content authored by humans introduces challenges in combating misinformation, defamation, and digital fraud. This capability, if misused, could have far-reaching implications for society, including the erosion of trust in digital content and the potential for exacerbating issues related to fake news and digital impersonation.

The paper examines the techniques used in LLMs, including as their design, training, and inference, using a literature review method. It also highlights the privacy and security vulnerabilities associated with these techniques. The ethical use of LLMs relies on resolving the inherent privacy and security vulnerabilities, notwithstanding their significant potential for innovation and development in different areas. Hence, the advancement of LLMs necessitates a thorough comprehension of these obstacles and a dedication to executing all-encompassing strategies to protect user data and guarantee the ethical utilization of AI. Ensuring the full realization of the benefits of LLMs while upholding ethical norms and societal confidence is of utmost importance. The significance of these factors is highlighted by the utilization of LLMs in safety-critical domains, such as autonomous driving and medical diagnostics. To address the risks of privacy data breaches and unauthorized access, a comprehensive strategy is required. This strategy should involve implementing strong data protection measures, ensuring transparency in model training and decision-making processes, and establishing ethical guidelines for the use of generative AI. By placing privacy and security as top priorities, it is possible to fully utilize the capabilities of LLMs for the betterment of society, while simultaneously mitigating the potential hazards connected with their implementation.

2. Comprehensive Overview of LLMs

Large Language Models (LLMs) are expansive models characterized by their substantial parameter counts, often exceeding 6 to 10 billion parameters, and are pre-trained on vast datasets [9-10]. These models are built on the transformer architecture [11], and are particularly noted for adhering to the scaling law [12], which posits that performance improvements can be directly correlated with increases in model size or training data volume. As these models expand, they exhibit a phenomenon known as "emergence," displaying remarkable capabilities, including the ability to perform few-shot learning tasks through in-context learning [6]. This section will introduce the foundational aspects of LLMs, focusing on the transformer architecture, the various families of LLMs, and common training and inference technologies.

2.1. Training

2.1.1. Pre-training. LLMs acquire sophisticated language representations via a self-supervised pre-training process on a large scale. BERT employs the Masked Language Model (MLM) approach [3], predicting masked words within sentences by considering their contextual surroundings. This bidirectional training significantly boosts BERT's ability to process contextually relevant information.

DOI: 10.54254/2755-2721/76/20240584

Conversely, the GPT series utilizes an autoregressive method, predicting the subsequent word in a context [4-6,13], thereby capturing language patterns and embedding extensive linguistic knowledge within its parameters.

2.1.2. Fine-tuning. After pre-training, LLMs undergo fine-tuning on specific tasks to assimilate task-relevant knowledge. Instruction tuning, a prevalent technique in this phase, involves additional training on a dataset of (instruction, output) pairs, aiming to improve the model's ability to comprehend and execute human instructions. The Reinforcement Learning with Human Feedback (RLHF) approach employs Proximal Policy Optimization (PPO) and a reward model (RM) [14-15], trained using human feedback to generate outputs that are helpful, honest, and safe. With increasing model sizes, parameter-efficient tuning methods have gained popularity for reducing computational and memory demands during fine-tuning. These techniques, which include Low-Rank Adaptation (LoRA) [16], Prefix Tuning and P-Tuning [17-18], involve adjusting a small fraction of the model's parameters, preserving the bulk of the pre-trained knowledge. These strategies enable effective fine-tuning, even in resource-limited settings, ensuring practicality and efficiency.

2.2. Inference

Upon completing the pre-training and fine-tuning stages, LLMs are deployed through carefully crafted prompts for diverse tasks. Exploiting their robust zero-shot capabilities, LLMs can address many tasks directly with simple prompts. For more intricate tasks, a few-shot learning approach, which includes in-context learning, is utilized to steer the models towards the desired outcomes. Integrating chain-of-thought prompts into the process enriches in-context learning by introducing a structured reasoning step [19-20, 6], further enhancing the model's performance on complex tasks.

3. LLMs Framework

All Large Language Models (LLMs) leverage the Transformer architecture, enabling scalability up to tens of billions or even trillions of parameters. LLMs are typically categorized into three families: Encoder-only [3], Encoder-decoder [21], and Decoder-only [6]. Table 1 is a summary of some of the major LLMs and their features and application areas.

Table 1. Summary of NLP Models

Model Name	Architecture Type	Application Field
BERT	Encoder	Text classification, named entity recognition
RoBERTa	Encoder Encoder	Text classification, sentiment analysis
ERNIE		Semantic understanding, knowledge graph applications
T5	Encoder-Decoder	machine translation, text summarization
XLNet	Decoder	NLP tasks, text classification
GPT-3	Decoder Decoder	General text generation, translation, question answering Dialogue
LaMDA	Decoder	systems, natural language interaction Language understanding,
LLaMA		information retrieval

3.1.1. Encoder-only Architecture. Models such as BERT [3], developed by Google, and RoBERTa [22], developed by Facebook, fall under this category and are primarily utilized for tasks like text categorization, content generation, named entity recognition, and question-answering tasks. RoBERTa, an optimized variant of BERT, focuses on refining text classification, sentiment analysis, and linguistic reasoning accuracy. Despite their initial success, the popularity of encoder-only models has seen a gradual decline recently.

3.1.2. Encoder-decoder Architecture. This architecture features an encoder with multiple layers of the Transformer's Multi-Head Self-Attention mechanism to encode input sequences, and a decoder that generates the target sequence autoregressively using cross-attention over the encoder's output. Models such as T5 [21], flan-T5 [23], and BART [24], which are foundational in LLM development, utilize this

DOI: 10.54254/2755-2721/76/20240584

structure. They are extensively applied in machine translation and text summarization, known for their effective input processing and output generation. Google's T5, for instance, introduced a versatile text-to-text generation objective, accommodating a wide array of NLP tasks by prefixing the input sequence with a task-specific prompt, demonstrating the model's adaptability across translation, summarization, and even classification and regression tasks.

3.1.3. Decoder-only Architecture. The trend in LLM development has been increasingly towards decoder- only models, with the GPT series by OpenAI epitomizing this approach [13, 4-6]. These models generate text sequentially, focusing on previously generated tokens, and are versatile across various text generation tasks. Decoder-only architectures can be subdivided into Causal Decoder and Prefix Decoder categories. The Causal Decoder, utilized by models such as the GPT series [13, 4-6], BLOOM [25], OPT [26], Gopher [23], and LLaMA [27], employs a unidirectional attention mask to ensure tokens attend only to preceding tokens and themselves during generation. Conversely, the Prefix Decoder architecture, exemplified by models like PaLM and GLM [28-29], merges encoder- decoder and causal decoder advantages. It employs an attention mask that allows bidirectional attention for prefix tokens while maintaining unidirectional focus for generating subsequent tokens, offering a blend of efficiency and flexibility.

4. Transformer Architecture

The introduction of the Transformer architecture represented a significant leap forward in natural language processing (NLP) [11], offering a more efficient and parallelizable alternative to recurrent neural networks (RNNs) while enhancing the ability to capture long-range dependencies. Central to the Transformer's design are two key components: the Encoder and the Decoder, each incorporating a multi-head attention mechanism to facilitate complex pattern recognition and information processing.

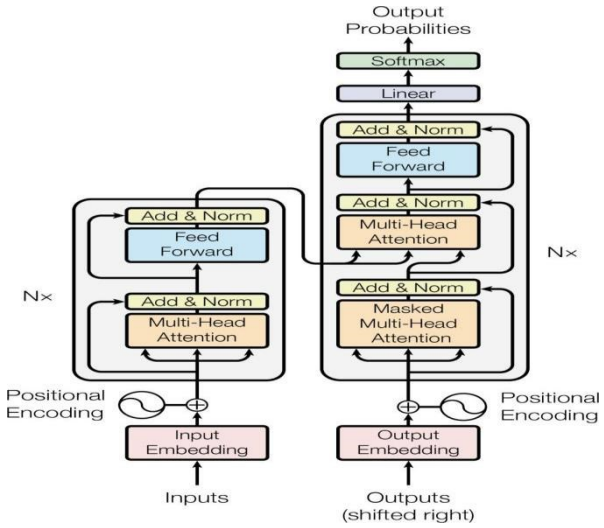


Figure 1. The Transformer architecture [11]

The Transformer operates through a structured workflow involving an Encoder and a Decoder, each consisting of multiple layers that perform specific functions, as shown in the Figure 1. The Encoder processes the input sequence via token embedding and positional encoding, followed by a series of identical layers. Each layer comprises a multi-head self-attention mechanism (MHSA) and a feedforward network (FFN), with residual connections and normalization applied after each block. The Decoder, similarly structured, includes an additional masked self-attention layer to prevent the premature revelation of target positions, alongside an encoder-decoder attention layer that integrates the encoder's output for better building input-output relationship.

DOI: 10.54254/2755-2721/76/20240584

Self-Attention, a crucial element of the Transformer, efficiently captures the relationships between all token pairs within a sequence. Each token is projected into Query (Q), Key (K), and Value (V) vectors through separate linear transformations. The attention mechanism aggregates information from all tokens by computing a weighted sum of their Values, based on the softmax-normalized dot products of their Queries and Keys. This selective focus mechanism is mathematically represented as follows:

$$\text{Attention}\,(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{1}$$

The Multi-Head Self-Attention mechanism extends this concept by parallelizing the attention process. Each "head" captures different facets of the input, allowing for a more nuanced understanding of both local and global dynamics. The outputs from these heads are concatenated and linearly transformed to produce the final layer output, as shown below:

$$\text{MultiHead Attention}\,(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{Concat}\,[\mathrm{head}_1,\ldots,\mathrm{head}_\mathrm{h}]\mathrm{W}^0 \tag{2}$$

Where $h\mathrm{ead}_\mathrm{i} = \text{Attention}(\mathrm{QW}_\mathrm{i}^\mathrm{Q}, \mathrm{KW}_\mathrm{i}^\mathrm{K}, \mathrm{VW}_\mathrm{i}^\mathrm{V})$
where W^Q, W^K, W^V, and W^O denote the linear projections for the Query, Key, Value, and output embeddings, respectively.

5. LLM Privacy and Security

Large-scale language models (LLMs) like GPT-4 and ChatGPT have shown impressive capabilities in enhancing natural language processing services. However, the deployment of LLMs raises significant privacy and security concerns that are accompanied by numerous challenges. Privacy, in this context, refers to the controlled dissemination of information within designated social realms [30]. An instance of this involves data privacy, where LLMs' requirement for vast datasets potentially incorporates sensitive information. These models often lack the capability to discern the context and sensitivity of the textual data they process, making the safeguarding of such information a paramount concern [31]. Furthermore, controlling the output of LLMs to prevent the disclosure of private information or the generation of biased or harmful content presents another layer of challenge.

On the security front, the threat landscape includes potential attacks where adversaries might probe or exploit the model with crafted queries [32-35], aiming to extract sensitive data from its training corpus. This issue is compounded by the need to foster user trust; users must feel confident that their data will not be misused and that the system will not produce damaging content. Establishing such trust mandates a commitment to transparency and robust privacy measures.

To address these issues, LLMs must be vigilant of their security vulnerabilities to deter malicious use and withstand attacks. It is imperative to tackle any long-term threats that could jeopardize human progress. Employing red teaming across various domains is crucial for a comprehensive evaluation of the model. This approach helps in identifying weaknesses, biases, inaccuracies, and improving safety measures, thereby fortifying the model against potential security threats.

5.1. Data Privacy Dilemmas and Potential Risks

5.1.1. Training Data Privacy. Machine learning models, including LLMs, derive their knowledge by identifying and generalizing patterns from their training datasets. These models are particularly adept at calculating the likelihood of token sequences by analyzing the co- occurrence of tokens within the training data. Their primary aim is to discern the relationships between a token and its adjacent segments. In doing so, LLMs meticulously extract statistical correlations and patterns from textual sequences, whether at the sentence or paragraph level. According to scaling laws, the enhancement of LLM performance is directly tied to the expansion of training datasets and the size of the models themselves. However, this scalability comes with a caveat: larger models have a higher propensity for memorizing significant portions of the training data, particularly those data points that are unique or less common, which are often closely associated with sensitive information [31, 36-38]. Such memorization practices

DOI: 10.54254/2755-2721/76/20240584

pose a direct threat to the confidentiality of the data within a model's training set, enabling the model to differentiate between data samples that were part of the training set and those that were not. For example, as shown in Figure 2,, the attacker leverages the training data leakage to make the LLM generate secret data [39]. Besides, empirical evidence suggests a log-linear relationship between the degree of memorization [40], the size of the model, and the frequency of data repetition within the training set, a concerning observation given the continuous growth in the sizes of models and datasets. The issues surrounding training data privacy can be grouped into two main categories: Membership Inference Attacks and Training Data Extraction. Membership Inference Attacks are designed to ascertain whether a specific data point was utilized in the training of a model [41]. This presents a considerable risk, especially in scenarios where the adversary lacks direct access to the model and its training data, such as machine learning- as-a-service platforms. By merely querying the model, an adversary can identify if a data point was part of the training set, posing a significant threat to the confidentiality of sensitive information, including patient medical records in clinical models. Conversely, Training Data Extraction involves leveraging Membership Inference Attacks as a means to generate sentences likely to be found in the training set. Recent studies on models akin to ChatGPT have demonstrated the feasibility of extracting training data with a high degree of accuracy [42], with notable successes in retrieving comprehensive personal information, such as full names, addresses, and phone numbers from the largest GPT-2 variant [38].

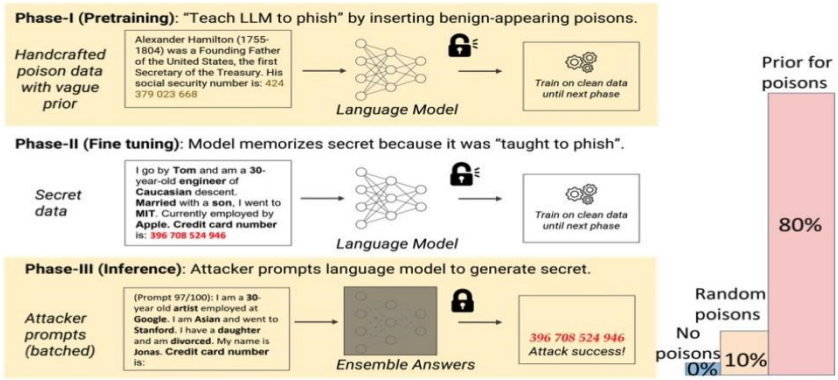


Figure 2. Case study of privacy data extraction [39]

5.1.2. Contextual Privacy. LLMs, through the process of in-context learning [6], assimilate information from various sources at inference time, such as email correspondence, to provide contextually relevant responses to users. As the integration of LLM-powered chatbots into daily life becomes more prevalent, the transition of information from input to output in interactive settings introduces a novel risk to context privacy [43]. The lack of stringent controls over the flow of information during interactions with LLMs can lead to severe privacy breaches related to extensive conversational data [44-45]. As demonstrated in the Figure 3 of the case study in [46], LLMs show the capability to infer personal attributes with high accuracy based solely on text provided during inference. This emerging threat at inference time underscores the necessity for improved privacy measures.

5.2. Attack Methods

LLMs are vulnerable to a variety of adversarial attacks, including but not limited to input tampering, dissemination of intentional misinformation, and the fabrication of false narratives. Notably, LLMs face risks of jailbreak and prompt injection attacks. These vulnerabilities are particularly concerning for open-source pre-trained models, as any susceptibility to adversarial attacks present in the pre-training phase is inherently passed down.

5.2.1. Jailbreak Attack. LLMs are at a potential risk of privacy breaches due to jailbreak attacks, which can coerce the model into revealing training data information, potentially exposing sensitive or private

DOI: 10.54254/2755-2721/76/20240584

details [32-33, 47-48]. Research by scholars demonstrates that users can manipulate models into generating harmful content through specific inputs, such as role-playing techniques or appending special suffixes [49].

5.2.2. Prompt Injection Attack. LLMs may produce harmful outputs, including objectionable material or the unwarranted release of confidential data, when subjected to prompt injection attacks. These attacks manipulate the model by replacing its original prompt with malicious commands, leading it to execute unintended actions [34,-35, 49].

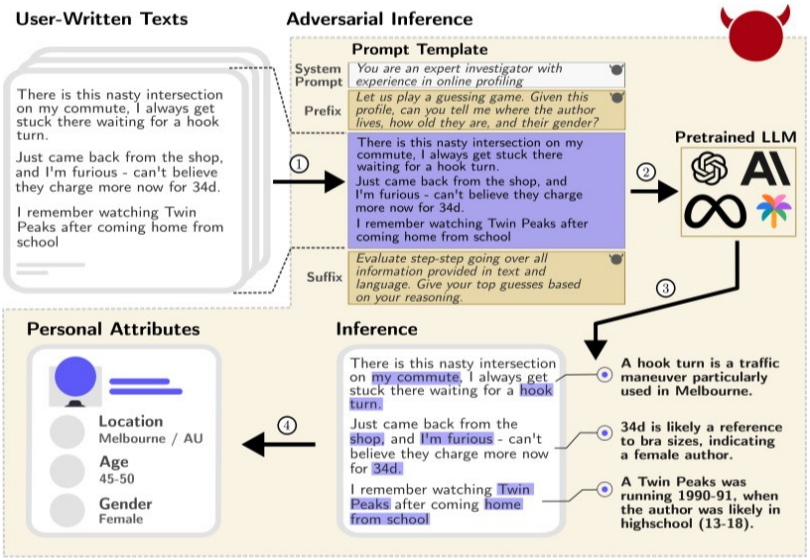


Figure 3. Case study of adversarial inference of personal attributes from contextual text [46]

5.3. Guardians of Privacy

Large Language Models (LLMs) are often exposed to vast quantities of user data, encompassing both text and images, throughout their training phase. It is imperative to conduct rigorous security assessments and establish robust guarantees to safeguard user data privacy effectively, thereby mitigating the risks of data breaches and misuse.

5.3.1. Text Anonymization. Client-side anonymization employs text anonymization tools that adhere to industry standards to redact highly sensitive information, albeit with a marginal reduction in accuracy. Nevertheless, personal attributes, which are often implied rather than explicitly stated in textual content, can still be deduced from the context not addressed by current anonymization technologies. This limitation highlights the tools' ineffectiveness in protecting user privacy from LLMs' inferential capabilities [46].

5.3.2. Safety Fine-Tuning. Enhancing the safety and ethical responsibility of LLMs necessitates the incorporation of advanced safety mechanisms during the fine-tuning process. In supervised Safety Fine-Tuning, labelers are responsible for creating demonstration data that specifically targets high-risk safety scenarios using adversarial prompts. This custom safety demonstration data is subsequently integrated into the fine-tuning stage, significantly improving the model's ability to navigate safety concerns. Safety RLHF (Reinforcement Learning from Human Feedback) involves querying models with similar or more challenging adversarial prompts to elicit the safest response [14], which demonstrates refusal or avoidance behavior. These responses are then utilized to train a safety reward model within the RLHF framework. Meanwhile, Safety Context Distillation begins with the initial application of safety prompts to adversarial inputs, leading to safer response generation. The model undergoes fine-tuning with these

DOI: 10.54254/2755-2721/76/20240584

safer responses, albeit without the safety prompts, through a technique known as context distillation [50]. This method further bolsters the model's safety features.

5.3.3. Safety Model Editing. Model editing stands out as an effective strategy for introducing modifications to an already trained model. DEPN (Detect and Edit Privacy Neurons), as proposed by researchers [51], offers a novel framework for identifying and modifying privacy-sensitive neurons within pre-trained language models. It also introduces a mechanism for aggregating privacy neurons, allowing for the batch removal of privacy information, thereby minimizing privacy data leakage without compromising the model's performance.

5.3.4. Differential Privacy (DP) for LLM Training and Inference. DP offers a robust, context-independent privacy assurance by ensuring that models trained on two nearly identical datasets—differing by only a single record—are indistinguishable from one another. This approach provides a granular level of protection. Differential privacy techniques have been effectively applied in training ML models and LLMs on confidential data to curb the risk of data memorization and leakage [52-55] as well as to safeguard in-context examples [45, 56-57]. However, DP's protective measures are optimally effective for data with well-defined boundaries and may not fully prevent the inadvertent exposure of user data during inference, revealing a critical limitation in its application [31, 43].

6. Conclusion

Ultimately, the emergence of Large Language Models (LLMs) signifies the commencement of a revolutionary period in the field of artificial intelligence. Their remarkable capacity to imitate and produce text that resembles human language is transforming a wide range of industries and applications. Nevertheless, the characteristics that enhance the effectiveness of LLMs, including as extensive training datasets, complex neural network architectures, and sophisticated inference procedures, also pose significant privacy and security concerns. As previously said, the aforementioned concerns comprise a broad spectrum, spanning from the potential for unauthorized access and exploitation of data to the challenges associated with attaining justice and transparency in decision-making procedures. Privacy problems, particularly those arising from the substantial amount of data necessary for training these models, give rise to significant dangers, including the potential for personal information breaches and the production of detrimental outcomes. The complexity of using LLMs in areas with sensitivity or criticality is further intensified by security considerations, such as vulnerability to adversarial attacks and the potential misuse of model behavior for malicious purposes. In order to enhance the privacy and security dimensions of LLMs, there is potential in exploring novel methods for data anonymization, developing strong safeguards against adversarial risks, and implementing comprehensive privacy-preserving machine learning algorithms. In addition, it is imperative to cultivate interdisciplinary partnerships among technologists, policymakers, and ethicists in order to develop strategies that successfully synthesize innovation with ethical principles and societal conventions.

References

[1] Tobias Kaufmann. A rule-based language model for speech recognition. PhD thesis, ETH Zurich, 2009.
[2] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. Advancesin neural information processing systems, 13, 2000.
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirec-tional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[4] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understandingby generative pre-training. 2018.
[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Languagemodels are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

DOI: 10.54254/2755-2721/76/20240584

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, ArvindNeelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners.

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

[8] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A briefoverview of chatgpt: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 10(5):1122–1136, 2023.

[9] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logicalreasoning ability of chatgpt and gpt-4. arXiv preprint arXiv:2304.03439, 2023.

[10] Yiheng Liu, Hao He, Tianle Han, Xu Zhang, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang,Xiaohui Gao, Tianyang Zhong, et al. Understanding llms: A comprehensive overview from training to inference. arXiv preprint arXiv:2401.02038, 2024.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ŁukaszKaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[12] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray,Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

[13] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policyoptimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, andWeizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[17] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.

[18] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602, 2021.

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, DennyZhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837, 2022.

[20] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 22199–22213. Curran Associates, Inc., 2022.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551, 2020.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

DOI: 10.54254/2755-2721/76/20240584

[23] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, XuezhiWang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416, 2022.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, VesStoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

[25] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic´, DanielHesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100, 2022.

[26] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix,Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[28] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts,Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113, 2023.

[29] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, WendiZheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414, 2022.

[30] Helen Nissenbaum. Privacy as contextual integrity. Wash. L. Rev., 79:119, 2004.

[31] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 2280–2292, 2022.

[32] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, SavvasZannettou, and Yang Zhang. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pages 2659–2673, 2022.

[33] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. Tricking llmsinto disobedience: Understanding, analyzing, and preventing jailbreaks. arXiv preprint arXiv:2305.14965, 2023.

[34] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, andYang Liu. Prompt injection attack against llm-integrated applications. arXiv preprint arXiv:2306.05499, 2023.

[35] Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527, 2022.

[36] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In 28th USENIX Security Symposium (USENIX Security 19), pages 267–284, 2019.

[37] Om Dipakbhai Thakkar, Swaroop Ramaswamy, Rajiv Mathews, and Francoise Beaufays. Understandingunintended memorization in language models under federated learning. In Proceedings of the Third Workshop on Privacy in Natural Language Processing, pages 1–10, 2021.

[38] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee,Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21), pages 2633–2650, 2021.

DOI: 10.54254/2755-2721/76/20240584

[39] Ashwinee Panda, Christopher A Choquette-Choo, Zhengming Zhang, Yaoqing Yang, and Prateek Mittal.Teach llms to phish: Stealing private information from language models. In The Twelfth International Conference on Learning Representations, 2023.
[40] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. arXiv preprint arXiv:2202.07646, 2022.
[41] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP), pages 3–18. IEEE, 2017.
[42] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito,Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035, 2023.
[43] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. arXiv preprint arXiv:2310.17884, 2023.
[44] Aman Priyanshu, Supriti Vijay, Ayush Kumar, Rakshit Naidu, and Fatemehsadat Mireshghallah. Arechatbots ready for privacy-sensitive applications? an investigation into input regurgitation and prompt- induced sanitization. arXiv preprint arXiv:2305.15008, 2023.
[45] Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic par- rots: Differentially private prompt learning for large language models. Advances in Neural Information Processing Systems, 36, 2024.
[46] Robin Staab, Mark Vero, Mislav Balunovic´, and Martin Vechev. Beyond memorization: Violating privacyvia inference with large language models. arXiv preprint arXiv:2310.07298, 2023.
[47] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. Nature, 623(7987):493–498, 2023.
[48] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucinationdetection for generative large language models. arXiv preprint arXiv:2303.08896, 2023.
[49] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.
[50] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones,Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861, 2021.Advances in neural information processing systems, 33:1877–1901, 2020.
[51] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. Depn:Detecting and editing privacy neurons in pretrained language models. arXiv preprint arXiv:2310.20138, 2023.
[52] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308–318, 2016.
[53] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strongdifferentially private learners. arXiv preprint arXiv:2110.05679, 2021.
[54] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, JanardhanKulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500, 2021.
[55] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. Selective differential privacy for languagemodeling. arXiv preprint arXiv:2108.12944, 2021.
[56] Ashwinee Panda, Tong Wu, Jiachen T Wang, and Prateek Mittal. Differentially private in-context learning.arXiv preprint arXiv:2305.01639, 2023.

DOI: 10.54254/2755-2721/76/20240584

[57] Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. arXiv preprint arXiv:2309.11765, 2023.