SCANDINAVIAN JOURNAL OF MILITARY STUDIES

GenAI in the Military: Trends and Opportunities

RESEARCH ARTICLE

LAURI VASANKARI

AAPO KOSKI

*Author affiliations can be found in the back matter of this article

SCANDINAVIAN MILITARY STUDIES

ABSTRACT

Generative artificial intelligence offers the potential for significant, far-reaching applications within the military domain. This article surveys current trends and explores emerging opportunities for GenAI in military operations, focusing on its integration into strategic planning, decision-making, and operational efficiency. Key areas of application include information extraction, decision support, mission simulations and information warfare. This paper further examines critical challenges such as ethical considerations, bias mitigation, hallucination management, and secure deployment within classified environments. By addressing these challenges and leveraging the opportunities presented by GenAI, military organizations can enhance their capabilities, maintain strategic advantages, and ensure preparedness for future conflicts. The paper concludes with recommendations for research, development, and collaboration across NATO and allied forces to fully harness the power of GenAI in defense and security.

CORRESPONDING AUTHOR:
Lauri Vasankari
University of Turku, Finland
lauri.vasankari@gmail.com

KEYWORDS:
generative AI; military; artificial intelligence; defense technology

TO CITE THIS ARTICLE:
Vasankari, L., & Koski, A. (2025). GenAI in the Military: Trends and Opportunities. *Scandinavian Journal of Military Studies*, 8(1), pp. 416–434. DOI: https://doi.org/10.31374/sjms.415

DOI: 10.31374/sjms.415

INTRODUCTION

Generative artificial intelligence (GenAI), "a transformer-based machine learning model trained in an unsupervised manner on extensive datasets and specifically optimized for generating valuable data through prompts" (Oniani et al., 2023), has rapidly emerged as a transformative force across industries (Maslej et al., 2024). Its potential within the military domain is profound (Grand-Clément, 2023).

The ability of GenAI to generate, synthesize, and analyze vast amounts of information has sparked growing interest among defense organizations seeking to enhance strategic planning, decision-making, and operational efficiency. From simulating war-gaming scenarios (Hua et al., 2024; Rivera et al., 2024) to enhancing intelligence analysis and acting as a force multiplier through disinformation and false signaling (Zarrar & Kakar, 2024), GenAI applications are increasingly becoming integral to modern defense strategies.

The integration of GenAI into military operations presents both unprecedented opportunities and complex challenges. On one hand, GenAI has the potential to revolutionize decision-support systems, mission planning, and autonomous warfare by processing and interpreting large datasets in real time. It can improve operational readiness through enhanced training simulations (Chuang & Cheng, 2022), optimize logistics and supply chain management (Boone et al., 2025), and contribute to cyber defense by detecting anomalies and novel threats (Uddin et al., 2025). On the other hand, the deployment of GenAI in military contexts raises critical concerns, including ethical considerations, questions of bias mitigation (Stebbins et al., 2024), adversarial vulnerabilities (Shayea et al., 2025), and the secure implementation of AI-driven technologies within classified environments.

While algorithmic bias is not solely a GenAI issue, since all AI systems may be skewed in some manner due to data, the design of the model's algorithms, the learning process, and objectives or performance metrics (Blanchard & Bruun, 2024), GenAI brings new challenges; being more relatable to human users, it is capable of greatly altering the way humans interact with technology (Schneider, 2025).

This paper introduces GenAI and the latest advances in its field before proceeding to explore the evolving landscape of GenAI in military applications. The focus is on identifying key trends, emerging opportunities, and associated challenges. By reviewing the latest research and technological advancements, we analyze how GenAI is shaping the future of defense operations. Further, we examine the implications of GenAI adoption in military contexts, including its technical challenges and limitations and its consequences for human-AI collaboration, ethical governance, and strategic policy frameworks. The study also discusses the role of NATO and allied forces in fostering research, development, and responsible implementation of GenAI-driven defense solutions.

This study provides an understanding of how GenAI is amplifying military capabilities and the strategic steps necessary for its successful deployment. The insights contribute to the discourse on AI-driven military innovations and inform future policy and research.

METHODS

This study presents a comprehensive survey and analysis of literature published on GenAI in the military and defense sectors between 2022 and 2025. The selected timeframe aligns with the rapid advancements in GenAI technology and its increasing adoption in defense applications. The research questions concern the current state of research and development of GenAI in the military context. What are the challenges? What are the assumed trends and recommended courses of action for stakeholders?

To ensure a structured and relevant review, publications were sourced from RAND Corporation reports and Royal United Services Institute (RUSI) publications due to their military focus, and from academic databases and search engines such as Google Scholar. The search query used to identify relevant literature included the following: "(Military OR Defense) AND (GenAI OR LLM OR Generative Artificial Intelligence)".

The results were filtered to focus on studies published within the last three years (2022–2025). The query resulted in 24,100 hits for this period. Manual filtering to reduce the number of initial search hits was conducted; apart from institutional platforms, only the first 200 returns,

sorted by 'relevance', were assessed, focusing on open- and institution-access papers. From this dataset, 49 publications were selected for closer examination based on abstracts and indicated relevance to military applications of GenAI. Studies that were duplicative, lacked direct relevance to GenAI, or were considered low-impact (e.g., lacking peer review, expert credibility or detailed methodology) were excluded. Of the 49 papers, 29 were selected for the review. Publications with a GenAI focus or relevance and a military context were selected. The distribution of publications is displayed in Figure 1. Six papers were included despite being pre-prints due to their applicative nature.

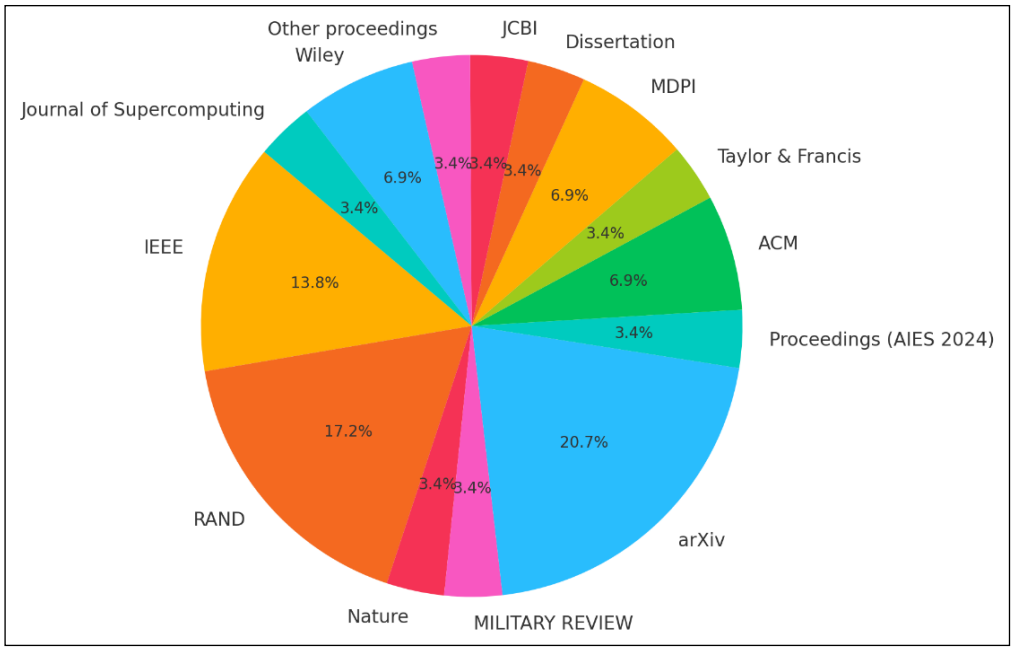

Figure 1 The distribution of publication venues for the reviewed papers.

Each selected publication was categorized based on its research method and primary contribution. The classification included six categories:

1. **Survey** – Studies presenting an overview of existing research without introducing new methods.
2. **Review** – Critical evaluations of past research with additional insights or critiques.
3. **Policy analysis** – Studies focused on policies, governance or strategic frameworks.
4. **Application** – Studies proposing tested and evaluated solutions for military use.
5. **Proposition** – Theoretical frameworks, new architectures, or conceptual models.
6. **Overview** – Papers addressing singular topics.

To enhance consistency in classification, each paper was analyzed by two researchers to validate its category assignment. Further, application-focused and propositional papers underwent deeper examination to assess readiness levels, technological feasibility, and real-world implementation challenges.

By systematically analyzing these sources, this study provides a comprehensive understanding of the current research landscape, identifies gaps, and highlights emerging opportunities and challenges for GenAI in military applications.

GENAI AND THE STATE OF THE ART

While GenAI's rise to prominence is a very recent phenomenon, its development can be traced back to the mid-20th century. Joseph Weizenbaum introduced ELIZA, an early chatbot that employed rule-based scripts to generate a human-like conversation, in 1966 (Weizenbaum, 1966). In the late 1980s probabilistic methods such as Bayesian networks were used to introduce complex probabilistic relationships (Pearl, 1988). This foundation was enhanced by advancements in neural networks such as recurrent neural networks (RNN) and long short-term

DOI: 10.31374/sjms.415

memory (LSTM; see Hochreiter & Schmidhuber, 1997). The modern era of generative AI can be deemed to have started in 2014, when Ian Goodfellow and his colleagues (2014) introduced generative adversarial networks (GAN). The idea of GAN is to train a discriminator and a generator network, where the generator aims to generate new data and the discriminator evaluates how closely the generated (i.e., synthetic) outputs match the source data. While the GAN proved to be very efficient and successful in many tasks, notably image generation (Radford et al., 2016), followed by even more capable diffusion methods (Bishop & Bishop, 2023). The current state of GenAI development began with the invention of transformer architecture with an attention mechanism (Vaswani et al., 2017).

Essentially, the transformer is an encoder–decoder architecture in which the input is encoded into a sequence of continuous representations by the encoder and then decoded into an output sequence that constitutes the generated content. This mechanism can be used in applications such as language translation. The attention mechanism enables the model to evaluate relationships between all parts of an input sequence regardless of their distance, in contrast to the sequential recurrence used in RNNs. Specifically, Vaswani and his colleagues (2017) introduced multi-head attention into the decoder-encoder structure – a system allowing multiple types of contextual relationships to be computed in parallel. For the decoder output, each word in the learned vocabulary is assigned a probability of being the next word in the generated sequence, conditioned on the context calculated by the decoder stack. Most current architectures do consist of only one of the two, the encoder or a decoder, depending on the task and scope.

For the time being, advances in technology are mainly due to the scaling up of the transformer models and extensive training with vast datasets, resulting in very large, multimodal models (where "multimodal" designates data types other than text and natural language). The research and development of transformer architecture has since been accelerated by at least eight major innovations, explained below.

MIXTURE OF EXPERTS

In 2017, a Google research team introduced a mixture of experts (MoE) model to leverage "outrageously" large neural networks more efficiently (Shazeer et al., 2017). The innovation, achieved in the same year as the transformer with attention mechanism (Vaswani et al., 2017), focuses on conditional computation that leverage very large neural networks efficiently. The proposed sparsely-gated MoE utilizes a gating network that determines a sparse combination of sub-networks for each input, achieving vastly better model capacity. For example, utilizing suggested solution enables the parameters to be scaled up more than 1000 times without significantly increasing computational costs during inference.

RETRIEVAL-AUGMENTED GENERATION (RAG)

After the launch of widely known GenAI applications such as ChatGPT, issues of GenAI reliability have become a concern. GenAI models are prone to generate unintended or plainly false responses, a phenomenon known as "hallucination" (Ji et al., 2023).

To combat this, Lewis et al. (2020) have introduced the *retrieval-augmented generation* (RAG) framework, which utilizes a vectorized database of source documents, from which top-*K* documents or chunks are retrieved for generation for each query. The aim is to reduce generative AI model hallucination and focus on the scope of the information that exists within the document database. Results show that RAG models are able to generate more specific and factual answers than parametric-only baselines. The approach has significantly improved the utilization of a particular, factual information database that needs to be queried with a minimum of errors – something significant in a military context.

FEW-SHOT LEARNING

Few-shot learning, a machine learning method in which models are trained or prompted using only a small number of examples, has been shown to produce results comparable to fine-tuning when applied to LLMs (Brown et al., 2020). In essence, few-shot learning utilizes K examples – in Brown and colleagues' study, between 10 and 100 – consisting of context-completion pairs, followed by a final example of context for which the model is expected to generate a completion. A one-shot approach uses K = 1, while zero-shot learning relies solely

DOI: 10.31374/sjms.415

on a natural language prompt without any task-specific examples. Brown et al. found that the few-shot learning approach nearly matches state-of-the-art fine-tuned results, providing a computationally efficient alternative to full fine-tuning. The experiment was carried out on GPT-3, which has 175 billion parameters, but the premise has been shown to hold for smaller models as well (Ateia & Kruschwitz, 2024).

CHAIN OF THOUGHT

To advance from replication to reasoning, Wei et al. (2022) have proposed a *chain of thought* (CoT) structure that utilizes a series of intermediate reasoning steps, improving the LLM's capacity for complex reasoning. This is achieved through CoT prompting, where the models employ few-shot prompting of the reasoning steps, aiding inference in complex questions. The prompts are triplets with ⟨input, CoT, output⟩, where the CoT is a series of reasoning steps leading to the output.

In effect, CoT enables the model to partition the problem into subproblems, effectively enriching the problem analysis with more tokens and thus applying more computation to it. This leads to improved performance as well as better explainability in the reasoning steps, as these steps can be viewed and evaluated by the user. Improved inference comes at the cost, however, of a greater consumption of resources, being computationally more intensive. While Wei et al. demonstrate a great performance boost for large models, they note that models smaller than 100 billion parameters would not benefit from a CoT approach. Hence, it can be argued that CoT approach and the resulting models, referred to as *large reasoning models* (LRMs), may remain unfeasible for low-resource on-premise or edge implementations unless new innovations achieve comparable performance in smaller models.

TEACHER-STUDENT KNOWLEDGE DISTILLATION

A very prominent and much-studied topic in the domain of GenAI focuses is the pursuit of large language models more computationally efficient in terms of inference and memory (Xu et al., 2024; Gupta & Agrawal, 2021). The proposed solutions include *model pruning*, *quantization*, and *knowledge distillation*. Through pruning, models are scaled down in ways that deliberately minimize effects on performance (removing low magnitude weights, for example); quantization eases the memory requirements by using lower bit size for the weights; "knowledge distillation" describes methods of transferring the performance of a large and widely capable model into a smaller model, creating a computationally feasible and easily deployable version with high performance metrics on the chosen field (Gupta & Agrawal, 2021).

The method of knowledge distillation using a teacher-student scheme, in which a smaller model learns from the outputs of a larger teacher model, was proposed by Ba and Caruana (2014). In essence, the smaller or shallower model learns to mimic the function of the deeper model, aiming to achieve similar performance with more attainable computational resources.

UNSUPERVISED REINFORCEMENT LEARNING

The AI landscape was shaken in late 2024 and early 2025 by the announcement of R1, developed by the Chinese company DeepSeek: a model with reported results comparable to top-tier models such as OpenAI-o1 (Guo et al., 2025). In their paper, Guo et al. showcase post-training as large-scale unsupervised reinforcement learning (RL) – a machine learning method in which a model learns to make decisions by interacting with an environment and receiving rewards or penalties based on its actions, without relying on supervised fine-tuning (SFT – a training process in which a pre-trained model is adjusted using labeled data to improve its performance). This is achieved by applying RL directly on to the V3 base model to create R1-Zero. The final production model, R1, used a multi-stage pipeline with cold-start SFT, reasoning-oriented RL, and rejection-sampled self-generated SFT. The demonstrated shift from all-human SFT is a major innovation, as it suggests that the laborious part of data preparation can be automated with a capable base model that can curate the data for fine-tuning without human involvement.

Guo et al. (2025) also demonstrate the distillation of the reasoning patterns of larger models into smaller ones, permitting performance and computational resources to be re-balanced in favor of performance. The distillation experiment was carried out by fine-tuning open-source models like Qwen and Llama, using 800,000 samples curated with DeepSeek-R1. From a

DOI: 10.31374/sjms.415

computationally restricted perspective, the performance of the smallest models, Llama 8B and Qwen 7B, are interesting: the benchmark results show that these distilled models are comparable to GPT-4o, Claude-3.5-Sonnet and even OpenAI o1-mini in performance, with some variance between benchmarks. The researchers note that only SFT was applied for distilled models and unsupervised RL could provide further, substantial boost in model performance.

TITANS ARCHITECTURE AND VARIANTS

Behrouz et al. (2024) have recently introduced a group of novel architectures named Titans, which aim to mimic the memory of humans with varying implementations of certain core ideas. In essence, the Titans architecture introduces three types of memory: short-term, long-term, and persistent. Short-term memory in the architecture core consists of an attention mechanism and resembles human working memory; long-term memory, meanwhile, a deep neural memory, stores information over extended periods of time. The persistent memory acts as a meta-memory that ensures patterns and frameworks are held in the memory by learning data-independent meta-information. The persistent memory weights remain fixed during test-time as these are responsible for storing the knowledge about the task, whereas neural memory learns during test time and working memory (core) uses in-context learning, enabled by the attention mechanism associated with transformer architectures.

The major improvements of the proposed architectures are, as the paper's title *Titans: Learning to Memorize at Test Time* indicates, their ability to learn during test time (i.e., during inference) and to scale beyond a two-million-token context window. The context window refers to the amount of input that can be processed within a single input–output interaction.

The neural long-term memory utilizes a surprise metric, which is estimated with the impact of the gradient so that a larger gradient indicates greater difference from the previous data. This surprise metric is divided into past surprise and momentary surprise to take both the recent past and the latest incoming data into account. Likewise, architecture incorporates a forgetting mechanism to forget past data that is no longer relevant.

The result is a more effective architecture with scalability beyond a two-million-token context window size with performance above baselines. The proposed architectures excel in handling long sequences, due to the surprise metric, and introduce test-time learning into the current state-of-the-art. The implications are extensive, as the proposed architectures may have tremendous impact for military applications: test-time learning may improve domain-specific tuning in use without standard approach of training and deployment.

AGENTIC AI

In a recent development, the AI models crafted and trained with the methods described above are turned into "agents" that can execute tasks unsupervised (Shavit et al., 2023). While OpenAI engineers define the degree of agentic AI in terms of goal complexity, environmental complexity, adaptability, and independent execution, the primary characteristic is the ability to "take actions which consistently contribute towards achieving goals over an extended period of time, without their behavior having been specified in advance" (Shavit et al., 2023). The key idea is to automate certain workflows to enhance the productivity and performance of the complete process, in which the AI model functions more like a human worker instead of a component within the workflow.

SUMMARY

This section traced the evolution of GenAI, grounding the current landscape in the foundational transformer architecture. The state of the art is characterized by a dynamic tension between two competing trends. On one hand, innovations like mixture of experts (MoE) and chain of thought (CoT) are pushing the boundaries of model scale and reasoning capability, primarily in large, resource-intensive settings. On the other hand, a parallel and equally important trend focuses on efficiency, accessibility, and adaptation. Techniques such as knowledge distillation are enabling the creation of smaller, specialized models that retain a meaningful level of performance compared to their larger counterparts but are feasible for deployment in constrained environments.

More recent breakthroughs promise to accelerate this trend. Unsupervised reinforcement learning offers a path to automate and scale model fine-tuning without laborious human oversight, while novel architectures like Titans introduce the capacity for models to learn continuously during use ("test-time learning").

Together, these advancements create a technological landscape of potential for the use of agentic AI, and the utilization of GenAI in military. The following literature review analyzes the extent to which these state-of-the-art capabilities have been translated into practical military applications, exploring the opportunities and the significant challenges that arise when applying these powerful but complex technologies in a defense context.

LITERATURE REVIEW

Table 1 displays all 29 articles classified into the categories according to the type presented in the Methods section. The majority of these papers go into the categories of survey and application, while policy analysis, proposition, and overview papers are present in near equal proportions. The "other" category, consisting of scarce methodologies, includes a case study and a policy analysis.

Table 1 Articles Analyzed in the Literature Review.

TYPE	COUNT	PAPERS
Survey	8	Andreoni et al. (2024), Geist et al. (2024), Liu G. et al. (2024), Stebbins et al. (2024), Feffer et al. (2024), Moy and Gradon (2023), Huang et al. (2025), Mikhailov (2023)
Proposition	4	Black et al. (2024), Oniani et al. (2023), Lee et al. (2025), Tian et al. (2025)
Overview	3	Brearcliffe et al. (2023), Kelly and Smith (2024), Rashid et al. (2023)
Application	10	Hua et al. (2024), Goecks and Waytowich (2024), Chuang and Cheng (2022), Liu X. et al. (2024), Liu T. et al. (2025), Barzyk et al., (2024), Lee et al. (2023), Ruiz and Sell (2024), Rivera et al (2024), Lin et al. (2024)
Review	2	Zarrar and Kakar (2024), Caballero and Jenkins (2025)
Other	2	Marcellino et al. (2023), Beauchamp-Mustafaga et al. (2024)
Total	29	

While the survey papers provide valuable insight into previous research, current trends and future prospects, this analysis focuses on the application studies as these are hypothesized to provide concrete examples of both challenges and opportunities in adopting GenAI into military use.

Table 2 lists the application studies with general topics and key contributions; proposition papers are briefly analyzed, as these papers address a problem with a proposition that, hypothetically, ought to give insight into future applications.

Table 2 GenAI Application (Type A) and Proposition (Type P) Papers.

PAPER	TYPE	TOPIC	KEY CONTRIBUTION
Lin et al. (2024)	A	Decision making and decision support	Historic battle analysis with LLM multiagent simulations
Hua et al. (2024)	A	Decision making and decision support	Historic strategic international conflict analysis with LLM multiagent simulations
Goecks and Waytowich (2024)	A	Decision making and decision support	Enhancing course of action (COA) generation with LLMs
Rivera et al. (2024)	A	Decision making and decision support	Examining escalation risks associated with LLM use in decision making
Chuang and Cheng (2022)	A	Decision support	Conversational AI systems for military training using intent detection and response generation techniques
Liu X. et al. (2024)	A	Information extraction and fine-tuning	LLM approach for extracting military equipment entities from unstructured text to build a military knowledge base
Barzyk et al. (2024)	A	Cybersecurity	GenAI methodology for automating data tagging in military zero trust architecture cybersecurity frameworks

(Contd.)

PAPER	TYPE	TOPIC	KEY CONTRIBUTION
Lee et al. (2023)	A	Decision making and decision support	Proposes Deep AI Military Staff (DAMS), a multi-agent AI system for battlefield decision- making and introduces Multi-Agent Collaboration Architecture to enhance situational awareness
Ruiz & Sell (2024)	A	Information extraction and fine-tuning	Presents TRACLM, a fine-tuned Large Language Model (LLM) developed for the U.S. Army to improve AI-driven decision-making and intelligence analysis.
Liu T. et al. (2025)	A	Cybersecurity	The use of GenAI for enhancing cross-layer covert communication in military networks.
Black et al. (2024)	P	Strategic advantage	Proposes a strategic framework that positions GenAI within the broader context of military competition
Oniani et al. (2023)	P	Ethical principles	Proposes ethical cross-domain principles for transparency, value-alignment and accountability
Lee et al. (2025)	P	Collaboration	Proposes a system-level architecture using FL as a collaborative framework for LLM training for allied nations.
Tian et al. (2025)	P	Unmanned systems	Explores the integration of LLMs and LVMs into UAVs

IN-DEPTH ANALYSIS OF GENAI APPLICATIONS AND PROPOSITIONS

The papers presented in Table 2 show that a majority of the applications are focused on decision making and decision support. The second group goes under cybersecurity with two papers on par with fine-tuning and information extraction.

DECISION MAKING AND DECISION SUPPORT

In the process of decision making, a choice is made according to constraints presented by the environment, available information, and cognitive capacity (Simon, 1977). Balis and O'Neill (2022) have noted that while AI has had a profound impact on military decision making, it has attracted little attention beyond specialists. To enhance decision making, decision-support systems (DSS) have been devised (Sprague, 1980), recently integrating AI into the military decision-making process (MDMP; see Nadibaidze et al., 2024).

Goecks and Waytowich (2024) have researched the use of GenAI in accelerating course of action (COA) development in military operations. The COA development is a critical component in most MDMPs, where several different COAs are developed for comparison to enhance decision making with regard to the situation and the objective. The research utilizes the Operation TigerClaw (Narayanan et al., 2021) scenario and StarCraft II learning environment (Vinyals et al., 2017) as the setting and environment, and OpenAI's GPT-4-Turbo as well as GPT-4-Vision which are prompted to act as a COA-GPT ("generative pre-trained transformer"). The research indicates that using LLMs for COA generation accelerates the process substantially, outperforming reinforcement learning baselines, with better correlation to the commander's intent. However, expert human scores are still comparable or superior to the baselines. The most striking difference is in friendly forces casualties, in which both textual and visual COA-GPT incur a far greater number of casualties.

Rivera et al. (2024) have studied the escalation risk of LLMs in military and diplomatic decision making. The research setting puts eight LLMs as nation agents in a simulation where the models engage in military and foreign policy. The LLMs are selected from five publicly available models – GPT-3.5, GPT-4, Claude 2.0, Llama-2-Chat and GPT-4-Base. The reported results indicate both that the models are more prone to escalate than de-escalate a situation, and show individual models to differ in their profiles and predictability. It was also shown that without instructions training, a base model will choose the most severe actions more often than other models – but it should be noted that none of the models were tuned in any way to follow a military or foreign policy doctrine, acting, instead, on the basis of their training data and system prompts.

In a similar setting, Hua et al. (2024) have researched an LLM-based multi-agent system named WarAgent in simulating historical events. The simulation used GPT-4 models as country agents, with secretary agents as safeguards to avoid the impact of fallacies. The agents were prompted to act according to specific roles. To evaluate the inevitability of war the models were also tuned with counterfactual information to avoid mere replication of history. The results suggested that avoiding war depends on changes in national policies.

DOI: 10.31374/sjms.415

In parallel with this strategic-level point of view, Lin et al. (2024) created BattleAgent, combining a large vision language model (VLM) and a multi-agent system to simulate historical battles to increase understanding. The setting zooms from the macro-level approach by Hua et al. (2024) to the battlefield, with commander and soldier agents acting in coordination with their respective side against the similarly designed adversary. The approach was tested in emulating three historic battles; on average the results aligned with historical data, albeit with a notable variance between the models GPT-4, GPT-4-vision, and Claude-3. While both WarAgent and BattleAgent focused on testing and improving historic analysis, the approach and initial results demonstrated how GenAI can be leveraged in simulating warfighting outcomes on different levels, from a tactical setting to strategy and diplomacy.

Chuang and Cheng (2022) designed a task-oriented dialogue system for military scenarios. The system integrated slot-filling, intent detection, and retrieval-based answering into answer generation, using a Chinese corpus of military training missions. The methodology leverages natural language processing (NLP) and natural language generation (NLG) methods without LLMs. The proposed system extracted features for intent detection and created intent-entity relations for question generation. The results showed a query satisfaction greater than 80% over 8 scenarios after two rounds of dialog.

Lee et al. (2023) demonstrated a multi-agent-based collaboration architecture for manned and unmanned assets. The focus of the paper is in enhancing battlefield data processing and situation awareness with AI and automation. A part of the framework uses a conditional variational auto-encoder with a GAN to complete images of objects that are hidden or occluded. The results indicate that AI solutions, including narrow applications of GenAI, can make the battlefield considerably more transparent than it currently is.

CYBERSECURITY

Barzyk et al. (2024) have studied supporting zero trust architecture implementations with GenAI-automated data tagging, proposing a novel labeling tool to remove error-prone and time-consuming manual tagging. The zero-trust approach does not assume any part of the system to be a trusted zone, applying appropriate control measures to connections and data access throughout, involving data tagging. This means, for example, assigning classifications or other criteria for the data. The research group tested fine-tuning BERT (Devlin et al., 2019) and TinyLlama (Zhang et al., 2024) for a classification task to identify messages in categories. The results indicated that comparatively simple GenAI methods with modest fine-tuning afforded adequate results and that, therefore, offer a strong foundation to improve utilization of zero trust architectures.

Liu T. et al. (2025) have researched GenAI in cross-layer covert communication, where the layers consist of physical, network, and application layers. They propose a GenAI framework in which a generative model is used for knowledge transfer for channel optimization, channel quality evaluation for communication path planning, and enhanced concealment for warden evasion. In this regard, the warden systematically collects and analyzes data to identify patterns of covert transmissions. The applicatory part of the study compared diffusion RL as diffusion soft actor-critic (DSAC) to the traditional soft actor-critic (SAC) approach, providing numeric results that showcased both training efficiency and performance advantage for the diffusion-based approach. However, the DSAC used is not described in detail and the experimental setup description leaves room for interpretation.

INFORMATION EXTRACTION AND FINE-TUNING

Liu X. et al. (2024) have researched military equipment entity extraction using LLMs, using what they call *chain-of-thought named entity recognition* (CoTNER). The research highlights difficulties presented by military terminology in contextual analysis and the data requirements and computational burdens of very large models. As a proposition, the researchers use GPT-3.5 for data-augmentation with chain-of-thought combined with high-quality data filtering in order to fine-tune a relatively small Llama-3-8B-Instruct model. In comparison with models such as BERT-MRC (Li et al., 2019) and GPT-NER (Wang et al., 2023), the proposed CoTNER showcases better performance in all metrics. In a way, the data-augmentation part can be viewed as knowledge distillation, as the GPT3.5-175B is considerably larger LLM used to produce the fine-tuning data for the eight billion parameter Llama3 model.

Ruiz and Sell (2024) have researched fine-tuning open-source LLMs for the army domain. While their experimental fine-tuning was performed on small, 3–7 billion scale LLMs, novel methods show promise for expansion of local fine tuning up to and beyond 70 billion parameter models. The best results were received when the fine-tuning was enhanced with a synthetic instruction tuning data set (knowledge distillation, created with the Mistral-8x-7B-Instruct-v0.1 model). A key takeaway regarding domain specific considerations was the relation of tokenization and contextual connections that result in domain-specific knowledge.

PROPOSED FRAMEWORKS, CONCEPTS AND ARCHITECTURES

In addition to technical applications, a subset of the reviewed literature focuses on the development of conceptual frameworks and architectures aimed at guiding the integration of GenAI in military contexts. These works are essential for setting strategic direction, enabling interoperability, and addressing ethical and organizational considerations beyond model performance alone.

While not solely GenAI focused, Black et al. (2024) propose a strategic framework that positions GenAI within the broader context of military competition, emphasizing the need for national policies that anticipate the disruptive nature of AI-enabled capabilities. Their work outlines how strategic advantages may stem not only from technological superiority but also from organizational adaptation and ethical governance.

Oniani et al. (2023) contribute a conceptual model for cross-domain ethical principles, drawing parallels between GenAI use in military and healthcare domains. Their framework emphasizes the importance of transparency, accountability, and value alignment in AI-enabled decision-making – principles that are especially relevant in high-stakes environments such as conflict scenarios and lethal autonomous systems.

Lee et al. (2025) explore a system-level architecture for federated learning (FL) (McMahan et al., 2017; Ji et al., 2024) among allied nations, proposing a collaborative framework for secure training of military LLMs. The use of FL as a collaborative framework for LLM training is interesting and has many upsides for military applications, including reduced communication overheads and enhanced data privacy (Sani et al., 2024, Iacob et al., 2024). The work of Lee et al. (2025) highlights the vulnerability of GenAI systems to prompt injection attacks,[1] especially in decentralized training environments, and offers mitigation strategies to ensure robustness and trust. This architecture represents an important conceptual advance in enabling secure distributed GenAI development within coalition contexts.

Tian et al. (2025) visualize LLMs along with vision foundation model (VFM)s and vision language models (VLM) as foundation models that can be leveraged for unmanned ariel vehicles (UAVs). The key difference from LLMs is that VLMs integrate textual and visual information, permitting operations such as visual reasoning, whereas VFMs have become a core technology in computer vision. In their work, Tian et al. (2025) examine the role and use of these foundation models and propose a framework for an agentic UAV capable of processing multimodal data from different sources to advance autonomy and swarm behavior through the cooperation of multiple agents. In essence, the paper maps current technologies to multimodal data sources and lays out a framework to create GenAI powered UAVs.

SUMMARY

In short, the ten analyzed GenAI application papers highlight two technical things: the use of proprietary LLMs in a surrogate task to experiment with and evaluate off-the-shelf capability; and the fine-tuning of smaller models into mission- and task-specific versions. While most of the papers had some decision-support point of view, it has also been argued that humans remain in a critical role in wargaming (Hinton, 2023), which is a key component of the military decision-making process.

While the relatively low number of cybersecurity topics is worth noting, most military cybersecurity research and development is, of course, not public, which may explain the discrepancy. Likewise,

1 A prompt injection attack is a security exploit where an adversary provides malicious input to an LLM to make it bypass its safety constraints or execute unintended commands (Perez & Ribeiro, 2022).

DOI: 10.31374/sjms.415

the number of increasingly complex and intelligent systems means that security is in increasingly high demand, and there exists plenty of research on GenAI cybersecurity topics (see, for example, Yigit et al., 2024) that, while not solely military-focused, are still applicable to military domain. In a similar manner, while the GenAI-specific algorithmic bias is not well covered, broader military AI bias has been researched (Bode and Bhila, 2024).

Together, these studies offer valuable building blocks for understanding how GenAI systems can be effectively integrated into defense infrastructure. However, a common theme across the literature is the lack of standardization in conceptual models and the absence of empirical validation for many of the proposed frameworks. As GenAI technologies evolve, there is a growing need for operationally grounded, interoperable, and security-conscious architectures that can guide real-world implementations across national and allied defense systems, as well as for military baselines and benchmarks that can be used to evaluate the performance of AI models and systems.

Out of the introduced state-of-the-art GenAI research, none except few-shot learning and agent-like approaches were experimentally present in the analyzed research papers as the focus of interest. Rather, certain technologies such as MoE and distillation have enabled a scaling up of the large, leading-edge proprietary models, indirectly influencing or enabling the analyzed application studies. Hence, it can be argued that the research cutting-edge, driven by industry, is yet to enter the military domain.

All in all, the analyzed research shows a tendency towards smaller models that underwent fine-tuning or the use of proprietary large models with system prompting to get favorable results. The state-of-the-art research from the past years enables both the scaling up of large models as well as size-reduction through distillation whilst maintaining performance. Known as the *scaling law*, performance increases linearly when model size increases exponentially (Matarazzo & Torlone, 2025). Proprietary models are driving the development, as they exhibit best performance and function as the stepping stone towards capable smaller models. The immense data and computation requirements make experimenting with in-house base models difficult (Matarazzo & Torlone, 2025). The resource-intensive nature of GenAI entrenches the importance of large cloud service providers that have, by the nature of their business, significant computation and memory capacity and large amounts of data. As an alternative, open-weight models can be leveraged for domain-specific purposes with smaller computational resources while reaching similar performance, but with considerably more effort, especially with regard to fine tuning. The challenges of military data sets (Rettore et al., 2024), military-tuned models, benchmarks, collaboration and frameworks hinders evaluation, validation, and incremental research and development.

The implications of frameworks greatly extends the conceptual scope of certain key issues. The need for things such as governance frameworks is evident in high-stakes environments, as the applicatory solutions need to be rooted at higher levels of guidance, conduct of operations, and doctrine. The strategic impact (Black et al., 2024), ethical framework (Oniani et al., 2023) and secure collaboration (Lee et al., 2025) serve as the fundamentals for the ground level capabilities such as unmanned systems (Tian et al., 2025) to be deployed into.

CURRENT STATE, TRENDS AND CHALLENGES

The landscape of GenAI research over the past two to three years reflects a dual trend: the continued scaling of large proprietary models and a growing ecosystem of smaller, open-weight models tailored for domain-specific use. As shown in Figure 2, state-of-the-art LLMs now span from massive, multimodal systems with tens or hundreds of billions of parameters to computationally cost-efficient, specialized models that can run in constrained environments. This trend is of particular interest to the military domain, where deployment constraints, security requirements, and operational reliability often preclude the use of cloud-based or opaque AI systems.

Despite major advancements in architectures (mixture of experts and Titans, for example), reasoning strategies (chain of thought, for example), and training paradigms such as unsupervised RL and distillation, the current research base reveals several important gaps:

DOI: 10.31374/sjms.415

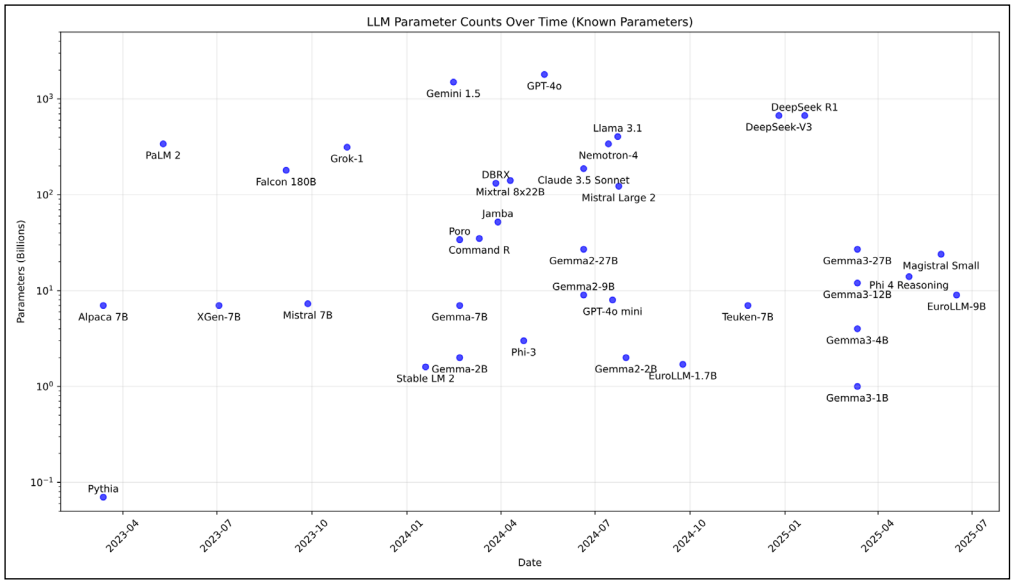

Figure 2 The known or estimated (uncertain) parameter counts for a selection of top models from 2023 to 2025 (Cardillo, 2025, Farabet & Warkentin 2025, LLM stats, 2025, Luukkonen et al., 2024, Ali et al. 2024, Martins et al., 2024, Martins et al., 2025).

- **Limited operational realism:** Most application studies rely on unclassified data, simulated settings, and public LLMs. There is a lack of experimentation with models trained on domain-specific or sensitive military corpora.
- **Lack of end-to-end system integration:** Research tends to focus on model capability (COA generation or information extraction, for example), but few studies demonstrate full system-level integration into command and control or intelligence, surveillance, and reconnaissance (ISR) workflows.
- **Underrepresentation of security and ethical risks:** While technical breakthroughs are covered, there is comparatively less focus on adversarial risks, governance frameworks, or explainability under operational pressure.

Despite these gaps, recent non-military-focused research has addressed several problems with GenAI, such as cognitive impairment (Kosmyna et al., 2025), explainability (Schneider, 2024), and the risks of agentic AI (Narajala & Narayan, 2025). The emergence of open-source ecosystems and the proliferation of modular, composable AI architectures present a tangible opportunity for military organizations to shape GenAI systems around their own operational needs. Techniques such as teacher-student distillation, federated training, fine-tuning, and test-time learning offer pathways to align GenAI performance with domain-specific constraints, particularly in low-resource environments or environments disconnected from information.

In general, the findings of this study illustrate a dynamic and rapidly evolving landscape in the application of GenAI to military contexts. While the potential of GenAI to transform key functions such as decision making, simulation, cybersecurity, and training is clear, the majority of the current research remains in exploratory or experimental stages. Most application studies reviewed rely on either publicly available large models or small-scale fine-tuned versions, typically operating outside of classified or high-assurance environments.

Importantly, the analysis highlights a disconnect between state-of-the-art GenAI capabilities and their practical military adoption. Breakthroughs such as mixture of experts architectures, unsupervised reinforcement learning, and scalable long-context models are primarily driven by commercial actors, and their integration into defense systems is slowed by operational, ethical, regulatory, and classification barriers. For example, Garcia (2024) notes that lack of international law on AI and algorithmic decision making compounds the existing ethical challenges.

While the research community is making rapid progress on model architecture and training efficiency, the military application of these innovations remains in its early stages. The limited availability of secure, in-domain datasets and systems further restricts experimentation with powerful LLMs in high-stakes scenarios. A key insight from the analysis is the strong orientation toward decision-support systems, which aligns with broader trends in both civilian and defense AI. GenAI shows considerable promise in accelerating COA generation, simulation and wargaming, information extraction, and situational awareness enhancement. However, as demonstrated in multiple studies, the outcomes generated by large language models (LLMs)

DOI: 10.31374/sjms.415

are still vulnerable to hallucinations, unintended escalation behaviors, and unpredictable outputs – particularly when models are not aligned with military doctrine or contextual nuance.

It can be argued that proprietary models, notwithstanding their impressive benchmark results, are not suitable for military use as an off-the-shelf product. This is demonstrated in both the escalation behavior that may stem from the disconnect between the training data and the attended use case in Rivera et al. (2024), as well as the COA-GPT (Goecks & Waytowich, 2024). The research highlights the issues with very large proprietary models, mainly the lack of explainability due to the vast size of the model, undisclosed model weights, and training data, which are linked to the expected behavior. Models of this size and scale are bound to have irregularities when applied to a specific, possibly ill-suited domain and its constraints, such as the military context. As an example, we can hypothesize that the escalation behavior may stem from the training data, as the public internet can be argued to be filled with input-output pairs that are more aggressive than actual military doctrines.

As a countertrend, the adoption of smaller, fine-tuned models is emerging as a pragmatic approach for military deployment. These models offer the benefits of interpretability, deployability in secure or constrained environments, and lower computational demands. Yet, fine tuning alone does not solve challenges related to contextual reasoning, long-term memory, or instruction-following. The performance of smaller models falls short, and the general applicability is naturally lower. Recent advances such as chain of thought prompting, knowledge distillation, and test-time learning offer promising pathways – but these methods remain underutilized in military settings, often due to resources, data (Menthe et al., 2024), or policy limitations. This disconnect between the model performance, explainability and suitability for a particular task in a constraint environment prompts a need for a larger, military specific base-model that can be used as both the military equivalent of large proprietary models and as a distillation base model for mission-specific use cases. Hence, the bottlenecks are data availability and computation.

Lee et al. (2025) note a potential solution for the creation of credible GenAI capability for allied nations in their proposition for a framework to secure system-level federated learning architecture used to train LLMs in collaboration. The FL paradigm would solve several issues, as it promotes privacy and the ownership of data, reduces communication overheads, and enables tuning for user-specific needs. As a result, allied nations would have their own local LLMs trained on local and possibly sensitive military data, and only model parameters would be shared, with differential privacy, to the global model for system-level updates. In the FL scheme, the global model is updated according to local model inputs and the global parameters are then distributed back to local use, without sharing data itself. This approach may eventually permit credible architecture to enhance GenAI capabilities while simultaneously dealing with certain military limitations.

The analysis thus points out two major obstacles in successful GenAI adaptation to military use: reliance on proprietary models, and a lack of resources, architecture and infrastructure. Even those open-weight models fine-tuned for military use originate either in industry or academia. Besides the AI behemoths in the technology industry, smaller models have been designed and trained by public instances, including EuroLLM (Martins et al., 2025) and Teuken (Ali et al., 2024). The military AI capability driven solely by academia or industry cannot be deemed acceptable in the long term. To design and train, for example, a NATO base model family, the military alliance requires resources, infrastructure and cooperation with industry and academia. Here "resources and infrastructure" denote the data, computation and personnel; "architecture" denotes the way these are connected to enable training and fine-tuning "military-grade" AI models. Bridging this gap requires a deliberate shift toward applied experimentation, secure deployment environments, and closer collaboration between AI researchers, defense technologists, and policy stakeholders.

Finally, while ethical concerns, human-machine teaming, and governance frame- works are acknowledged in the literature, they appear underrepresented in empirical application studies. This gap underscores the need for multidisciplinary collaboration that goes beyond technical validation to include normative, strategic, and policy-oriented perspectives. The role of NATO and allied defense organizations in setting standards, enabling trusted infrastructure, and coordinating joint research efforts will be critical in closing this gap.

DOI: 10.31374/sjms.415

DISCUSSION

The focus on publicly available research presents a clear limitation of this study. Further research and development conducted by individual nations and coalitions lie beyond the scope of the present analysis. While this limitation is undeniable, it can also be argued that the pace of innovation is primarily set by industry. The cutting edge of research is typically disseminated through open-access academic publications, with observable trends gradually moving from civilian to military applications, as documented in this paper.

Although not exhaustive, this study highlights evidence-based trends and challenges at the structural level related to the military adaptation of GenAI. The current rate of advancement in GenAI appears to divide stakeholders into two groups: early adopters and onlookers. It can be hypothesized that the current craze for GenAI may hinder the adaptation of current state-of-the-art due to the promise of what might come next. Some research already considers the possible invention of artificial general intelligence (AGI; Black et al., 2024), which can be perceived in two ways. AGI may be seen as a strategic capability requiring further effort if it is to be attained; or it may be held to be an inevitable outcome requiring that we simply wait. As development is already driven primarily by industry, this study does not advocate waiting for future technological advancements. Instead, it encourages the creation and development of functional frameworks and architectures, the accumulation of high-quality data, and the establishment of mission-relevant benchmarks to facilitate the adaptation of GenAI and other AI methods sooner rather than later.

CONCLUSION

This study has presented a structured review and analysis of current research on the use of GenAI in military applications, covering 29 publications from 2022 to early 2025. The literature reveals both significant potential and clear limitations in the application of GenAI across defense-related domains such as decision making, cybersecurity, training, simulation, and information processing.

A key finding is that while large language models offer valuable capabilities, especially in accelerating planning and improving information synthesis, their use in mission-critical military environments remains constrained by challenges in trust, reliability, and operational security. Many current applications rely on either general- purpose public models or fine-tuned small-scale models, with limited alignment to military doctrine, context, or classification requirements.

At the same time, recent advances in GenAI such as chain of thought prompting, knowledge distillation, unsupervised reinforcement learning, and architectures enabling test-time learning are rapidly expanding the technological frontier. These innovations offer viable strategies to bridge the gap between model performance and deployability in defense contexts. However, these methods have yet to be widely adapted for military use, particularly in secure or resource-constrained environments.

To realize the full potential of GenAI in the military domain, several priorities emerge as recommendations to stakeholders. Foremost, operationalization of AI governance frameworks needs to be instated to ensure responsible deployment, especially in high-stakes scenarios. The usage of GenAI capabilities necessitates development of human-AI teaming models that integrate explainability and oversight into decision-making workflows.

From a technical point of view, the adaptation requires developing and deploying secure and robust system architectures that enable and boost AI development and deployment at scale, including data pipelines and management. Fundamentally for GenAI in the military domain, the design and training of a family of military base models, mimicking the current field of proprietary models, should be a top priority for allied nations. This should be done in collaboration with trusted industry partners and academia and aligned with investment in the research and development of scalable, domain-specific fine-tuning pipelines for small and medium-sized LLMs that can operate in classified or disconnected environments. This technical endeavor should be compounded by strengthening of NATO and allied research collaborations to foster joint development, experimentation, and standard-setting for GenAI in defense and security.

The trajectory of GenAI in the military domain is promising but uneven. Continued research and cross-sector coordination are essential for the move from theoretical promises to robust, mission-ready capabilities. The growing body of literature provides a strong foundation for future work that integrates technical, ethical, strategic, and operational perspectives into a cohesive vision for AI-enabled defense.

ABBREVIATIONS

AI – artificial intelligence

COA – course of action

CoT – chain of thought

GAN – generative adversarial networks

GenAI – generative artificial intelligence

FL – federated learning

LLM – large language model

LRM – large reasoning model

LSTM – long short-term memory

MDMP – military decision-making process

MoE – mixture of experts

NLG – natural language generation

NLP – natural language processing

RL – reinforcement learning

RNN – recurrent neural network

SFT – supervised fine-tuning

VFM – vision foundation model

VLM – vision language model

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Lauri Vasankari orcid.org/0009-0007-0107-3809
61N Solutions Oy, University of Turku, Finland

Aapo Koski orcid.org/0000-0002-4000-8491
61N Solutions Oy, The Faculty of Information Technology and Communication Sciences, Tampere University, Kalevantie 4, Tampere, 33100, Finland

REFERENCES

Ali, M., Fromm, M., Thellmann, K., Ebert, J., Weber, A. A., Rutmann, R., Jain, C., Lübbering, M., Steinigen, D., Leveling, J., Klug, K., Schulze Buschhoff, J., Jurkschat, L., Abdelwahab, H., Stein, B. J., Sylla, K.-H., Denisov, P., Brandizzi, N., Saleem, Q., Bhowmick, A., Helmer, L., John, C., Ortiz Suarez, P., Ostendorff, M., Jude, A., Manjunath, L., Weinbach, S., Penke, C., Filatov, O., Asaadi, S., Barth, F., Sifa, R., Küch, F., Herten, A., Jäkel, R., Rehm, G., Kesselheim, S., Köhler, J., & **Flores-Herr, N.** (2024). *Teuken-7B-Base & Teuken-7B-Instruct: towards European LLMs*. arXiv. https://doi.org/10.3233/FAIA251328

Andreoni, M., Lunardi, W. T., Lawton, G., & **Thakkar, S.** (2024). Enhancing autonomous system security and resilience with Generative AI: A comprehensive survey. *IEEE Access*, *12*, 109470–109493. https://doi.org/10.1109/ACCESS.2024.3439363

Ateia, S., & **Kruschwitz, U.** (2024). *Can open-source LLMs compete with commercial models? Exploring the few-shot performance of current GPT models in biomedical tasks*. arXiv. https://arxiv.org/abs/2407.13511

Ba, L. J., & **Caruana, R.** (2014). *Do deep nets really need to be deep?* arXiv. https://arxiv.org/abs/1312.6184

Balis, C., & **O'Neill, P.** (2022). *Trust in AI: Rethinking future command* (Occasional paper). Royal United Services Institute for Defence and Security Studies.

Barzyk, C., Hickson, J., Ochoa, J., Talley, J., Willeke, M., Coffey, S., Pavlik, J., & **Bastian, N. D.** (2024). A generative artificial intelligence methodology for automated zero-shot data tagging to support tactical zero trust architecture implementation. *Proceedings of the Annual General Donald R. Keith Memorial Conference*. https://doi.org/10.37266/ISER.2025v12i2.pp83-88

Beauchamp-Mustafaga, N., Green, K., Marcellino, W., Lilly, S., & **Smith, J.** (2024). *Dr. Li Bicheng, or how China learned to stop worrying and love social media manipulation* (Technical report). RAND Corporation.

Behrouz, A., Zhong, P., & **Mirrokni, V.** (2024). *Titans: Learning to memorize at test time*. arXiv. https://arxiv.org/abs/2501.00663

Bishop, C. M., & **Bishop, H.** (2023). *Deep learning: foundations and concepts*. Springer Nature. https://doi.org/10.1007/978-3-031-45468-4

Black, J., Eken, M., Parakilas, J., Dee, S., Ellis, C., Suman-Chauhan, K., Bain, R., Fine, H., Aquilino, M. C., Lebret, M., & **Palicka, O.** (2024). *Strategic competition in the age of AI: Emerging risks and opportunities from military use of artificial intelligence*. RAND.

Blanchard, A., & **Bruun, L.** (2024). *Bias in military artificial intelligence*. SIPRI Publications. https://doi.org/10.55163/CJFT9557

Bode, I., & **Bhila, I.** (2024, September 3). The problem of algorithmic bias in AI-based military decision support systems. *Humanitarian Law and Policy*.

Boone, T., Fahimnia, B., Ganeshan, R., Herold, D. M., & **Sanders, N. R.** (2025). Generative AI: Opportunities, challenges, and research directions for supply chain resilience. *Transportation Research Part E: Logistics and Transportation Review*, *199*, 104135. https://doi.org/10.1016/j.tre.2025.104135

Brearcliffe, D. K. (2023). *Computational military science: Complexity, broken symmetry, and artificial intelligence* [Unpublished doctoral dissertation]. George Mason University.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Girish, S., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paperfiles/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Caballero, W. N., & **Jenkins, P. R.** (2025). On large language models in national security applications. *Stat*, *14*(2), e70057. https://doi.org/10.1002/sta4.70057

Cardillo, A. (2025). *Best 39 large language models (LLMs) in 2025*. Exploding Topics. https://explodingtopics.com/blog/list-of-llms

Chuang, H. M., & **Cheng, D. W.** (2022). Conversational AI over military scenarios using intent detection and response generation. *Applied Sciences*, *12*(5), 2494. https://doi.org/10.3390/app12052494

Devlin, J., Chang, M. W., Lee, K., & **Toutanova, K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

Farabet, C., & **Warkentin, W.** (2025, May 12). *Introducing Gemma 3: The most capable model you can run on a single GPU or TPU*. Google. https://blog.google/technology/developers/gemma-3/

Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & **Heidari, H.** (2024). Red-teaming for generative AI: Silver bullet or security theater? In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 7, pp. 421–437). ACM. https://doi.org/10.1609/aies.v7i1.31647

Garcia, D. (2024). Algorithms and decision-making in military artificial intelligence. *Global Society*, *38*(1), 24–33. https://doi.org/10.1080/13600826.2023.2273484

Geist, E., Frank, A. B., & **Menthe, L.** (2024). *Limits of artificial intelligence for warfighters* (Technical report). RAND Corporation.

Goecks, V. G., & **Waytowich, N.** (2024). COA-GPT: Generative pre-trained transformers for accelerated course of action development in military operations. In *2024 International Conference on Military Communication and Information Systems (ICMCIS)* (pp. 1–10). IEEE. https://doi.org/10.1109/ICMCIS61231.2024.10540749

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & **Bengio, Y.** (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27, pp. 2672–2680). Curran Associates, Inc.

Grand-Clément, S. (2023). *Artificial intelligence beyond weapons: Application and impact of AI in the military domain*. UNIDIR.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., & Zhang, Z. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://doi.org/10.1038/s41586-025-09422-z

Gupta, M., & **Agrawal, P.** (2021). *Compression of deep learning models for text: A survey*. arXiv. https://arxiv.org/abs/2008.05221

Hinton, P. (2023). Generative AI and wargaming: What is it good for? *The RUSI Journal*, *168*(7), 34–41. https://doi.org/10.1080/03071847.2023.2282863

Hochreiter, S., & **Schmidhuber, J.** (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hua, W., Fan, L., Li, L., Mei, K., Ji, J., Ge, Y., Hemphill, L., & **Zhang, Y.** (2024). *War and peace (WarAgent): Large language model-based multi-agent simulation of world wars*. arXiv. https://arxiv.org/abs/2311.17227

Huang, Z., Zhang, X., Tang, Z., Xu, F., Datcu, M., & **Han, J.** (2025). Generative artificial intelligence meets synthetic aperture radar: A survey. *IEEE Geoscience and Remote Sensing Magazine*. https://doi.org/10.1109/MGRS.2024.3483459

Iacob, A., Sani, L., Marino, B., Aleksandrov, P., Shen, W. F., & **Lane, N. D.** (2024). Worldwide federated training of language models. *Proceedings of the FL@FM-NeurIPS*.

Ji, S., Tan, Y., Saravirta, T., Yang, Z., Vasankari, L., Pan, S., Guodong, L., & **Walid, A.** (2024). Emerging trends in federated learning: From model fusion to federated X learning. *International Journal of Machine Learning and Cybernetics*, *15*, 3769–3790. https://doi.org/10.1007/s13042-024-02119-1

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & **Fung, P.** (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kelly, P., & **Smith, H.** (2024). How to think about integrating generative AI in professional military education. *Military Review*.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., Braunstein, I., & **Maes, P.** (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task*. arXiv. https://arxiv.org/abs/2506.08872

Lee, C. E., Baek, J., Son, J., & **Ha, Y. G.** (2023). Deep AI military staff: Cooperative battlefield situation awareness for commander's decision making. *The Journal of Supercomputing*, *79*(6), 6040–6069. https://doi.org/10.1007/s11227-022-04882-w

Lee, Y., Park, T., Lee, Y., Gong, J., & **Kang, J.** (2025). *Exploring potential prompt injection attacks in federated military LLMs and their mitigation*. arXiv. https://arxiv.org/abs/2501.18416

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M. N., Yih, W., Rocktäschel, T., Riedel, S., & **Kiela, D.** (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paperfiles/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & **Li, J.** (2019). *A unified MRC framework for named entity recognition*. arXiv. https://doi.org/10.18653/v1/2020.acl-main.519

Lin, S., Hua, W., Li, L., Chang, C.-J., Fan, L., Ji, J., Hua, H., Jin, M., Luo, J., & **Zhang, Y.** (2024). *BattleAgent: Multi-modal dynamic emulation on historical battles to complement historical analysis*. arXiv. https://doi.org/10.18653/v1/2024.emnlp-demo.18

Liu, G., Van Huynh, N., Du, H., Hoang, D. T., Niyato, D., Zhu, K., Kang, J., Xiong, Z., Jamalipour, A., & **Kim, D. I.** (2 L. (2024). *Generative AI for unmanned vehicle swarms: Challenges, applications and opportunities*. arXiv. https://arxiv.org/abs/2402.18062

Liu, T., Liu, J., Zhang, T., Wang, J., Wang, J., Kang, J., Niyato, D., & **Mao, S.** (2025). *Generative AI-driven cross-layer covert communication: Fundamentals, framework and case study*. arXiv. https://doi.org/10.1109/MCOM.004.2500025

Liu, X., Yu, Z., Liu, X., Miao, L., & **Yang, T.** (2024). Military equipment entity extraction based on large language model. *Applied Sciences*, *14*(19), 9063. https://doi.org/10.3390/app14199063

LLM Stats. (2025). *LLM leaderboard 2025 – Verified AI rankings*. Retrieved July 2, 2025, from https://llm-stats.com/

Luukkonen, R., Burdge, J., Zosa, E., Talman, A., Komulainen, V., Hatanpää, V., Sarlin, P., & **Pyysalo, S.** (2024). *Poro 34b and the blessing of multilinguality*. arXiv. https://arxiv.org/abs/2404.01856

Marcellino, W., Beauchamp-Mustafaga, N., Kerrigan, A., Chao, L. N., & **Smith, J.** (2023). *The rise of generative AI and the coming era of social media manipulation 3.0: Next-generation Chinese astroturfing and coping with ubiquitous AI* (Technical report). RAND Corporation.

Martins, P. H., Alves, J., Fernandes, P., Guerreiro, N. M., Rei, R., Farajian, A., Klimaszewski, M., Alves, D. M., Pombal, J., Boizard, N., Faysse, M., Colombo, P., Yvon, F., Haddow, B., de Souza, J. G. C., Birch, A., & **Martins, A. F. T.** (2025). *EuroLLM-9B: Technical report*. arXiv. https://arxiv.org/abs/2506.04079

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J. P., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., C. de Souza, J., Birch, A., & **Martins, A.** (2024). *EuroLLM: Multilingual language models for Europe*. arXiv. https://doi.org/10.1016/j.procs.2025.02.260

Maslej, N., et al. (2024). Chapter 4: Economy. In *Artificial Intelligence Index Report 2024* (pp. 46–68). Stanford Institute for Human-Centered Artificial Intelligence.

Matarazzo, A., & **Torlone, R.** (2025). *A survey on large language models with some insights on their capabilities and limitations*. arXiv. https://arxiv.org/abs/2501.04040

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & **y Arcas, B. A.** (2017). *Communication-efficient learning of deep networks from decentralized data*. arXiv. https://arxiv.org/abs/1602.05629

Menthe, L., Zhang, L., Geist, E., Steier, J., Frank, A. B., van Hegewald, E., Briggs, G. J., Scholl, K., Ashpari, Y., & **Jacques, A.** (2024). *Limits of artificial intelligence for warfighters* (Technical report). RAND Corporation.

Mikhailov, D. (2023). *Optimizing national security strategies through LLM-driven artificial intelligence integration*. arXiv. https://doi.org/10.14293/PR2199.000136.v1

Moy, W. R., & **Gradon, K. T.** (2023). Artificial intelligence in hybrid and information warfare. In M. Cristiano, et al. (Eds.), *AI and international conflict in cyberspace* (pp. 48–74). RAND Corporation. https://doi.org/10.4324/9781003284093-4

Nadibaidze, A., Bode, I., & **Zhang, Q.** (2024). *AI in military decision support systems: A review of developments and debates* (Review). Center for War Studies, University of Southern Denmark.

Narajala, V. S., & **Narayan, O.** (2025). *Securing Agentic AI: A comprehensive threat model and mitigation framework for Generative AI agents*. arXiv. https://arxiv.org/abs/2504.19956

Narayanan, P., Vindiola, M., Park, S., Logie, A., Waytowich, N., Mittrick, M., Richardson, J., Asher, D., & **Kott, A.** (2021). *First-year report of ARL directors strategic initiative (fy20–23): Artificial intelligence (AI) for command and control (C2) of multi-domain operations (MDO)* (ARL-TR-9192). DEVCOM Army Research Laboratory.

Oniani, D., Hilsman, J., Peng, Y., Poropatich, R. K., Pamplin, J. C., Legault, G. L., & **Wang, Y.** (2023). Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digital Medicine*, *6*(1), 225. https://doi.org/10.1038/s41746-023-00965-x

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann. https://doi.org/10.1016/C2009-0-27609-4

Perez, F., & **Ribeiro, I.** (2022). *Ignore previous prompt: Attack techniques for language models*. arXiv. https://arxiv.org/abs/2211.09527

Radford, A., Metz, L., & **Chintala, S.** (2016). *Unsupervised representation learning with deep convolutional generative adversarial networks*. arXiv. https://arxiv.org/abs/1511.06434

Rashid, A. B., Kausik, A. K., Al Hassan Sunny, A., & **Bappy, M. H.** (2023). Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International Journal of Intelligent Systems*, *2023*(1), 8676366. https://doi.org/10.1155/2023/8676366

Rettore, P. H., Zißner, P., Alkhowaiter, M., Zou, C., & **Sevenich, P.** (2024). Military data space: Challenges, opportunities, and use cases. *IEEE Communications Magazine* (pp. 70–76). https://doi.org/10.1109/MCOM.001.2300396

Rivera, J. P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., & **Schneider, J.** (2024). Escalation risks from language models in military and diplomatic decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 836–898). ACM. https://doi.org/10.1145/3630106.3658942

Ruiz, D. C., & **Sell, J.** (2024). *Fine-tuning and evaluating open-source large language models for the Army domain*. arXiv. https://doi.org/10.48550/arXiv.2410.20297

Sani, L., Iacob, A., Cao, Z., Marino, B., Gao, Y., Paulik, T., Zhao, W., Shen, W. F., Aleksandrov, P., Qiu, X., & **Lane, N. D.** (2024). The future of large language model pre-training is federated. *Proceedings of the FL@FM-NeurIPS*.

Schneider, J. (2024). Explainable generative AI (GenXAI): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, *57*(11), 289. https://doi.org/10.1007/s10462-024-10916-x

Schneider, J. (2025). Mental model shifts in human-LLM interactions. *Journal of Intelligent Information Systems*. https://doi.org/10.1007/s10844-025-00960-6

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Vallone, A., Passos, A., & **Robinson, D. G.** (2023). *Practices for governing agentic AI systems*. OpenAI. https://openai.com/index/practices-for-governing-agentic-ai-systems/

Shayea, G., Zabil, M., Habeeb, M., Khaleel, Y., & **Albahri, A.** (2025). Strategies for protection against adversarial attacks in AI models: An in-depth review. *Journal of Intelligent Systems*, *34*(1), 20240277. https://doi.org/10.1515/jisys-2024-0277

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & **Dean, J.** (2017). *Outrageously large neural networks: The sparsely-gated mixture-of-experts layer*. arXiv. https://arxiv.org/abs/1701.06538

Simon, H. A. (1977). *The new science of management decision* (Rev. ed.). Prentice Hall.

Sprague, R. H. (1980). A framework for the development of decision support systems. *MIS Quarterly*, *4*(4), 1–26. https://doi.org/10.2307/248875

Stebbins, D., Girven, R. S., Parker, T., Deen, T., De Bruhl, B., Ryseff, J., Paige, J. W., Kleiman, A. Y., Bhatt, S. D., Sousa, É. M., Kepe, M., & **Fay, M.** (2024). *Exploring artificial intelligence use to mitigate potential human bias within US Army Intelligence Preparation of the Battlefield processes* (Report No. RRA2763-1). RAND Corporation.

Tian, Y., Lin, F., Li, Y., Zhang, T., Zhang, Q., Fu, X., Huang, J., Dai, X., Wang, Y., Tian, C., Li, B., Lv, Y., Kovács, L., & **Wang, F. Y.** (2025). UAVs meet LLMs: Overviews and perspectives toward agentic low-altitude mobility. *Information Fusion*, *122*, 103158. https://doi.org/10.1016/j.inffus.2025.103158

Uddin, M., Irshad, M. S., Kandhro, I. A., Alanazi, F., Ahmed, F., Maaz, M., Hussain, S., & **Ullah, S. S.** (2025). Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations. *Artificial Intelligence Review*, *58*(8), 1–39. https://doi.org/10.1007/s10462-025-11219-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & **Polosukhin, I.** (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., & **Schrittwieser, J.** (2017). *StarCraft II: A new challenge for reinforcement learning*. arXiv. https://arxiv.org/abs/1708.04782

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & **Wang, G.** (2023). *GPT-NER: Named entity recognition via large language models*. arXiv. https://arxiv.org/abs/2304.10428

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & **Zhou, D.** (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., & **Zhou, T.** (2024). *A survey on knowledge distillation of large language models*. arXiv. https://arxiv.org/abs/2402.13116

Yigit, Y., Buchanan, W. J., Tehrani, M. G., & **Maglaras, L.** (2024). *Review of Generative AI methods in cybersecurity*. arXiv. https://arxiv.org/abs/2403.08701

Zarrar, H., & **Kakar, S. A.** (2024). Generative artificial intelligence & its military applications by the US and China – lessons for south Asia. *Journal of Computing & Biomedical Informatics*.

Zhang, P., Zeng, G., Wang, T., & **Lu, W.** (2024). *TinyLlama: An open-source small language model*. arXiv. https://arxiv.org/abs/2401.02385

TO CITE THIS ARTICLE:
Vasankari, L., & Koski, A. (2025). GenAI in the Military: Trends and Opportunities. *Scandinavian Journal of Military Studies*, 8(1), pp. 416–434. DOI: https://doi.org/10.31374/sjms.415

Submitted: 04 April 2025
Accepted: 30 October 2025
Published: 24 November 2025

Scandinavian Journal of Military Studies is a peer-reviewed open access journal published by Scandinavian Military Studies.