7-3-2024

On large language models in national security applications

William N. Caballero
Air Force Institute of Technology

Philip R. Jenkins
Air Force Institute of Technology

On Large Language Models in National Security Applications

William N. Caballero[a], Phillip R. Jenkins[a]

[a]*Department of Operational Sciences, Air Force Institute of Technology, WPAFB, OH 45433*

Abstract

The overwhelming success of GPT-4 in early 2023 highlighted the transformative potential of large language models (LLMs) across various sectors, including national security. This article explores the implications of LLM integration within national security contexts, analyzing their potential to revolutionize information processing, decision-making, and operational efficiency. Whereas LLMs offer substantial benefits, such as automating tasks and enhancing data analysis, they also pose significant risks, including hallucinations, data privacy concerns, and vulnerability to adversarial attacks. Through their coupling with decision-theoretic principles and Bayesian reasoning, LLMs can significantly improve decision-making processes within national security organizations. Namely, LLMs can facilitate the transition from data to actionable decisions, enabling decision-makers to quickly receive and distill available information with less manpower. Current applications within the US Department of Defense and beyond are explored, e.g., the USAF's use of LLMs for wargaming and automatic summarization, that illustrate their potential to streamline operations and support decision-making. However, these applications necessitate rigorous safeguards to ensure accuracy and reliability. The broader implications of LLM integration extend to strategic planning, international relations, and the broader geopolitical landscape, with adversarial nations leveraging LLMs for disinformation and cyber operations, emphasizing the need for robust countermeasures. Despite exhibiting "sparks" of artificial general intelligence, LLMs are best suited for supporting roles rather than leading strategic decisions. Their use in training and wargaming can provide valuable insights and personalized learning experiences for military personnel, thereby improving operational readiness.

Keywords: Large language models, applied statistics, artificial intelligence, national security, strategic planning

1. Introduction

The profound success of GPT-4 in early 2023 [56] demonstrated the revolutionary nature of large language models (LLMs) to a global audience. Their capabilities were immediately recognized as a potentially disruptive force in numerous industries, e.g., healthcare [30], education [38], and finance [37]. However, as with other artificial intelligence (AI) models, LLMs are not limited in scope to civilian applications; their capabilities may also transform national security operations. Indeed, as Richard Moore

of MI6 highlighted, AI, including LLMs, is already a critical factor in current threat environments, acting as a force multiplier for many tools and practices [62].

The nature and degree of this transformation have yet to be determined, but speculation abounds. For example, using AI in defense applications is a disquieting topic for some observers, conjuring pop-culture images of *Skynet* [39]. However, numerous applications lacking autonomous weapon employment can be drastically reshaped through AI integration. Such applications overlap with civilian activities (e.g., natural language processing tasks, computer-vision implementation, and decision-analytic problems), domains where LLMs are being rapidly commercialized. Thus, opinions on using LLMs in defense applications run the gamut; proponents contend that they may be used to create a virtual Clausewitz, whereas dissidents believe their use in any defense setting would result in dangerous hallucinations. Similar contentions arise when considering using LLMs outside of military activity and within the use of other national security instruments, e.g., diplomatic, information, and economic measures [75].

From the perspective of a statistician or data scientist, applying LLMs to national security problems will feel quite familiar despite the domain's derived idiosyncrasies. Nearly every civilian activity has a national security analog, implying that LLMs developed for a particular civilian task may be extended to a national security setting after domain-specific modification. Military physicians may receive decision support from LLMs, provided these systems have been tuned to the realities of combat. Likewise, LLMs may support the decision-making of military commanders via automatic summarization, sentiment analysis, and topic modeling, provided they have been trained on domain-specific vocabulary, acronyms, and jargon. However, the national security setting modifies the relative importance of select LLM design challenges. For example, even if an LLM is only used for automatic summarization during armed conflict, the effect of hallucinations may be catastrophic if these summaries inform senior-level decision-making. On another extreme, if an LLM is used to provide strategy or planning recommendations, then model interpretability and output transparency are paramount to ensuring effective human oversight. The same applies to understanding emergent LLM abilities in national security applications as well [83, 59]. Moreover, when an LLM has access to classified information, the importance of data privacy and the relevance of prompt-injection attacks [e.g., see 33] increases dramatically; if the underlying information is not appropriately safeguarded, security incidents (e.g., data spillage) that endanger national security personnel and the civilian population pose a significant risk. The national security setting is high stakes and leaves little margin for error in LLM applications.

Therefore, within this manuscript, we explore the ramifications of LLMs on national security applications viewed in terms of the strategic competition continuum set forth by the US Joint Chiefs of Staff [73]. This framework does not view international conflict in binary terms (e.g., war and its absence) but interprets interactions between rivals continuously varying from peaceful cooperation to armed conflict across multiple domains. We synthesize varying perspectives on the use of LLMs for national security applications, examine the current usage of LLMs in defense settings, explore their efficacious use in defense applications, and forecast their more general effect on national security writ large. Additionally, whereas LLMs are the focus of this manuscript, emphasis is also

placed on the interface of LLMs with alternative probabilistic, statistical, and machine learning (ML) methods to determine how this interplay may affect national security operations.

The remainder of this manuscript is structured as follows. Section 2 inspects the contemporary application of LLMs by the national security establishment and reviews its relation to scholarly research on the topic. In this manner, we tangibly explore the reception of LLMs among a subset of national security organizations and investigate how they relate to the views of other national security scholars. Based upon this information, Section 3 provides our views on how LLMs can be best integrated into defense applications and discusses their broader repercussions across the other instruments of national power. Section 4 provides concluding remarks and posits avenues of future inquiry.

2. A Rapidly Evolving and Uncertain Landscape

Whereas the impressiveness of LLMs is self-apparent, at the time of this writing, their utility in application is less certain. This is typified by Google's difficulties incorporating Gemini, a multi-modal LLM, into its search engine [12] and recent surveys on generative AI utilization [88]. Nevertheless, the LLM landscape is rapidly evolving, inducing numerous stakeholders to experiment with associated transformations to national security operations. This section reviews such experimentation and their relation to the scholarly literature.

2.1. Recent LLM Initiatives and Applications by National Security Organizations

From the perspective of the US Department of Defense (DoD), the emergence of high-functioning LLMs was perfectly timed for innovation. After nearly two decades focused on counterterrorism operations, the 2018 National Defense Strategy (NDS) refocused the DoD on inter-state strategic competition. This strategy recognized that the military advantage enjoyed by the US after the Cold War had eroded [40] and, as one means to rectify this realty, stated that the DoD would "invest broadly in military application of autonomy, artificial intelligence [AI], and machine learning [ML], including rapid application of commercial breakthroughs." Six years after its publication, the acquisition practices exposed by the 2018 NDS, which are also expressed in the 2022 NDS, have taken root, and their effects are apparent in the DoD's approach to LLMs.

In August 2023, Task Force (TF) Lima was created under the DoD Chief Data and AI Office (CDAO) with the express purpose of identifying low-risk applications for generative AI and LLMs [78]. Such a focus on low-risk areas coincides with recent strategic guidance from Biden [5] and the US Department of Defense [70] on responsible AI use; in the view of Deputy Secretary of Defense Kathleen Hicks, "most commercially available systems enabled by [LLMs] aren't yet technically mature enough to comply with our DoD ethical AI principle." At the time of this writing, TF Lima's work continues; however, it has highlighted the DoD's use of LLMs to streamline staff operations within its internal bureaucracy via, e.g., the automatic summarization of doctrine, instructions, and policy manuals. TF Lima is also endeavoring to identify how LLMs affect classification guidance, especially given the effects of classification by combination, i.e., a national

security particularity whereby specified unclassified information becomes classified upon combination [81]. In line with such concerns, TF Lima is also constructing a "virtual sandbox" through which military personnel can experiment with generative AI tools [80]; suitably, TF Lima has its own ChatGPT instance that can be used to discuss its activities [67].

Despite the ongoing development of a unified experimentation platform, the pace of LLM development has necessitated that other DoD agencies work in parallel with TF Lima. The United States Air Force (USAF) has been particularly active, and LLMs featured chief among the topics discussed at the USAF's 2024 Data, Analytics, and AI Forum. Multiple USAF organizations have developed proof of concepts whereby LLMs are leveraged to expedite myriad coding and administrative tasks. For example, Air Mobility Command has leveraged LLMs to generate campaign simulations based on user-defined, text-based narratives. Moreover, US Air Forces Central, among other USAF organizations, use LLMs to expedite routine maintenance of their endemic software tools. More recently, at the Bravo 11 Hackathon, a US Pacific Air Forces team spearheaded an ambitious use case whereby LLMs automatically summarize incoming mission reports to rapidly distill insights for commanders [32, 57]. In so doing, the team illustrated how an LLM can readily handle a task that, when performed manually, requires multiple operators and weeks of man-hours. Similar time savings were shown at the Air Force Test Center's Data Hackathon when an LLM pipeline was created to automatically generate flight test documents, e.g., plans and reports [11]; the prototypes were based on the relatively small MPT-7b, MPT-30b, Falcon-7b, and Falcon-40b models augmented with a customized retrieval augmented generation (RAG) architecture.

Additionally, Air University's innovation center (AUiX) has been developing a GPT framework for wargaming called the Comprehensive Heuristic Utility for Combat Knowledge (CHUCK), which is currently in a beta stage but shows exciting potential for the future. This initiative is part of a broader collaboration with Stanford University's Hoover Institution and the MIT Artificial Intelligence Accelerator (MIT/AIA), aimed at advancing wargaming techniques and exploring LLM's impact on crisis decision-making. These efforts align with the Air Force Futures office's AI initiatives, exploring how AI can enhance wargaming by running thousands of iterations to optimize strategies and decision-making [21]. Such initiatives aim to transform traditional wargaming by incorporating AI capabilities to improve strategic analysis and operational planning. In conjunction with these wargaming efforts, the Air Force Research Laboratory also developed NIPRGPT, an LLM recently cleared and deployed by the USAF for installation on designated unclassified systems [49].

Alternatively, the USAF's most innovative use cases derive from its academic institutions. Cadets in the Data Science program at the USAF Academy constructed an LLM-based prototype to modernize the user interface in the USAF's Envision platform [77]. The prototype is a "no-code" solution allowing users to submit statistical queries via text-based narratives and receive outputs with corresponding explanations. For example, a user may request a graphical summary of pilot training graduation rates by commissioning source, and the tool may output a box plot with an associated narrative on how to interpret the figure. The USAF Academy has also used LLMs for assessments

within its statistical courses. In the Spring 2024 semester, LLMs were used in its applied statistical modeling class (i.e., MATH 378) to conduct "oral boards" instead of examinations; that is, students were asked to interface with an LLM on a specified set of topics, and the associated transcripts were used to assess student knowledge. LLMs are being further integrated into USAF Academy instruction via the development of a virtual teaching assistant. Notably, seniors in the Data Science program created the QuantaIQ prototype that uses LLMs to generate assignments, conduct assessments, and interface with students via question-and-answer sessions.

Other branches of the United States military are also exploring LLMs, though their approaches and applications vary. The United States Army is experimenting with generative AI in military video games to improve battle planning by using LLMs, including OpenAI's GPT-4 Turbo and GPT-4 Vision models, to provide information on battlefield terrain and details on friendly and enemy forces [61]. Additionally, the Army is developing new policy guidance to guide the department's use of generative AI to streamline operations while addressing security concerns [22]. In collaboration with Scale AI, the Marine Corps created an LLM named Hermes to assist in military planning by synthesizing data, generating hypotheses, and refining courses of action [29]. At Marine Corps University, experts are incorporating LLMs into simulations and wargames to test whether they improve analytical products and ease of use for military students [79]. These experiments demonstrated that LLMs can enhance the planning process and provide valuable insights but also require human oversight to manage current limitations, such as hallucinations and structural biases, to ensure accuracy.

In contrast, the United States Navy has adopted a more cautious approach to using LLMs. According to a recent memo published by the Navy's acting Chief Information Officer, "the use of proprietary or sensitive information poses a unique security risk and has the potential to lead to data compromise when employed by commercial generative AI models" [53]. The Navy's policy advises against using commercial LLMs for operational purposes until security requirements are fully investigated and approved [14]. This cautious stance is reflected in efforts to secure access to LLM technology through Jupiter, the Navy's enterprise data and analytics platform, to ensure safe employment [50]. Similarly, the United States Space Force implemented a temporary ban on using generative AI and LLM tools for official purposes in October 2023 due to concerns about safeguarding sensitive data [23]. As of the time of this writing, the ban remains in place as the service continues to evaluate the best path forward for securely integrating generative AI capabilities into its mission [24].

Beyond the DoD, other US national security agencies are also exploring LLM applications. The Central Intelligence Agency (CIA) began exploring generative AI and LLM applications more than three years before the widespread popularity of ChatGPT. For example, generative AI was leveraged in a 2019 CIA operation called Sable Spear to help identify entities involved in illicit Chinese fentanyl trafficking [2]. The CIA has since used generative AI to summarize evidence for potential criminal cases, predict geopolitical events such as Russia's invasion of Ukraine, and track North Korean missile launches and Chinese space operations [2]. In fact, Osiris, a generative AI tool developed by the CIA, is currently employed by thousands of analysts across all eighteen U.S. intelligence

agencies. Osiris operates on open-source data to generate annotated summaries and provide detailed responses to analyst queries [2]. The CIA continues to explore LLM-incorporation in their mission sets and recently adopted Microsoft's generative AI model to analyze vast amounts of sensitive data within an air-gapped, cloud-based environment to enhance data security and accelerate the analysis process [64]. Other agencies, including the Defense Advanced Research Projects Agency (DARPA), which uses LLMs to detect and fix critical software vulnerabilities, and MITRE, which collaborates with NATO to enhance AI security, are also exploring the effects of LLMs in their application areas [76, 68].

Moreover, allied countries are actively investigating LLMs to enhance military capabilities across various domains as well. For example, the 2024 TIDE Hackathon, co-hosted by NATO Allied Command Transformation (ACT) and the Dutch Ministry of Defence, included an LLM wargaming challenge [48]. This challenge focused on using LLMs to improve military wargaming by creating dynamic scenarios and providing real-time feedback to enhance interoperability, decision-making, and strategy development among allied forces. LLM capabilities have garnered attention from other NATO entities aside from ACT as well. Namely, the NATO Communications and Information Agency (NCI Agency) is developing an AI cognitive agent to automate routine IT-support tasks [20]. Similarly, Hadean, a British deep tech start-up company, secured a contract with the Defence and Security Accelerator (DASA) to develop an LLM for the British Army's virtual training space [25]. This initiative seeks to create a dynamic virtual environment with realistic human terrain and social media simulation, providing real-time feedback, generating complex scenarios, and assisting in after-action reports to enhance military training and decision-making. Additionally, the United States and Australia are leveraging generative AI for strategic advantage in the Indo-Pacific, focusing on applications such as enhancing military decision-making, processing sonar data, and augmenting operations across vast distances [3]. These efforts are accelerating data processing, improving the identification and response to threats, and helping maintain technological and operational superiority by investing in AI talent and reskilling military personnel. These examples demonstrate the transformative potential of LLMs on modern military strategy and operations.

Besides the US DoD and its allies, the nation's strategic competitors (e.g., China, Russia, North Korea, and Iran) are also exploring the national security applications of LLMs. For example, China employs Baidu's Erni Bot, an LLM similar to ChatGPT, to predict human behavior on the battlefield to enhance combat simulations and decision-making [41]. Some reports indicate Baidu's Ernie Bot surpasses ChatGPT in accuracy and real-time information processing but struggles with political inquiries due to restrictions [66]. Additionally, the CopyCat network, suspected to be aligned with the Russian government, leverages LLMs to manipulate media content for disinformation [19, 44]. The CopyCat network plagiarizes, translates, and edits content from legitimate media outlets to spread biased political messages aligned with Russian interests and has been linked to generating sophisticated narratives on the Ukraine conflict and US politics that are particularly difficult for public officials to counteract. Moreover, suspected state-backed hackers from China, Iran, North Korea, and Russia are accused of experimenting

with LLMs to assist in cyber operations, e.g., by generating malicious code and content for phishing campaigns [18]. For instance, North Korea's Kimsuky group uses LLMs to advance their military cyber capabilities by generating content for phishing campaigns, targeting organizations focused on North Korean defense and nuclear issues, thereby enhancing their ability to gather intelligence and exploit vulnerabilities [54]. The widespread exploration of LLMs in such capacities highlights the threat the technology poses, particularly in cyber and information warfare, and necessitates the development of countermeasures to safeguard national security.

2.2. Relation to Scholarly Perspectives

Strictly speaking, LLMs are trained to predict the next word in a phrase conditioned upon previously observed words. However, the surveyed use of LLMs for national security applications is not concerned with this task in and of itself. Instead, national security organizations are often concerned with the *emergent* abilities of LLMs. The existence of such emergent LLM abilities as generally conceived is not universally accepted [e.g., see 35, 59]. Nevertheless, it is undeniable that most national security applications using LLMs seek to extend the technology beyond the task for which the LLMs were trained.

Natural language tasks are heavily featured in the current efforts of national security organizations, and, in this regard, the capabilities of LLMs are well-established. Namely, Min et al. [42] survey the state-of-the-art performance of pre-trained language models on multiple natural language tasks. This survey was recently updated by Minaee et al. [43] and, given the rapid pace of development, another survey will likely be warranted in the near future. Conspicuously relevant in the defense setting are the capabilities of LLMs for automatic summarization. Zhang et al. [90] study LLM performance on news article summarization, finding that a variety of LLMs performed on par with their human counterparts. The authors also determined that prompting and instruction tuning, not model size, dictated model performance in zero-shot summarization. This result presages the general importance of prompt engineering when using LLMs in defense settings. Moreover, given the immense scale of national security operations, it is important to understand how well LLMs summarize longer documents or corpora; these documents often exceed the context window size of standard LLMs. Chang et al. [7] construct and validate a metric that evaluates LLM performance on such documents, and initial results suggest they perform similarly to humans on the task. Schmidt and Robert [60] specifically study LLM summarization performance in the DoD acquisition setting with favorable results.

Alternatively, the self-evident efficacy of LLMs for text generation also implies their efficacy for information warfare. Goldstein et al. [17] and Low [34] explore how LLMs allow for the automation of influence operations, circumventing the need for labor-intensive *troll farms*, once a feature of the space [84]. Coupled with the disinformation threat of other generative AI models [85], the need for an effective detection mechanism for LLM-generated propaganda is imminent. Wan et al. [82] propose a prototypical pipeline for doing so. However, the work of Chen and Shu [8] suggest that such messages may be harder to detect than their man-made counterparts, a task of considerable difficulty in

and of itself [89]. Uchendu et al. [69] also discuss the identification of deepfake texts, concluding that this difficult task may be facilitated via human collaboration.

Interestingly, the desire to leverage AI for strategy development predates the development of LLMs. Bazin [4] contrived the concept of a virtual Clausewitz, a cognitive computing system with access to the whole of human military thought, that would advise senior military leaders. The development of LLMs has since increased the relevance of such ideas. Notably, COA-GPT [15] is an LLM-based tool designed to automate the COA development phase of the joint planning process [74], and the authors test it against baseline reinforcement learning methods in StarCraft II with favorable results. Goecks and Waytowich [16] provide a similar planning tool focused specifically on disaster-relief operations. However, it is worth noting that both Simmons-Edler et al. [63] and Hunter and Bowen [28] warn against such systems, the former arguing against LLM-recommender systems and the latter against AI more generally in military command decisions. Simmons-Edler et al. [63] cites the potential for LLMs to escalate and the intractability of their recommendations. Hunter and Bowen [28] assert that reliance on AI for command decisions may lower the standards and practice of strategy development. Recent empirical research in tabletop war games corroborates some of these concerns. Rivera et al. [55] perform wargaming experiments whereby each nation state is modeled with a commercial LLM; GPT-4, GPT-3.5, Claude 2, Llama-2 (70B) Chat and GPT-4-Base are tested. Therein, the authors find that the LLMs tend to develop arms-race dynamics, leading to greater conflict and, in rare cases, the use of nuclear weapons. Lamparth et al. [31] similarly compare how LLMs and humans conduct themselves in a US-China wargame. The authors find that, while the groups behaved similarly, the LLMs were more aggressive. Additionally, Lamparth et al. [31] found that prompt verbosity affected LLM behavior. Such findings are highly relevant to the future of automated wargaming [e.g., see the *Snow Globe* system of 26].

It may be argued that these more ambitious efforts rely upon an LLM having some form of innate knowledge or, at the least, the ability to reflect human logical reasoning. Diverse perspectives exist in the literature on both of these topics. Yildirim and Paul [87] discuss the nature of LLM knowledge, juxtaposing LLM instrumental knowledge with the worldly knowledge of humans. Saba [58] emphasizes that LLMs cannot be relied upon for factual information because, as currently trained, factual and non-factual information is not differentiated. In this light, if an LLM cannot distinguish fact from fiction, the nature of its knowledge is in question. Moreover, Huang and Chang [27] consider if and how LLMs execute logical reasoning, concluding that, for LLMs on the scale of GPT-3 175B or higher, reasoning is an emergent behavior, despite their difficulties with complex logical tasks. Alternatively, Ellis [13] argues that efficient human learning is intrinsically linked to their ability to reason inductively. He claims that out-of-the-box LLMs are deficient in this task but that this shortcoming can be addressed by layering a Bayesian model on top of a pre-trained LLM. Setting aside the actual nature of LLM knowledge and reasoning, many researchers and practitioners seek to leverage LLMs in various game-playing and strategic decision-making tasks. Xu et al. [86] and Chen and Chu [10] provide comprehensive surveys, respectively, that provide further context on LLM use for national security planning.

Finally, in view of the momentous effects associated with national security operations, we would be remiss if we excluded interpretable, explainable, or adversarial machine learning from this discussion. Each of these sub-disciplines is consequential to using LLMs for strategic decision-making. Both interpretable and explainable machine learning focus on tractability but through distinct means. Interpretable ML seeks to build simple yet powerful models, and explainable ML endeavors to identify simple yet faithful approximations of a black-box model. To date, LLMs are best viewed as a black-box model, implying that explainable techniques [e.g., see the Deep SHAP values of 36] are most readily applicable; however, the development of interpretable image classifiers [e.g., see 9] suggest that the development of high-performing, interpretable language models may be feasible. Applying either approach is useful if it can successfully identify undesirable behavior, e.g., bias in automatic summarization [91]. Alternatively, adversarial ML focuses on the corruption and defense of ML methods in the presence of an opponent. In the context of LLMs, training-time attacks can be used to modify output in accordance with the attacker's objectives. Bagdasaryan and Shmatikov [1] provide a concrete example whereby poisoning attacks to the training data encourage LLMs to spin automatically generated summaries according to the adversary's point of view. Deployment-time attacks against a pre-trained LLM focus on modifying input data to affect a desired output. Raina et al. [52] explore such attacks against assessment LLMs whereby a short universal phrase is appended to a phrase to disrupt model performance. The threat of such attacks to national security is self-evident. If an attacker can manipulate an LLM that informs command decision-making, traps may be laid in advance so that their adversary acts per their desires. For additional information on adversarial ML, we point the interested reader to Oprea and Vassilev [51].

3. Discussion

National security professionals are generally proceeding cautiously with LLM-based technology. Although some seek to use LLMs as a foundation upon which to build a virtual Clausewitz, the primary focus of national security organizations is to use LLMs for natural language tasks, a function in which their capabilities are well-established. This calculated approach is enshrined in the strategic documents and international agreements adopted by many countries. For example, the US Department of Defense [70] set forth a framework for AI implementation conditioned upon its responsible, equitable, traceable, reliable, and governable use. Similar tenets are foundational to the approach toward emerging and disruptive technologies set forth by NATO [47]. The Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, endorsed by 54 countries, also echoes analogous themes [72]. In our estimation, such deliberate action is appropriate given the nascent nature of LLMs and our incipient understanding of their performance.

Utilizing LLMs for standard natural language tasks, e.g., automatic summarization, is admittedly less intriguing than their use for generating automated campaign plans. However, we contend that if these less alluring functions were embedded across the national security ecosystem, the effects would be transformational. Defense organizations

are characterized by massive bureaucracies. In these bureaucracies, many offices are tasked with aggregating information and distilling insights to send to the next level in the bureaucratic hierarchy. Others are responsible for aggregating information, proposing courses of action for a commander's decision, and disseminating these decisions via written documentation. When conducted exclusively by humans, these tasks are incredibly slow and cumbersome. Despite the adoption of information-age technology, many defense organizations have struggled to effectively manage data [71], forcing staffs to rely on manual inquiry and collection. Moreover, many national security operations, particularly those below armed conflict, inhibit the development and use of structured databases. National objectives are often qualitatively defined without a clear quantitative analog [e.g., see the Hamlet Evaluation System discussed by 45], allocated resources may rapidly change due to shifting national priorities, and personnel are transient (e.g., service-member deployments).

Much of the data available to national security organizations is thus unstructured, often in the form of text or images. Current practices require a human to ingest, summarize, and distill insight from this information, thereby limiting decision-making to the speed of human comprehension. Therefore, LLM-based automatic summarization may be particularly well-suited for national security applications. Such models can ingest and summarize a significant amount of unstructured information much quicker than their human counterparts. Alternatively, in use cases where structured databases exist, LLMs may be used as a component within broader systems that translate text inquiries into code, thereby allowing a layman to readily interact with data. For example, a non-technical analyst may enter a text query that triggers the LLM to generate structured query language (SQL) code for data set creation, R code for visualization, and text suggestions for more follow-up statistical analyses.

Integrating LLMs with alternative probabilistic, statistical, and ML methods can further streamline information processing and analysis by breaking down queries into manageable steps and leveraging external tools for precise calculations. For instance, combining LLMs with statistical forecasting methods can enhance the accuracy of intelligence predictions, and using LLMs alongside supervised ML techniques can improve data classification and analysis. Moreover, through interaction with an LLM that has access to the relevant information, an analyst can conversationally explore a compendium of structured and unstructured data to derive novel insights. By facilitating knowledge-based work, LLMs may allow commanders to reduce their bureaucratic footprint and reallocate personnel from administrative to operational roles. LLMs may also enable the deconstruction of information silos fashioned within the bureaucracy due to limited human awareness and communication. Coupling these potentialities with the speed of LLM processing, national security organizations may be able to rapidly accelerate and improve their decision-making cycles; Sundar [65] provides an example in USAF operations wherein an LLM needed 10 minutes to complete a task that typically requires days of man-hours.

Nevertheless, hallucinations remain a potential threat, and this risk must be appropriately mitigated. Recent research by Nahar et al. [46] suggests that simply warning users about hallucinations improves their ability to detect them. Future academic studies

should also compare LLM-induced hallucinations to errors induced by transmission chaining in large bureaucracies; such studies would provide a baseline to determine the degree to which LLM hallucinations deviate from errancy in the status quo. Alternative comparative studies could view the problem from a decision-theoretic perspective, investigating how well each method summarizes data and how this affects decision-making. LLMs may fall short of expectations, or their utilization in national security staffs may evoke unforeseen consequences; however, they hold great promise in expediting sluggish bureaucratic processes.

Conversely, the direct use of LLMs for real-world, strategic-level planning is more fraught. Recently developed LLMs have shown "sparks" of artificial general intelligence (AGI) [6], obtaining near human performance on tasks from a variety of fields (e.g., medicine, law, and psychology); however, we contend that national security planning is a function of a different character. No textbook contains the right answers from which an LLM can learn. The work of military historians, along with the treatises of Sun Tzu, Jomini, and Clausewitz, contain valuable information that may allow an LLM to craft a written strategic document comparable to the status quo. However, such documents are aspirational and are useful insofar as the humans executing the strategy achieve unified action across domains throughout the strategic, operational, and tactical levels of command. Achievement of this goal has eluded even the most skilled and experienced national security leaders. Thus, we are skeptical that a recipe for its attainment is embedded even within even the whole of written human text. That is not to say that AI cannot attain superhuman abilities for the task or that LLMs are not the building blocks to achieve it, but we believe that more than sparks of AGI would be required. Alternatively, we are more optimistic about utilizing contemporary LLMs to train national security strategists and planners. Their use in wargaming allows for individual study and personalized instruction heretofore lacking in professional military education. Automated, qualitative wargames [e.g., see 26] may be calibrated for specific scenarios and adversaries, transcripts may be reviewed later with instructors to facilitate learning, and students may replicate the same wargame autonomously to rectify their mistakes. We find this last feature of replication particularly compelling. Military planners often struggle to cope with the vast degree of uncertainty inherent in their operations and do not understand the distinction between decision and outcome quality. If properly executed during wargame training, LLMs may be able to partially rectify these issues.

Based on the widespread efforts by national security organizations, we believe LLMs will further the AI-induced change in the character of conflict. In particular, LLMs will disrupt the status quo under the threshold of armed conflict. This is already being seen through their use in the development of nation-state-sponsored propaganda. Disinformation campaigns are becoming cheaper at scale, and automated propaganda is harder to discern. This trend will only continue as LLMs are paired with alternative generative AI methods (e.g., those creating images and videos). Moreover, as LLMs are incorporated into national security processes, they will further expand the relevance of adversarial machine learning and instantiate a new cyber-warfare arena. The black-box nature of existent LLMs increases this threat, making the incorporation of interpretable

and/or explainable models imperative to enabling attack-and-vulnerability detection. Given the breadth of contemporary LLM initiatives and applications, the aforementioned threats and mitigation efforts are relevant across the competition continuum.

4. Conclusion

Integrating LLMs into national security operations presents unprecedented opportunities and significant challenges. As evidenced by their varied applications across the DoD and other world defense organizations, LLMs have the potential to revolutionize the efficiency (and effectiveness) of national security operations. The potential benefits are substantial. LLMs can automate and accelerate information processing, enhance decision-making through advanced data analysis, and reduce bureaucratic inefficiencies. Their automatic summarization capabilities can streamline the creation of operational documents and reports, while their integration with probabilistic, statistical, and ML methods can improve accuracy and reliability, e.g., combining LLMs with Bayesian techniques can provide more robust threat predictions. Conversely, deploying LLMs into national security organizations does not come without risks. More specifically, the potential for hallucinations, the ensuring of data privacy, and the safeguarding of LLMs against adversarial attacks are significant concerns that must be addressed. These risks are particularly concerning in high-stakes decision-making environments where the accuracy and integrity of information are crucial (e.g., armed conflict). Addressing these challenges is critical to protecting sensitive information, preventing malicious exploitation, and avoiding potential national security catastrophes.

The broader implications of LLM integration into national security organizations are profound. Beyond immediate defense applications, LLMs have the potential to influence strategic planning, international relations, and the broader geopolitical landscape. The purported ability of nations to leverage LLMs for disinformation campaigns and cyber operations emphasizes the need to develop appropriate countermeasures and continuously scrutinize and update AI security protocols. While LLMs have demonstrated "sparks" of AGI, we contend that their current capabilities are best suited for supporting roles rather than leading strategic decisions. Their use in training and wargaming can provide valuable insights and personalized learning experiences for military personnel, helping to bridge knowledge gaps and improve operational readiness.

A cautious and calculated approach must be taken by national security professionals with regard to integrating LLMs into their operations. Deliberate integration, guided by frameworks for responsible AI use, is both appropriate and necessary for harnessing the potential benefits of LLMs while mitigating associated risks. Moving forward, continued research and collaboration between defense, academic, and commercial entities is essential to fully realize the benefits of LLMs. National security professionals must ensure that, while they explore how these powerful tools can be leveraged to enhance national security capabilities, they simultaneously focus on safeguarding against potential threats. This balanced approach will enable them to navigate the complexities of AI integration and establish strategic advantage in an increasingly contested and technologically advanced world.

Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, United States Department of Defense, or United States Government. This work is approved for public release (distribution unlimited) in accordance with PA# USAFA-DF-2024-514.

References

[1] Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 769–786. IEEE, 2022.

[2] Frank Bajak. U.s. intelligence agencies' embrace of generative ai is at once wary and urgent, 2024. URL `https://www.pbs.org/newshour/world/u-s-intelligence-agencies-embrace-of-generative-ai-is-at-once-wary-and-urgent`. Accessed: 2024-06-18.

[3] Ylli Bajraktari. The us and australia need generative ai to give their forces a vital edge. *The Strategist*, 2024. URL `https://www.aspistrategist.org.au/the-us-and-australia-need-generative-ai-to-give-their-forces-a-vital-edge/`. Accessed: 2024-06-18.

[4] Aaron Bazin. How to build a virtual clausewitz. *The Strategy Bridge*, March 2017. URL `https://thestrategybridge.org/the-bridge/2017/3/21/how-to-build-a-virtual-clausewitz`.

[5] Joe Biden. Executive order 14110: Safe, secure, and trustworthy development and use of artificial intelligence, October 2023.

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[7] Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*, 2023.

[8] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

[9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[10] Yuwei Chen and Shiyong Chu. Large language models in wargaming: Methodology application and robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2894–2903, 2024.

[11] Jordan Conner, Luis Moros, Riley Livermore, Danny Riley, Troy Soileau, Ben Faircloth, Tim Lortz, and Li Yu. Us air force hackathon: How large language models will revolutionize usaf flight test. *Databricks Blog*, March 2024. URL `https://www.databricks.com/blog/us-air-force-hackathon-how-large-language-models-will-revolutionize-usaf-flight-test`.

[12] Michael Dobuski. Google makes adjustments to ai overviews after a rocky rollout. ABC News, June 2024. URL `https://abcnews.go.com/Technology/google-makes-adjustments-ai-overviews-rocky-rollout/story?id=99876432`.

[13] Kevin Ellis. Human-like few-shot learning via bayesian reasoning over natural language. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Kirsten Errick. Navy discourages military generative ai, llm usage, October 2023. URL `https://federalnewsnetwork.com/artificial-intelligence/2023/10/navy-discourages-military-generative-ai-llm-usage/`. Accessed: 2024-06-18.

[15] Vinicius G Goecks and Nicholas Waytowich. Coa-gpt: Generative pre-trained transformers for accelerated course of action development in military operations. *arXiv preprint arXiv:2402.01786*, 2024.

[16] Vinicius G Goecks and Nicholas R Waytowich. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. *arXiv preprint arXiv:2306.17271*, 2023.

[17] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

[18] Elias Groll. State-backed hackers are experimenting with openai models. CyberScoop, February 14 2024. URL `https://www.cyberscoop.com/openai-microsoft-apt-llm/`.

[19] Insikt Group. Russia-linked copycat uses llms to weaponize influence content at scale. `https://www.recordedfuture.com/russia-linked-copycop-uses-llms-to-weaponize-influence-content-at-scale`, 2024. Accessed: 2024-06-15.

[20] John Harper. Nato on the hunt for new ai cognitive agent. DefenseScoop, November 14 2023. URL `https://defensescoop.com/2023/11/14/nato-on-the-hunt-for-new-ai-cognitive-agent/`.

[21] Jon Harper. Ai wargaming: Air force futures at mit. https://defensescoop.com/2024/04/12/ai-wargaming-air-force-futures-mit/, April 12 2024. Accessed: 2024-06-17.

[22] Jon Harper. Army set to issue new policy guidance on use of large language models, 2024. URL https://defensescoop.com/2024/05/09/army-policy-guidance-use-large-language-models-llm/. Accessed: 2024-06-19.

[23] Unshin Lee Harpley. Space force pumps the brakes on chatgpt-like technology with temporary ban. https://www.airandspaceforces.com/space-force-chatgpt-technology-temporary-ban/, 2023. Accessed: 2024-06-19.

[24] Unshin Lee Harpley. Air force launches its own generative ai chatbot. experts see promise and challenges. https://www.airandspaceforces.com/air-force-launches-generative-ai-chatbot/, 2024. Accessed: 2024-06-12.

[25] John Hill. Hadean builds large language model for british army virtual training space. https://www.army-technology.com/news/hadean-builds-large-language-model-for-british-army-virtual-training-space/, 2024. Accessed: 2024-06-15.

[26] Daniel P Hogan and Andrea Brennen. Open-ended wargames with large language models. *arXiv preprint arXiv:2404.11446*, 2024.

[27] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[28] C. Hunter and B. E. Bowen. We'll never have a model of an ai major-general: Artificial intelligence, command decisions, and kitsch visions of war. *Journal of Strategic Studies*, pages 1–31, 2022.

[29] Benjamin Jensen and Dan Tadross. How large-language models can revolutionize military planning, 2023. URL https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/. Accessed: 2024-06-19.

[30] Jenelle Jindal, Suhana Bedi, Akshay Swaminathan, Michael Wornow, Jason Fries, Akash Chaurasia, and Nigam Shah. Large language models in healthcare: Are we there yet?, May 2024. URL https://hai.stanford.edu/news/large-language-models-healthcare-are-we-there-yet. Stanford University Human-centered Artificial Intelligence.

[31] Max Lamparth, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, and Harold Trinkunas. Human vs. machine: Language models and wargames. *arXiv preprint arXiv:2403.03407*, 2024.

[32] 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.

[33] Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*, 2024.

[34] Jwen Fai Low. *Automated Information Warfare: For and Against Saturation Attacks*. PhD thesis, McGill University, 2023.

[35] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*, 2023.

[36] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[37] Duncan MacRae. Large language models could 'revolutionise the finance sector within two years, March 2024. URL `https://www.artificialintelligence-news.com/2024/03/27/large-language-models-could-revolutionsise-the-finance-sector-within-two-years/`. AINews.

[38] Anne J. Manning. What is 'original scholarship' in the age of ai?, May 2024. URL `https://www.news.harvard.edu/gazette/story/2024/05/how-is-generative-ai-changing-education-artificial-intelligence/`. The Harvard Gazette.

[39] John Markoff and Matthew Rosenberg. The pentagon's 'terminator conundrum': Robots that could kill on their own. New York Times, October 25 2016. URL `https://www.nytimes.com/2016/10/26/us/pentagon-artificial-intelligence-terminator.html`.

[40] Jim Mattis. Summary of the 2018 national defense strategy of the united states of america. Technical report, Department of Defense Washington United States, 2018.

[41] Christopher McFadden. China is training ai to predict human behavior on the battlefield. `https://interestingengineering.com/military/china-training-ai-predict-humans`, 2024. Accessed: 2024-06-15.

[42] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[43] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[44] Phil Muncaster. Ai-powered russian network spreads fake news. `https://www.infosecurity-magazine.com/news/aipowered-russian-network-fake-news/`, 2024. Accessed: 2024-06-15.

[45] Emily Mushen and Jonathan Schroden. Are we winning? a brief history of military operations assessment. *DTIC Document, DOP-2014-U-008512-1Rev*, 2014.

[46] Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations. *arXiv preprint arXiv:2404.03745*, 2024.

[47] NATO. Nato 2022 strategic concept. https://www.nato.int/nato_static_fl2014/assets/pdf/2022/6/pdf/290622-strategic-concept.pdf, 2022. Accessed: 2024-06-19.

[48] NATO Allied Command Transformation. TIDE Hackathon: Spotlight on the Wargaming LLM Challenge. NATO ACT, October 20 2023. URL https://www.act.nato.int/article/tide-hackathon-spotlight-wargaming-llm-challenge/.

[49] Anastasia Obis. Air force unveils new generative ai platform. https://federalnewsnetwork.com/defense-main/2024/06/air-force-unveils-new-generative-ai-platform/, June 2024. Accessed: 2024-06-17.

[50] Brian O'Connell. U.s. navy not ready to set sail on ai just yet. AI Finance Today, October 2023. URL https://aifinancetoday.com/u-s-navy-not-ready-to-set-sale-on-ai-just-yet/. Accessed: 2024-06-18.

[51] Alina Oprea and Apostol Vassilev. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2023.

[52] Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*, 2024.

[53] Jane O. Rathbun. Guidance on the use of generative ai and large language models. Memorandum, Department of the Navy, Chief Information Officer, September 2023. URL file:///C:/Users/Cpt%20Jenkins1/Downloads/DONCIOMemoGuidanceonUseofGenerativeAIandLargeLanguageModels06Sept20231.pdf. Accessed: 2024-06-18.

[54] Shreyas Reddy. North korean hackers take phishing efforts to next level with ai tools. https://www.nknews.org/2024/02/north-korean-hackers-take-phishing-efforts-to-next-level-with-ai-tools-report/#:~:text=called%20%E2%80%9Cfrightening.%E2%80%9D-,Microsoft%20reported%20that%20it%20observed%20North%20Korean%20threat%20group%20Kimsuky,generate%20content%20for%20phishing%20campaigns, 2024. Accessed: 2024-06-15.

[55] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation risks from language models in military and diplomatic decision-making. *arXiv preprint arXiv:2401.03408*, 2024.

[56] Kevin Roose. Gpt-4 is exciting and scary. *The New York Times*, 15, 2023.

[57] Justin Ross. Bravo 11 hackathon trip report. Technical report, USAF, Air Mobility Command, 2024.

[58] Walid S Saba. Stochastic llms do not understand language: Towards symbolic, explainable and ontologically based llms. In *International Conference on Conceptual Modeling*, pages 3–19. Springer, 2023.

[59] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

[60] Douglas Schmidt and John Robert. Applying large language models to dod software acquisition: An initial experiment. Carnegie Mellon University, Software Engineering Institute's Insights, Apr 2024. URL `https://doi.org/10.58012/s32x-gc46`. Accessed: 2024-Jun-21.

[61] Science Desk. Us army experimenting with generative ai chatbots in war games: Report, 2024. URL `https://indianexpress.com/article/technology/science/us-army-ai-chatbot-war-games-9199456/`. Accessed: 2024-06-19.

[62] Jim Sciutto. *The Return of Great Powers: Russia, China, and the End of American Exceptionalism*. Penguin Press, New York, 2023. ISBN 9780593474132.

[63] Riley Simmons-Edler, Ryan Badman, Shayne Longpre, and Kanaka Rajan. Ai-powered autonomous weapons risk geopolitical instability and threaten ai research. *arXiv preprint arXiv:2405.01859*, 2024.

[64] Peter Suciu. Cia adopts microsoft's generative ai model for sensitive data analysis, 2024. URL `https://www.clearancejobs.com/news/cia-adopts-microsofts-generative-ai-model-for-sensitive-data-analysis/`. Accessed: 2024-06-18.

[65] Sindhu Sundar. Air force and defense department using ai and llm with humans. `https://www.businessinsider.com/air-force-defense-department-using-ai-artificial-intelligence-llm-humans-2023-7`, July 2023.

[66] Azmi Tamin. Baidu's ernie bot better at accuracy than chatgpt but lingers in politics. `https://interestingengineering.com/innovation/baidus-ernie-bot-better-at-accuracy-than-chatgpt-but-lingers-in-politics`, 2024. Accessed: 2024-06-15.

[67] TF Lima. Task force lima and gpt. `https://chatgpt.com/g/g-v12me2Sha-task-force-lima-gpt`, 2024.

[68] Catherine Trifiletti. Nato leverages mitre's ai expertise. MITRE, 2023. URL `https://www.mitre.org/news-insights/impact-story/nato-leverages-mitres-ai-expertise`.

[69] Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174, 2023.

[70] US Department of Defense. DoD Responsible AI (RAI) Strategy and Implementation Pathway, 2022. Accessed 17 Jun 2024.

[71] U.S. Department of Defense. Dod data analytics and ai adoption strategy. https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF, November 2 2023.

[72] U.S. Department of State. Political declaration on responsible military use of artificial intelligence and autonomy. https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/, 2024. Accessed: 2024-06-19.

[73] US Joint Chiefs of Staff. Joint doctrine note 1-18. Technical report, US Joint Chiefs of Staff, 2018. URL https://www.jcs.mil/Portals/36/Documents/Doctrine/jdn_jg/jdn1_18.pdf.

[74] US Joint Chiefs of Staff. *Joint Publication 5-0: Joint Planning*. U.S. Department of Defense, 2020. URL https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp5_0.pdf.

[75] US Joint Chiefs of Staff. Joint publication 3-0: Joint operations. Technical report, US Joint Chiefs of Staff, 2022. URL https://www.jcs.mil/Doctrine/Joint-Doctrine-Pubs/3-0-Operations-Series/.

[76] David Vergun. Darpa aims to develop ai autonomy applications warfighters can trust, 2024. URL https://www.defense.gov/News/News-Stories/Article/Article/3722849/darpa-aims-to-develop-ai-autonomy-applications-warfighters-can-trust/#:~:text=An%20example%20of%20that%2C%20he,software%20that%20underlies%20critical%20infrastructure. Accessed: 2024-06-20.

[77] Brandi Vincent. Air and space forces lean into data-informed decision making. https://defensescoop.com/2023/03/22/air-and-space-forces-lean-into-data-informed-decision-making/, March 22 2023. Accessed: 2024-06-17.

[78] Brandi Vincent. Inside task force lima's exploration of 180-plus generative ai use cases for dod. *DefenseScoop*, 2023. URL https://defensescoop.com/2023/11/06/inside-task-force-limas-exploration-of-180-plus-generative-ai-use-cases-for-dod/.

[79] Brandi Vincent. How marine corps university is experimenting with generative ai in simulations and wargaming, 2023. URL https://defensescoop.com/2023/06/28/how-marine-corps-university-is-experimenting-with-generative-ai-in-simulations-and-wargaming/. Accessed: 2024-06-18.

[80] Brandi Vincent. Task force lima preps new space for generative ai experimentation. DefenseScoop, April 2 2024. URL https://defensescoop.com/2024/04/02/task-force-lima-preps-new-space-generative-ai-experimentation/#:~:text=AI-,Task%20Force%20Lima%20preps%20new%20space%20for%20generative%20AI%20experimentation,to%20unleash%20an%20experimental%20sandbox.

[81] Brandi Vincent. Cdao developing 'classification guide' for large language models. *DefenseScoop*, 2024. URL https://defensescoop.com/2024/02/21/cdao-classification-guide-large-language-models-lugo/.

[82] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*, 2024.

[83] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[84] Emma Woollacott. Russian trolls outsource disinformation campaigns to africa. *Forbes*, 2020. Senior Contributor.

[85] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9291–9298, 2023.

[86] Xinrun Xu, Yuxin Wang, Chaoyi Xu, Ziluo Ding, Jiechuan Jiang, Zhiming Ding, and Börje F Karlsson. A survey on game playing agents and large models: Methods, applications, and challenges. *arXiv preprint arXiv:2403.10249*, 2024.

[87] Ilker Yildirim and LA Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 2024.

[88] Marc Zao-Sanders. How People Are Really Using GenAI. Harvard Business Review, March 2024. URL https://hbr.org/2024/03/how-people-are-really-using-genai.

[89] Chaowei Zhang, Ashish Gupta, Christian Kauten, Amit V Deokar, and Xiao Qin. Detecting fake news for reducing misinformation risks using analytics approaches. *European Journal of Operational Research*, 279(3):1036–1052, 2019.

[90] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.

[91] Karen Zhou and Chenhao Tan. Characterizing political bias in automatic summaries: A case study of trump and biden. *arXiv preprint arXiv:2305.02321*, 2023.