Cybersecurity Challenges and Mitigations for LLMs in DoD Applications

Corinne Yorkman and Mark Reith
Air Force Institute of Technology, WPAFB, USA

corinne.yorkman.1@us.af.mil
mark.reith.3@us.af.mil

Abstract: Great power competition has escalated globally, making it increasingly important for the Department of Defense (DoD) to adopt artificial intelligence (AI) technologies that are advanced and secure. Large language models (LLMs), which generate text, code, images, and other digital content based on data sets used in training have gained attention for their potential in DoD applications such as data analysis, intelligence processing, and communication. However, due to the complex architecture and extensive data dependency of LLMs, integrating LLMs into defense operations presents unique cybersecurity challenges. These risks, if not properly managed, could pose severe threats to national security and mission integrity. This survey paper categorizes these challenges into vulnerability-centric risks, such as data leakage, and misinformation, and threat-centric risks, including prompt manipulation and data poisoning, providing a comprehensive framework for understanding the potential risks of LLMs in DoD settings. Each category is reviewed to identify the primary risks, current mitigation strategies, and potential gaps, ultimately identifying where further research is needed. By summarizing the state of the art in LLM cybersecurity, this paper offers a foundational understanding of LLM security within the DoD. By advocating for a dual approach that considers both the evolving nature of cyber threats and the operational needs of the DoD, it aims to provide actionable recommendations to guide ongoing research in the integration of LLMs to DoD operations.

Keywords: Large language models, Cybersecurity challenges, Department of defense

1. Introduction

The integration of artificial intelligence (AI) into critical operations has transformed numerous sectors, and the Department of Defense (DoD) is no exception. Large language models (LLMs) represent significant advancements in natural language processing (NLP). These models have the potential to revolutionize decision-making processes, intelligence analysis, and communication strategies within the DoD (Caballero & Jenkins, 2024). However, with these opportunities come substantial cybersecurity risks, as the DoD operates in environments with high stakes for national security and sensitive data protection. The expansive capabilities of LLMs also introduce novel vulnerabilities and threats that demand rigorous examination and tailored mitigation strategies (Department of Defense, 2023). While the field of cybersecurity has seen a growing body of research, the application of this work to DoD use cases involves additional layers of complexity. The purpose of this paper is to survey the cybersecurity challenges associated with LLMs in DoD applications, evaluate current mitigation strategies, and propose a research agenda for addressing gaps in existing solutions. By compiling existing research and bridging gaps in previous work, this paper aims to serve as a reference for policymakers, researchers, and practitioners seeking to secure LLMs in defense settings.

1.1 Framework

A key aspect of this survey is the categorization of risks into two primary frameworks: vulnerability-centric risks and threat-centric risks. Vulnerability-centric risks address the systemic weaknesses within LLMs, such as data leakage, unintended biases, and misinformation (Ganguli et al., 2023). In contrast, threat-centric risks pertain to external adversarial threats, including prompt manipulation and data poisoning. Threat centric and vulnerability centric approaches are among the most common perspectives to address risk (Silva & Jacob, 2018). This dual framework is rooted in established practices within cybersecurity and risk management, which emphasize the interplay between mitigating internal weaknesses and addressing external threats. Further, it follows the NIST Cybersecurity Framework, which identifies vulnerabilities and threats as key risk factors. The classification also adheres to the ISO/IEC 27005, which lays out six key components of the risk management process and identifies vulnerability and threat analysis within its risk identification component (ISO & IEC, 2022). This categorization not only reflects existing risk assessment methodologies but also aligns with the DoD's strategic emphasis on resilience and proactive threat mitigation in cyberspace operations (Air Force Doctrine Publication 3-12, 2011).

2. Background

LLMs are "a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks"

(IBM, 2024). These capabilities include text summarization, question answering, language translation, and more. They are built on deep learning architectures and trained on vast datasets which enable them to perform complex NLP tasks.

Recent advancements in LLMs have pushed the boundaries of what these systems can achieve and highly diversified their applications. The introduction of transformers has revolutionized LLM architectures by enabling models to understand context through mechanisms like self-attention, significantly enhancing their performance (Vaswani et al., 2017). Advancements in model fine-tuning and transfer learning also allow LLMs to specialize in specific tasks without requiring training from scratch, reducing resource demands while improving efficiency (Bommasani et al., 2021). These improvements have made LLMs indispensable in domains requiring rapid and reliable text processing.

2.1 Role of LLMs in DoD Operations

The DoD has recognized the transformative potential of LLMs for enhancing operational efficiency and decision-making processes. As stated by the Deputy Secretary of Defense (2023), "the DoD faces an imperative to explore the use of this technology and the potential of these models' scale, speed, and interactive capabilities to improve the Department's mission effectiveness while simultaneously identifying proper protection measures and mitigating a variety of related risks." As an effort to advance safe and responsible AI technology within the USAF, the Air Force Research Laboratory developed NIPRGPT, an LLM cleared for installation on unclassified USAF systems (Secretary of the Air Force Public Affairs, 2024). Additional LLM use cases include those by the Air Mobility Command which has leveraged LLMs to generate campaign simulations and those by the US Air Forces Central which use LLMs to expedite routine maintenance of software tools (Caballero & Jenkins, 2024). LLMs have the potential to transform processes like information, planning, and decision-making and aid in military exercises. They can be used to synthesize intelligence reports, summarize classified documents, and assist in threat detection by identifying patterns in textual data (Caballero & Jenkins, 2024). At the Bravo 11 Hackathon, the US Pacific Air Forces team displayed the potential of LLM deployment in intelligence processing and report summarization (Leidos, 2024). These capabilities can enhance real-time operational decision-making and can be applied in adaptive mission planning. Recent work by Weller et al. (2024) also claims that LLMs can be a useful tool for enhanced wargaming in multi-domain operations by efficiently generating realistic and real-time scenario narratives. Additionally, by assisting the human operator in collecting, analyzing, and assessing data, LLMs could aid Intelligence, Surveillance, Target Acquisition, and Reconnaissance (ISTAR) processes (Chochtoulas, 2024). Finally, LLMs may allow for a reduction in commanders' bureaucratic footprints and a reallocation of personnel from administrative to operational roles (Caballero & Jenkins, 2024). Their scalability and adaptability make them ideal for high-demand, data-intensive environments like national defense operations. However, as described in the AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense (2019), it is still the early days for AI models, and the DoD treats these models as a virtual assistant, not a replacement for humans.

3. Vulnerability-Centric Risks for LLMs

LLMs are imperative in DoD applications, but, as noted by Yao et al. (2023), the very architecture that makes LLMs so powerful also introduces vulnerabilities that adversaries can exploit. This section defines specific vulnerability-centric risks and evaluates current mitigation strategies, focusing on data leakage, unintended bias, misinformation, and misalignment with DoD policies.

3.1 Data Leakage and Confidentiality

In the context of LLMs, data leakage occurs when LLMs inadvertently retain and disclose sensitive information encountered during training. Carlini et al. (2021) explain, language models can memorize verbatim text from their training data and reproduce it in response to specific prompts. The DoD manages highly sensitive information which contains classified research, military secrets, national security intelligence, and personal data from service members. While it is expected that any LLM trained on classified data would itself be subject to classification and associated access controls, the inclusion of such data in training still presents unique risks. These risks include the potential for unintended disclosure if access controls are improperly implemented or if the model is later used outside of its intended context. Even in non-adversarial contexts, unintended disclosures could erode trust in the use of LLMs for defense applications. The potential consequences range from reputational damage to the exposure of critical infrastructure vulnerabilities.

Dataset pre-processing is a fundamental approach where training datasets are thoroughly processed to remove sensitive or identifiable information. This step ensures that the model is less likely to memorize data that could

be inadvertently exposed during use (Kucharavy et al., 2024). In some cases, the sensitive information is critical to an LLMs functionality, and including it is necessary. Aligning with the DoD's Zero Trust cybersecurity model (Department of Defense, 2022), controlled access mechanisms can be applied at the infrastructure and application levels to manage how LLMs interact with sensitive data sources. Cryptographic methods, which encrypt data during LLM training, could be implemented into DoD use cases, as they are ideal for collaborative training and applications requiring strict confidentiality (Song and Zhao, 2024). A more novel approach to this risk includes research by Li et al. (2023) who present a model which utilizes differential privacy. Differential privacy, "gives a formal guarantee that individual-level information about participants in the database is not leaked," (*Differential Privacy*, n.d.) by adding random noise to query results, making it difficult to determine whether any individual's data is included or excluded. Despite expansive research and advancements, as of the time this paper was written, no provably secure design principles exist that completely prevent private information memorization and leakage in LLMs. As a result, continuous monitoring of model outputs using anomaly detection systems are crucial to identifying and flagging potential data leaks in real time.

3.2 Unintended Bias

LLMs are trained on massive datasets which include inherent biases (Ferrara, 2023). These biases can arise in outputs when models reflect and amplify the patterns of discrimination or underrepresentation embedded in their datasets (Bommasani et al. 2021). For the DoD, biased outputs from LLMs could result in discriminatory targeting, inappropriate mission prioritization, or inadequate resource allocation (Chochtoulas, 2024). Inaccurate or biased intelligence that overlooks demographic nuances can erode decision-maker trust in LLM-generated insights, diminishing their utility in high-stakes contexts.

Data balancing is a cornerstone strategy to address bias in LLMs and involves the careful curation of training datasets to ensure they are representative of diverse perspectives and demographics. For DoD applications, this entails sourcing data that reflects the complexities of global geopolitics, cultural contexts, and operational requirements. Red teaming, which identifies vulnerabilities in AI systems by crafting adversarial prompts and evaluating responses, is an effective strategy to detect biases. Recent studies by Perez et al. (2022) have proposed that LLMs themselves can be used to automate red teaming. Bias detection algorithms which flag potentially problematic outputs could also be implemented in LLM models.

3.3 Misinformation and AI Hallucination

Misinformation generated by LLMs can stem from inaccuracies in training data, inconsistent data, or the models' inherent limitations in reasoning and factual correctness. Regarding inconsistent data, the model may receive contradicting data whose validity requires specific context (e.g. "the sky is blue" and "the sky is grey"). LLMs are prone to generating plausible-sounding yet factually incorrect information due to their inability to differentiate truth from fabrication (Banerjee, 2024). This phenomenon is commonly referred to as AI hallucination, and it highlights the model's tendency to produce content that appears coherent while lacking factual grounding. The high degree of confidence with which LLMs generate misinformation further exacerbates the problem, as users may accept false information without verification (Abdali et al., 2024). Another subset of misinformation is misalignment, where outputs diverge from the values or operational requirements of the deploying organization (Yi et al., 2024). It occurs when LLM outputs, though they may be factually correct, fail to adhere to organizational guidelines, ethical standards, or regulatory frameworks. The dissemination of misinformation by LLMs poses significant risks to DoD operations, where decision-making depends on accurate and reliable data.

Combating misinformation and AI hallucination requires enhancing both the accuracy of LLMs and the processes for validating their outputs. Periodic model fine-tuning with curated datasets that emphasize DoD-approved terminologies, operational guidelines, and ethical considerations ensures that outputs remain consistent with policy standards (Florida International University et al., 2024). Human-in-the-loop (HITL) systems are indispensable in this context, enabling operators to review and verify LLM outputs before deployment in critical scenarios. Retrieval-Augmented Generation (RAG) can serve as a powerful defense by integrating LLMs with external knowledge sources. Instead of relying solely on pre-trained data, RAG retrieves relevant, verified information in real-time to ground responses in factual content, reducing susceptibility to misinformation (Prompting Guide, n.d.). It is also possible to prompt LLMs to express confidence levels, however the downfall of this approach is that it often leads to high error rates (Abdali et al., 2024). Enhanced transparency in LLM design, including features that trace the sources of generated information, can further aid in distinguishing accurate outputs from misinformation. LLMs are continuously adapting, and even with the implementation of effective mitigation practices, regularly evaluating LLM performance against DoD benchmarks to ensure that models remain responsive to evolving guidelines and operational context is essential to mitigate misinformation.

Additionally, robust training in AI literacy for DoD personnel ensures that operators remain vigilant about potential errors, fostering a culture of critical evaluation rather than blind reliance on LLMs.

4. Threat-Centric Risks of LLMs

In addition to internal weaknesses, LLMs are also subject to external adversarial threats. These threat-centric risks represent deliberate efforts by adversaries to manipulate, corrupt, or deceive using LLM capabilities. This section explores key threat-centric risks—prompt manipulation, input attacks, data poisoning, and model corruption—along with the associated risks and mitigation strategies.

4.1 Prompt Manipulation and Adversarial Input Attacks

Prompt manipulation involves an adversary crafting malicious input prompts deliberately designed to coerce LLMs into generating harmful or unintended outputs by exploiting their sensitivity to contextual phrasing (Abdali et al., 2024). Prompt manipulation poses significant risks in DoD contexts by facilitating the generation of disinformation or misinformation, ultimately creating operational confusion or undermining strategic decision-making. Adversaries could use manipulated prompts to generate convincing yet false intelligence reports, operational plans, or propaganda. The inherent sensitivity of LLMs to nuanced prompts makes them particularly vulnerable to these attacks, and the consequences could include operational disruption, reputational damage, and the proliferation of inaccurate information in critical systems (Caballero & Jenkins, 2024).

Mitigating prompt manipulation requires a dual-sided approach that integrates monitoring, detection, and resilience measures. A key defensive strategy is response monitoring, which involves analyzing LLM outputs in real time to detect and flag anomalies indicative of adversarial influence. These systems compare generated content against established patterns of safe and valid responses, triggering alerts when outputs deviate from acceptable norms (Gan et al., 2024). However, current keyword-based filtering techniques have limited efficacy against sophisticated adversarial prompts that strategically bypass flagged keywords (Springer, 2024). To enhance resilience, adversarial training plays a crucial role by exposing LLMs to manipulated prompts during the training phase, enabling them to recognize and resist malicious inputs more effectively (Alvarez and Marsal, 2024). Additionally, fine-tuning models with domain-specific data tailored to DoD applications reduces the likelihood of success for generalized prompt manipulation attacks. Beyond these methods, AI developers are increasingly employing red-teaming, a systematic testing process in which security researchers attempt to jailbreak models in a controlled environment to identify weaknesses before deployment. This proactive strategy helps uncover vulnerabilities in existing guardrails, allowing developers to implement stronger defenses against adversarial inputs (Ji, 2024).

4.2 Data Poisoning and Model Corruption

Data poisoning is where malicious users deliberately inject adversarial samples into training data with the aim of corrupting the learned model (Steinhardt et al., 2017). Poisoned data can bias LLM outputs, degrade their reliability, or even implant backdoors that adversaries exploit later. In one study, data poisoning was used to embed biases into text generation models, resulting in outputs that violated ethical standards and operational requirements (Florida International University et al., 2024). Specific to the DoD, corrupted models could produce compromised intelligence, introduce systemic vulnerabilities, or fail to function under operational conditions. The difficulty of detecting subtle manipulations in large datasets further amplifies this risk, as even minor corruptions can have outsized effects on LLM behavior (Bommasani et al., 2021).

Mitigating data poisoning attacks relies on securing the training process which involves controlling access to training datasets and ensuring that only verified, high-quality data is included. For the DoD, maintaining an isolated and secure training environment for model development reduces exposure to external interference, aligning with its strategic focus on safeguarding data integrity and minimizing external threats (Department of Defense, 2023). Implementing robust verification protocols is equally critical, requiring systematic validation of datasets and model updates before deployment (Gan et al., 2024). These protocols often employ anomaly detection tools to identify unusual patterns or discrepancies in data. Data sanitization techniques, which detect and remove anomalous or suspicious data points prior to training, are essential in preventing the incorporation of malicious inputs (Pal et al., 2024). Additionally, adversarial training, as described in section 4.1, has been shown to improve their robustness against data poisoning attacks (Steinhardt et al., 2024).

5. Discussion of Current Mitigation Strategies

Current mitigation strategies, such as adversarial training, data sanitization, response monitoring, and secure training pipelines, have effectively reduced vulnerabilities by hardening LLMs against adversarial tactics. For instance, adversarial training helps models recognize and resist malicious prompts, while secure training environments mitigate risks of data poisoning (Abdali et al., 2024). Data sanitization has shown promise in preventing sensitive information from being inadvertently embedded in LLM outputs, addressing concerns around data leakage and prompt injection (Steinhardt et al., 2017).

In the defense context, the effectiveness of these strategies is constrained by operational and security requirements. Defense environments require solutions that prioritize confidentiality, real-time responsiveness, and resilience under high-stakes conditions. For example, adversarial training, though effective in principle, often fails to account for the nuanced operational and linguistic complexities of military data, necessitating extensive customization for defense applications (Shah et al., 2020). Similarly, while secure model training protocols protect against external threats, the real-time demands of DoD operations can strain these processes, particularly when integrating classified and unstructured data streams (Department of Defense, 2023).

Current solutions exhibit several gaps when applied in defense contexts. First, existing models struggle with adaptability to the strict data security protocols mandated by the DoD, including handling classified information and ensuring compliance with military-specific operational standards. Additionally, response monitoring systems often lack the speed and accuracy needed to detect threats in real-time during critical missions. Using data balancing to combat unintended bias has resulted in LLMs compromising accuracy to incorporate diverse training data, as noted by Google's Gemini which created images of historical figures with the incorrect race (Robertson, 2024). Lastly, adversarial training and monitoring systems often require substantial computational resources, limiting their feasibility in resource-constrained or remote defense environments (Rajhi & Pissinou, 2023).

Implementing LLM mitigation strategies in DoD environments is fraught with logistical and operational challenges, not to mention political and doctrinal hurdles. The distributed and decentralized nature of military operations complicates the standardization of cybersecurity measures (Defense Science Board, 2019). For example, ensuring consistent security practices across geographically dispersed units requires substantial coordination and resource allocation (Singer & Friedman, 2014). Moreover, real-time response requirements, such as those in active combat zones, demand highly reliable and adaptable mitigation systems that current solutions may not provide.

From a technical perspective, the integration of LLM cybersecurity measures into legacy defense systems is a significant challenge. Many DoD systems operate on outdated architectures that lack the computational capacity to support modern AI applications, let alone advanced mitigation frameworks (Riviera et al., 2024). Scalability is another concern. For example, implementing adversarial training at the scale necessary to cover all potential use cases in DoD operations would require specialized hardware, extensive training data, and ongoing updates to keep pace with evolving threats (Wang et al., 2024). Additionally, the resource-intensive nature of these processes can divert funding from other critical defense priorities, creating a trade-off between immediate cybersecurity needs and long-term operational goals.

Compared to the DoD, civilian-sector approaches to LLM security prioritize cost-efficiency, scalability, and user experience. In contrast, DoD requirements for data confidentiality, real-time performance, and operational resilience necessitate more stringent and specialized approaches. For instance, while civilian models may tolerate occasional inaccuracies in outputs, DoD applications demand near-perfect reliability, particularly in mission-critical contexts (Caballero & Jenkins, 2024). Despite these differences, the DoD can draw valuable lessons from civilian practices. The private sector's emphasis on automation and efficiency offers insights into reducing the resource demands of LLM mitigation strategies. For instance, leveraging automated tools for bias detection and output validation could streamline processes and reduce manual oversight requirements (Florida International University et al., 2024). Additionally, the use of federated learning, a technique that enables decentralized training without sharing raw data, could enhance the DoD's ability to train LLMs on classified datasets while maintaining data security (Florida International University et al., 2024). Overall, a hybrid approach that combines the scalability of civilian solutions with the rigor of military standards offers the most promising path forward.

6. Gaps in Research and Recommended Roadmap for Future Research

While efforts have been made to address the cybersecurity risks associated with LLMs, important research challenges remain. Further investigation is necessary to ensure the safe and effective deployment of LLMs in DoD operations. Research must focus on improving techniques such as privacy-preserving training and dynamic data sanitization to address issues like data leakage and unintended bias. These methods should ensure compliance with DoD data security protocols while maintaining the performance of LLMs. Adaptive defenses, such as continuous adversarial training and real-time threat detection, must evolve to address increasingly sophisticated attacks. Leveraging ML to detect anomalous behaviors during LLM interactions can enhance resilience against prompt manipulation and data poisoning. Research into adaptive security frameworks tailored to defense scenarios could improve the robustness of LLMs under operational conditions (Defense Science Board, 2019). LLMs must align with DoD policies and ethical standards, like the DoD Artificial Intelligence Strategy and Implementation Pathway, to prevent misaligned outputs. Additionally, LLMs must be implemented in accordance with the foundational tenets of Responsible Artificial Intelligence (RAI) governance, warfighter trust, AI production and acquisition lifecycle, requirements validation, responsible AI ecosystem, and AI workforce (DoD, 2022). Developing comprehensive policy compliance frameworks, including automated checks that ensure that LLM-generated content adheres to operational guidelines, is essential (National Defense Authorization Act, 2022). Considering the human component, HITL mechanisms, where DoD personnel oversee and adjust LLM responses in sensitive applications, can reduce risks associated with automated decision-making. Ensuring a balance between human judgment and machine efficiency is critical for operations requiring precision and accountability (Air Force Doctrine Publication 3-12, 2011). Finally, equipping DoD personnel with an understanding of LLM limitations and best practices for secure use is essential. Tailored training programs should emphasize recognizing adversarial outputs, understanding LLM biases, and implementing effective safeguards during deployment (Joint Artificial Intelligence Center, 2021).

7. Conclusion

This paper examined the cybersecurity challenges of LLMs in DoD applications, categorizing and defining vulnerability-centric and threat-centric risks and analyzing mitigation strategies. While current mitigation strategies address many of these issues, gaps remain in scalability, real-time effectiveness, and adaptability to defense-specific needs. Recommendations include enhancing privacy protections, strengthening adversarial defenses, aligning models with DoD policies, and prioritizing HITL mechanisms and personnel training.

Disclaimer: The views expressed are those of the authors and do not reflect the official policy or position of the US Air Force, Department of Defense, or the US Government.

Ethics Declaration: Ethical clearance was not required for this research.

AI Declaration: ChatGPT was to conduct background research on relevant topics and to summarize content and key findings of reference papers before they were thoroughly reviewed. Upon completion of the paper, ChatGPT was used to review its contents, ensuring that the paper adheres to standard English conventions.

References

Abdali, S., Anarfi, R., Barberan, C., & He, J. (2024, March). *Securing large language models: threats, vulnerabilities and responsible practices*. https://arxiv.org/html/2403.12503v1#bib.bib184

AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. (2019). Defense Innovation Board. https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUM ENT.PDF

Air Force Doctrine Publication 3-12. (2011). *Cyberspace Operations: Catalog of Doctrine Topics*. Retrieved from https://www.doctrine.af.mil/Portals/61/documents/AFDP_3-12/3-12-AFDP-CYBERSPACE-OPS.pdf

Alvarez & Marsal. (2024). *Training data and prompt manipulation: How to keep your organization safe against LLM*. Alvarez & Marsal. https://www.alvarezandmarsal.com/insights/training-data-and-prompt-manipulation-how-keep-your-organization-safe-against-llm

Banerjee, S. (2024). LLMs will always hallucinate, and we need to live with this. In *Arxiv*. United We Care. https://arxiv.org/html/2409.05746v1#S3

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Caballero, W. N., & Jenkins, P. R. (2024). *On Large Language Models in National Security Applications*. arXiv. Retrieved from https://arxiv.org/pdf/2407.03453

Carlini, N., et al. (2021). Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*.

Chochtoulas, A. (2024). How Large Language Models are Transforming Modern Warfare - Joint Air Power Competence Centre. *Joint Air Power Competence Centre - NATO's Advocate to Air and Space Power*, *37*. https://www.japcc.org/articles/how-large-language-models-are-transforming-modern-warfare/

Defense Science Board. (2019). *Task Force Report: Cyber Resiliency in Weapon Systems*. U.S. Department of Defense.

Department of Defense. (2022, November). *DoD Zero Trust Strategy*. https://dodcio.defense.gov/Portals/0/Documents/Library/DoD-ZTStrategy.pdf

Department of Defense. (2023). *Department of Defense Data, Analytics, and Artificial Intelligence Adoption Strategy: Accelerating Decision Advantage*. Retrieved from https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_ST RATEGY.PDF

Deputy Secretary of Defense. (2023). *Establishment of Chief Digital and Artificial Intelligence Officer Generative AI and Large Language Models Task Force, Task Force Lima* (OSD006491-23). U.S. Department of Defense.

Differential privacy. (n.d.). Harvard University Privacy Tools Project. https://privacytools.seas.harvard.edu/differential-privacy

DoD Responsible AI Working Council. (2022). U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway. In *https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artifici al-Intelligence-Strategy-and-Implementation-Pathway.PDF*. Department of Defense.

Ferrara, E. (2023). *Should ChatGPT be biased? Challenges and risks of bias in large language models*. arXiv. https://arxiv.org/abs/2304.03738

Florida International University, Amini, M. H., & Wu, Y. (2024). Security and privacy challenges of large language models: A survey. *arXiv preprint*.

Gan, Y., Yang, Y., Ma, Z., He, P., Zeng, R., Wang, Y., Li, Q., Zhou, C., Li, S., Wang, T., Gao, Y., Wu, Y., & Ji, S. (2024, November 5). *Navigating the Risks: A survey of security, privacy, and ethics threats in LLM-Based agents*. https://arxiv.org/html/2411.09523v1

Ganguli, D., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., ... & Ziegler, D. (2023). *Red Teaming Large Language Models to Assess Risks and Vulnerabilities.* arXiv preprint arXiv:2303.08774

IBM. (2024b, October 23). Large language models. *IBM*. https://www.ibm.com/topics/large-language-models

International Organization for Standardization & International Electrotechnical Commission. (2022). ISO/IEC 27005:2022. In *iso.org*. International Standard.

Ji, J., Goldstein, J. A., Lohn, A. J., & Center for Security and Emerging Technology. (2023). Controlling large language model outputs: a primer. In *Center for Security and Emerging Technology*. https://cset.georgetown.edu/wp-content/uploads/CSET-Controlling-Large-Language-Model-Outputs-A-Pri mer.pdf

Joint Artificial Intelligence Center. (2021). *DoD AI Education Strategy*. U.S. Department of Defense.

Kucharavy, A., Plancherel, O., Mulder, V., et al. (2024). Large language models in cybersecurity: Threats, exposure, and mitigation. Springer.

Leidos. Hackathon produces ai-enabled cjadc2 solutions for the battlefield. Leidos Insights, 2024. URL https://www.leidos.com/insights/hackathon-produces ai-enabled-cjadc2-solutions-battlefield.

Li, Y., Tan, Z., & Liu, Y. (2023, May 10). *Privacy-Preserving prompt tuning for large language model services*. arXiv.org. https://arxiv.org/abs/2305.06212

National Defense Authorization Act. (2022). *Public Law 117-81*. Retrieved from https://www.congress.gov National Institute of Standards and Technology (NIST). (2022). *Artificial Intelligence Risk Management Framework (AI RMF).*

Pal, A., Bhargava, R., Hinsz, K., Esterhuizen, J., & Bhattacharya, S. (2024). *The empirical impact of data sanitization on language models*. arXiv. https://arxiv.org/abs/2411.05978

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). *Red teaming language models with language models*. arXiv. https://arxiv.org/abs/2202.03286

Prompting Guide. (n.d.). *Retrieval-augmented generation (RAG).* Prompting Guide. Retrieved February 20, 2025, from https://www.promptingguide.ai/research/rag

Rajhi, M., & Pissinou, N. (2023). Resource-Limited Localized Adaptive Adversarial Training for Machine Learning Model. *IEEE Intl Conference on Parallel & Distributed Processing*, 1113–1120. https://doi.org/10.1109/ispa-bdcloud-socialcom-sustaincom59178.2023.00178

Riviera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., & Schneider, J. (2024). Escalation risks from language models in military and diplomatic decision-making. *arXiv preprint*. Retrieved from https://arxiv.org/pdf/2401.03408

Robertson, A. (2024, February 21). Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis. *The Verge*. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

Secretary of the Air Force Public Affairs. (2024, June 10). *Department of the Air Force launches NIPRGPT*. U.S. Air Force. https://www.af.mil/News/Article-Display/Article/3800809/department-of-the-air-force-launches-niprgpt/

Shah, R., et al. (2020). The value of adversarial robustness. *Advances in Neural Information Processing Systems, 33*, 3456-3467.

Silva, F. R. L., & Jacob, P. (2018). Mission-Centric risk Assessment to improve cyber situational awareness. *Proceedings of the 17th International Conference on Availability, Reliability and Security*, 1–8. https://doi.org/10.1145/3230833.3233281

Singer, P. W., & Friedman, A. (2014). *Cybersecurity and Cyberwar: What Everyone Needs to Know*. Oxford University Press.

Song, E., & Zhao, G. (2024, November). *Privacy-Preserving Large Language Models: Mechanisms, Applications, and Future Directions*. arXiv. https://arxiv.org/abs/2412.06113

Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified defenses for data poisoning attacks. *Advances in Neural Information Processing Systems, 30*, 3517-3529.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. *"Attention Is All You Need."* arXiv, 12 June 2017, https://arxiv.org/pdf/1706.03762.

Wang, Z., Li, X., Zhu, H., & Xie, C. (2024). Revisiting adversarial training at scale. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24675–24685. https://doi.org/10.1109/cvpr52733.2024.02330

Weller, D., Meltschack, M., & Schwindling, D. (2024). *Leveraging large language models for enhanced wargaming in multi-domain operations*. NATO Science and Technology Organization. https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-MSG-217/MP-MSG-217-16.pdf

Wu, Q., & Xu, Y. (n.d.). *Surveying Attitudinal Alignment between Large Language Models vs. Humans Towards 17 Sustainable Development Goals*. https://arxiv.org/html/2404.13885v1

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Yi, J., 1, Ye, R., 2, Chen, Q., 1, Zhu, B., 3, Chen, S., 2, Lian, D., 1, Sun, G., 1, Xie, X., 3, University of Science and Technology of China, Shanghai Jiao Tong University, & Microsoft. (2024). On the Vulnerability of Safety Alignment in Open-Access LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 9236–9260) [Journal-article]. https://aclanthology.org/2024.findings-acl.549.pdf