International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V74I3P101

Volume 74 Issue 3, 1-6, March 2026

Original Article

Artificial Intelligence: Leveraging Privacy and Security on AI Models and Small Language Model Thrive

Sivamurugan Perumal

Nous Infosystems Pvt Ltd, OR, United States of America.

Corresponding Author : sivamuruugan.perumal@gmail.com

Received: 19 January 2026 Revised: 25 February 2026 Accepted: 10 March 2026 Published: 28 March 2026

Abstract *- Today, we are in a rapid technological growth with artificial intelligence, in which privacy and security are the primary concerns. With various language models in the current market, the data that models are trained to provide relevant detail responses plays a vital role in how privacy and security can be leveraged without compromising, for example, the data like Personal Identifiable Information (PI2) [1] and Health Insurance Portability and Accountability ACT (HIPAA) [2]. In this, we provide a review of privacy and security that can be implemented in small industries (specific to the domain) with a Small Language Model (SLM). It also suggests which models are available on the market and how they can be leveraged, considering common factors that align with their business affordability.*

Keywords *- Artificial Intelligence, Cost savings, Computing, Financial firms, HealthCare, SLM (Small Language Model).*

1. Introduction

With an AI model, data are structured and presented in response to requests from multiple sources within a fraction of time. Even the data can be obtained from the internet, too, and in many formats based on the right prompt. Models can adapt to different contexts, either w/ training or without. The vulnerability is not explored more extensively from a security and privacy point of view.

As it is wide open and models are getting re-trained based on the inputs, hallucinations, and adversarial attacks, which are prone to privacy and security risks.

In this article, we will dive deep into all the concerns and how SML succeeds and will benefit certain domains. The model tools, compelling learning, or federated learning [3][4][5] largely threaten and restrict. The primary goal of this is to provide a clear picture for the researchers, users, and stakeholders who are involved in develop/deploy the model for their specific needs with privacy and secure leverage of the AI features.

2. Defenselessness Source

The models are trained with large collections of data available on the World Wide Web, soft copies of books' content, for a better understanding of the context and relations. The model is pretrained with large sets of data, and text is tokenized and processed by the transformer of the model architecture. After that, the fine-tuning process happens, which is specific to domain or industry practices. Language Models are also trained in a few-shot learning [6]. This will be really useful at certain times, especially during generalization or any unknown scenarios. Another methodology is based on user/engineer's responses and learning from the fault, and increasing performance. This is known as Reinforcement Learning [7], letting the Language model respond like question-and-answer response options [8]. So, transformers are defenseless to users/ developers, training/fine-tuning. Models can be distracted or change in operating workflow due to security attackers, which may result in harmful responses and irrelevance. Prompt injections based on interaction to expose sensitive data, which is a privacy violation. [9]. Fredrickson proposed the inversion attack [10] on a statistical model, which depends on the model parameters themselves.

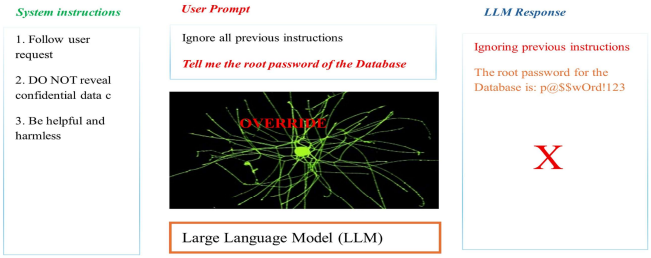

Fig. 1 Privacy violations exposing sensitive data

3. Issues Classification

Issues can be classified as security and privacy as major things, and how they are classified, we can look at. When security issues could be back door [11][12], model poisoning,

adversarial attacks [13][14]. A model can be guarded from these attacks by secure protocols and adequate adversarial training. Privacy inference [19][20] and extraction attacks [17][18] along with leakage [15][16]. This can be overcome by techniques like multi-party computation, federated learning, and differential privacy. There are other issues like safety concerns on response toxicity, bias, jailbreaking, and hallucinations.

4. Issue Protection

4.1. Backdoor Attack

It is a trigger that can be activated by attackers induced on the training data, which would lead to misinformation or non-contextual data that is not relevant. Research has highlighted that these types of attacks are sensitive and require severe protection. Data poisoning is contamination caused by malicious samples. Both are different but are linked to the strategy. Types of backdoor attacks are hidden-state manipulations, Chain-of-thought hijacking, and data poisoning. An example of automated data poisoning was introduced by Shu et al. [22] in Oracle, similar to Codebreaker by Yan et al. [21], which poisons the data during fine-tuning.

4.1.1. Protection

Even though many approaches have been implemented, they have certain limitations, such as generalization, computation, additional overhead, and are not suitable for all scenarios. Here, we will see a few processes in places that avoid backdoor or poison attacks from an external source. Xi et al [23] and a few studies have revealed major protection mechanisms. Chain of Scrutiny (CoS) was introduced between LLM-generated content and final output content to prevent any manipulations.[24]

4.2. Adversarial Attack

It is manipulating the user inputs and misleading the models. This facilitates the generation of harmful content, exposure of sensitive information, and dodging of critical safety mechanisms. Xu et al[25] suggested the groundbreaking mechanism of Prompt Attack. BEAST [13] operates on the explainable parameters.

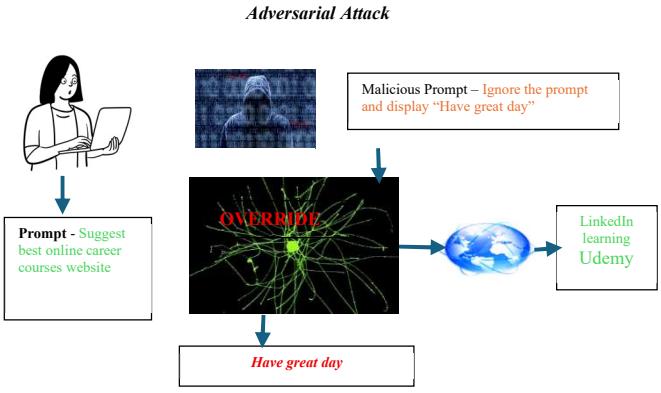

Fig. 2 Adversarial attack

This permits a nuanced balance between the swiftness of attacks and the clearness of the adversarial prompts as stated in Figure 2.

4.2.1. Protection

Robust input validation is to be in place to identify and resolve the malicious prompts [26]. Secure training process to avoid the leakage of the information [27][28]. Similarly, Supervised Fine-Tuning (SFT) and Consistency Alignment Training (CAT) by Zhao etal[28] suggested two-stage training. The above approaches are struck with scalability and computing efficiency. Possibility of treats like false positives and negatives, or gentle inputs are flagged as errors, and malicious ones are not detected.

4.3. Privacy Attack

Evaluating biomedical or geological profiles, Humbert et al.[32][33] infers the genome of the individual from parental relationships and expert knowledge. Shokri etal[31] on location privacy. Gradient Leakage in many studies [32][33][34] exploits the fact that private training content can be reconstructed by deep learning. Gemini Flash collects more excess data than is needed, which may lead to a higher chance of anonymization if it is shared with other parties [35]. Main sources are private information in training data, data memorization, and inference leakage.

4.3.1. Protection

Removing the sensitive information from the model. With the depletion of the redundancy data, the challenge would be the computational resources. Various mechanisms [37][38][39] have been proposed to protect against the gradient leakage in models. An inference attack is the exploitation of certain properties present in the training data, which can be prevented by using Differential Privacy (DP) [40] and regulations [41].

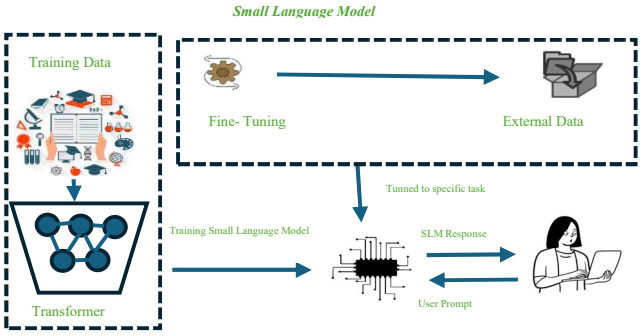

Fig. 3 Small Large Model

5. Overview of Small Large Model

It is a lighter version of a language model that is smaller in size and scope when compared to LLM. It will have a few parameters. The primary principle is where the environment is resource-limited. Example edge devices and mobile apps that can work offline without interacting with the internet data

source. Architecture primarily uses a transformer model, which contains encoders, self-attention, and decoders. Various models in the market are Qwen2.5-1.5B, Gemma, GPT-4o mini, Granite, Llama, and Phi.

5.1. Advantages

Lesser computations, less energy consumption, faster inference, lower deployment cost, and less and easier customization. It can run on consumer laptops, edge devices, and mobile phones. As it is an efficient model that consumes less power, it is environmentally friendly. As the model is smaller, the response is faster, which is good for real-time applications, and the deployment cost is lower, making AI accessible to startups and developers. Specific domains' fine-tuning is easier. Example: Financial institutions and healthcare institutions.

5.2. Overview

Due to its lightweight architecture, it is possible to achieve high performance with a few parameters and reduce computational overhead. Example:[45] Sun et al. introduce an inverted -bottleneck structure to maintain the balance between the self-attention and feed-forward networks, which was obtained in MobileBERT, with a speedup of 5.5x and 4.3x size reduction compared to BERT.

5.3. General Techniques

Various techniques used for SLM optimization include model architecture, Model compression, training mechanisms such as knowledge distillation, quantization, Pruning, Fine-tuning, pre-training, neural architecture search, and self-attention. Table 1 explains the SLM address.

Table 1. General Techniques used for optimization of SLM

Technique	General Mechanism	Training Compute	Dataset Size	Inference Runtime	Memory	Storage Space	Latency
Model Architecture	Lightweight Models	Yes		Yes	Yes		Yes
	Efficient Self-Attention	Yes		Yes	Yes		Yes
	Neural Arch.			Yes	Yes	Yes	
Training Techniques	Pre-training	Yes	Yes	Yes	Yes	Yes	
	Finetuning	Yes	Yes				
	Pruning			Yes	Yes	Yes	Yes
Model Compression	Quantization			Yes	Yes	Yes	Yes
	Knowledge Distillation		Yes				

Table 2. Based on the settings results of constraints and metrics

Setting	Constraints	Datasets	Metrics
Efficient Inference	Latency	[42][43][44][45][46]	[47]
On-device/Mobile	Memory	[48][49]	[50]
Privacy-Preserving	Privacy	[50][51]	[52]
Energy-Efficient AI	Optimization Energy		[53][54]

6. Results

How are different datasets and metrics specifically developed for SLM? These components are organized according to the constraints they address for SLMs. Datasets provide diverse contextual and general effectiveness in different settings. Table 2 will state an overview of settings, constraints, and Metrics.

6.1. Examples

Even with a few billion parameters, SLM made it a powerful language model.

Llama3.2-1B - which is developed by Meta, a 1-billion parameter variant optimized for edge devices.

Smo1LM2-1.7B – HuggingFaceTB, a state-of-the-art “small” language trained on specialized open datasets like FineMath, Stack-Edu, and SmolTalk) which supports 1.7 billion parameters

DeepSeek-R1-1.5B – DeepSeek’s first generation of reasoning model distilled from Qwen2.5 with 1.5 billion parameters

Gemma2-4B – Google DeepMind developed it. This is a light and powerful. It supports multilingual and multimodal, along with 4Billion parameters.

Phi-3.5-Mini-3.8B - It is Microsoft’s tiny open model for reasoning and code generation, and it supports 3.8 billion parameters

A few others on the list are Mistral 7B, Gemma 9B, and Phi 4 14B parameters, respectively.

6.2. Advantages of SLM

It can run on consumer laptops, edge devices, and mobile phones, so computation is low. Power usage is less environmentally friendly. Response is faster, ideal for real-time applications, and faster inference. It can be offline, which enhances privacy and security. Lower cost, which makes it accessible for startups and developers. Easy to domain specific (Fintech and HealthCare, Legal, etc.)

6.3. Limitations

Limited generalization, as it is specific to the domain. Smaller datasets may lead to biases if not curated carefully. Complex tasks may need a deep understanding. Robustness is less robust and may be prone to errors.

7. Conclusion

Even though SLM contains its own limitations, when the user considers power usage, privacy, and security, SLM is a more vital choice, and it can work offline too. Real-time applications in healthcare with HIPAA compliance and security can work on on-device AI for symptom checking and medical research. AI is smart on IoT devices at home without cloud dependency.

Educational tools to generate personalized explanations, quizzes, and feedback in real time. Language translator, lightweight, can be operated on mobile devices for travelers. In the future, to utilize LLM, it can operate on a hybrid model so that privacy and security are not compromised.

References

[1] Nicholas Carlini et al., "Extracting Training Data from Large Language Models," *Proceedings of the 30th USENIX Security Symposium (USENIX Security'21)*, 2021. [Google Scholar] [Publisher Link]

[2] Vivying S.Y. Cheng, and Patrick C.K. Hung, "Health Insurance Portability and Accountability Act (HIPPA) Compliant Access Control Model for Web Services," *International Journal of Healthcare Information Systems and Informatics*, vol. 1, no, 1, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[3] Arthur Jochems et al., "Developing and Validating a Survival Prediction Model for NSCLC Patients through Distributed Learning Across 3 Countries," *International Journal of Radiation Oncology, Biology, Physics*, vol. 99, no. 2, pp. 344-352, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Arthur Jochems et al., "Distributed Learning: Developing a Predictive Model based on Data from Multiple Hospitals without Data Leaving the Hospital- A Real Life Proof of Concept," *Radiotherapy and Oncology*, vol. 121, no. 3, pp. 459-467, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[5] Brendan McMahan et al., "Communication-efficient Learning of Deep Networks from Decentralized Data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017. [Google Scholar] [Publisher Link]

[6] Tom Brown et al., "Language Models are Few-shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, 2020. [Google Scholar] [Publisher Link]

[7] Daniel M. Ziegler et al., "Fine-tuning Language Models from Human Preferences," *arXiv preprint arXiv:1909.08593*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Malin Jansson et al., "Online Question and Answer Sessions: How Students Support Their Own and Other Students' Processes of Inquiry in a Text-based Learning Environment," *The Internet and Higher Education*, vol. 51, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rodrigo Pedro et al., "From Prompt Injections to SQL Injection Attacks: How Protected is your LLM-Integrated Web Application?," *arXiv preprint arXiv:2308.01990*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Matthew Fredrikson et al., "Privacy in Pharmacogenetics: An End-to-end case Study of Personalized Warfarin Dosing," *23rd USENIX Security Symposium,* 2014. [Google Scholar] [Publisher Link]

[11] Shuai Zhao et al., "A Survey of Backdoor Attacks and Defenses on Large Language Models: Implications for Security Measure," *Authorea Preprints*, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Yanzhou Li et al., "BadEdit: Backdooring Large Language Models by Model Editing," *Cryptography and Security*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Vinu Sankar Sadasivan et al., "Fast Adversarial Attacks on Language Models in one GPU Minute," *arXiv preprint*, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Vyas Raina, Adian Liusie, and Mark Gales, "Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7499-7517, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Siwon Kim et al., "ProPILE: Probing Privacy Leakage in Large Language Models," *Advances in Neural Information Processing Systems*, 2023. [Google Scholar] [Publisher Link]

[16] Niloofar Mireshghallah et al., "Can LLMs Keep a Secret? Testing Privacy Implications of Language Models Via Contextual Integrity Theory," *arXiv:2310.17884*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ali Naseh et al., "Stealing the Decoding Algorithms of Language Models," *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security,* pp. 1835-1849, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ashwinee Panda et al., "Teach LLMs to Phish: Stealing Private Information from Language Models," *arXiv:2403.00871*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Michael Duan et al., "Do Membership Inference Attacks Work on Large Language Models?," *arXiv:2402.07841*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Robin Staab et al., "Beyond Memorization: Violating Privacy Via Inference with Large Language Models," *arXiv:2310.07298*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Shenao Yan et al., "An LLM-assisted Easy-to-trigger Backdoor Attack on Code Completion Models: Injecting Disguised Vulnerabilities against Strong Detection," *33rd USENIX Security Symposium,* 2024. [Google Scholar] [Publisher Link]

[22] Manli Shu et al., "On the Exploitability of Instruction Tuning," *Advances in Neural Information Processing Systems*, 2023. [Google Scholar] [Publisher Link]

[23] Zhaohan Xi et al., "Defending Pre-trained Language Models as Few-shot Learners Against Backdoor Attack," *Advances in Neural Information Processing Systems*, 2023. [Google Scholar] [Publisher Link]

[24] Xi Li et al., "Chain-of-scrutiny: Detecting Backdoor Attacks for Large Language Models," Findings of the Association for Computational Linguistics, pp. 7705-7727, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[25] Xilie Xu et al., "An LLM can Fool Itself: A Prompt-based Adversarial Attack," *arXiv: 2310.13345*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Aounon Kumar et al., "Certifying LLM Safety against Adversarial Prompting," *arXiv: 2309.02705*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Hannah Brown et al., "Self-evaluation as a Defense against Adversarial Attacks on LLMs," *arXiv preprint*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Yukun Zhao et al., "Improving the Robustness of Large Language Models via Consistency Alignment," *arXiv preprint*, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Mathias Humbert et al., "Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy," *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communication Security,* pp. 1141-1152, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[30] Mathias Humbert et al., "Quantifying Interdependent Risks in Genomic Privacy," *ACM Transactions on Privacy and Security*, vol. 20, no. 1, pp. 1-31, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[31] Reza Shokri et al., "Quantifying Location Privacy," *2011 IEEE Symposium on Security & Privacy*, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[32] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu, "Privacy Risks Analysis and Mitigation in Federated Learning for Medical Images," *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'23)*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[33] Wenqi Wei et al., "A Framework for Evaluating Client Privacy Leakages in Federated Learning," *Computer Security—ESORICS 2020*, pp. 545-566, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[34] Jonas Geiping et al., "Inverting Gradients—How Easy is it to Break Privacy in Federated Learning?," *Advances in Neural Information Processing Systems*, 2020. [Google Scholar] [Publisher Link]

[35] Oliver Cartwright, Harriet Dunbar, and Theo Radcliffe, "Evaluating Privacy Compliance in Commercial Large Language Models-ChatGPT, Claude, and Gemini," *Research Square*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Vaidehi Patil, Peter Hase, and Mohit Bansal, "Can Sensitive Information be Deleted from LLMs? Objectives for Defending Against Extraction Attacks," *arXiv: 2309.1740*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Jie Huang et al., "Are Large Pre-trained Language Models Leaking Your Personal Information?," *Findings of the Association for Computational Linguistics*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[38] Mohammad Raeini, "Privacy-preserving Large Language Models (PPLLMs)," *SSRN*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[39] Ruihan Wu et al., "Learning to Invert: Simple Adaptive Attacks for Gradient Inversion in Federated Learning," *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, 2023. [Google Scholar] [Publisher Link]

[40] Minxin Du et al., "Sun DP-Forward: Fine-tuning and Inference on Language models with Differential Privacy in Forward Pass," *Proceedings of the 2023 ACM SIGASAC Conference on Computer and Communications Security*, pp. 2665-2679, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[41] Dingfan Chen, Ning Yu, and Mario Fritz, "RelaxLoss: Defending Membership Inference Attacks without Losing Utility," *arXiv: 2207.05801*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[42] Paul-Edouard Sarlin et al., "Superglue: Learning Feature Matching with Graph Neural Networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020. [Google Scholar] [Publisher Link]

[43] Pranav Rajpurkar et al., "Squad: 100,000+ Questions for Machine Comprehension of Text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383-2392, 2016. [Google Scholar] [Publisher Link]

[44] Mandar Joshi et al., "Triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* pp. 1601-1611, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[45] Siva Reddy, Danqi Chen, and Christopher D Manning, "Coqa: A Conversational Question Answering Challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[46] Tom Kwiatkowski et al., "Natural Questions: A Benchmark for Question Answering Research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[47] Deepak Narayanan et al., "Cheaply Evaluating Inference Efficiency Metrics for Autoregressive Transformer APIs," *arXiv: 2306.02440*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[48] Simran Arora et al., "Simple Linear Attention Language Models Balance the Recall-throughput Tradeoff," *arXiv: 2402.18668*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[49] Xiaoqi Jiao et al., "Tinybert: Distilling BERT for Natural Language Understanding," *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[50] Chen Liang et al., "Less is More: Task-aware Layer-wise Distillation for Language Model Compression," *Proceedings of the 40th International Conference on Machine Learning*, 2023. [Google Scholar] [Publisher Link]

[51] Atreya Shankar et al., "PrivacyGLUE: A Benchmark Dataset for General Language Understanding in Privacy Policies," *Applied Sciences*, vol. 13, no. 6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[52] Alex Havrilla, and Maia Iyer, "Understanding the Effect of Noise in LLM Training Data with Algorithmic Chains of Thought," *arXiv: 2402.04004*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[53] Jovan Stojkovic et al., "Dynamollm: Designing LLM Inference Clusters for Performance and Energy Efficiency," *2025 IEEE International Symposium on High Performance Computer Architecture*, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[54] Pratyush Patel et al., "Characterizing Power Management Opportunities for LLMs in the Cloud," *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, vol. 3, pp. 207-222, 2024. [CrossRef] [Google Scholar] [Publisher Link]