Commentary: The Dangers of Overreliance on Generative AI in the CT Fight

By Nicholas Clark

The explosive rise in generative artificial intelligence (AI) use has sparked debate over its applicability in military domains such as counterterrorism (CT). This article critically evaluates the role of large language models (LLMs) in CT, arguing that their utility remains limited and potentially detrimental when applied indiscriminately. After providing a high-level overview of the mathematical foundations of LLMs, the article demonstrates how these tools can produce misleading or confidently incorrect outputs. Through case studies and empirical findings, this article underscores the cognitive risks of overreliance on AI in CT planning and intelligence operations, including reduced analytical engagement and inhibited creativity among operators. While generative AI may assist in automating routine tasks, it lacks the capacity for nuanced judgment, uncertainty quantification, and dynamic responsiveness critical to effective CT work. The article concludes by advocating for a shift in focus toward enhancing education in probabilistic reasoning, such as Bayesian inference, and building robust data governance infrastructures. Such foundational improvements are prerequisites for any effective or responsible integration of AI into CT domains.

According to Brad Lightcap, OpenAI's chief operating officer, the number of weekly users of ChatGPT has now surpassed 400 million, up from 30 million only two years ago.[1] Given this reality, coupled with the drum beat of constant news stories extolling the virtues of artificial intelligence (AI), it is natural to question whether the counterterrorism (CT) community should expand its use of generative AI in general and large language models (LLMs) in particular. Indeed, the common thought is that the use of these tools will allow organizations such as U.S. Special Operations Command (SOCOM) to gather and analyze large amounts of data.[2]

Yet, in practice, the actual utility of AI remains narrow, especially in high-stakes or variable environments. In operational planning and intelligence analysis, overreliance on algorithms risks thwarting creativity and hindering the intellectual growth that is a hallmark of organizations within SOCOM. This is not to say that generative AI does not have a role in these types of organizations; however, this article argues for a recalibration of AI deployment: focusing on narrow, clearly beneficial, public-interest uses while resisting the temptation to adopt AI indiscriminately or unquestioningly. Rather than a vast investment in generative AI tools, the CT community would benefit more from an increased educational investment in probabilistic reasoning and data governance.

This article provides a high-level overview of the math behind LLMs that will highlight the limitations of these algorithms. The article then discusses how LLMs could potentially be employed in CT operational planning and intelligence analysis, arguing that the tools may not be beneficial in many cases. The article concludes by discussing other areas that would be more beneficial for the CT community to focus on than generative AI.

High-Level Overview of the Math Behind LLMs

While readers may be tempted to avoid the mathematics behind the algorithms, it is only through having a basic understanding of what these models are doing that allows users to understand the limitations of the tools. The more mathematically inclined reader will note that what follows is certainly not a full treatment of the algorithms, but it is very easy for even those who have advanced degrees in computer science, statistics, or mathematics to get lost in the full architecture of the algorithms.

The key to understanding LLMs is that they are built off auto-regressive models; that is, the models provide probabilistic output based off the words that are provided into the prompt. Each word in the English language is a potential next word and the algorithm assigns probabilities to all of the corpus. The algorithm then returns the word with the highest probability. The whole process then restarts to generate the second word and so on.

As a toy example, consider the prompt to an LLM to "Provide the next word in the statement 'The quick brown fox jumps over the lazy…'". What the algorithm does is takes the statement and first converts it into a vector.[a] This is what is referred to as tokenization.[3] The algorithm then assigns each token a weight and uses the weights to provide a probabilistic output. For instance, if we ask ChatGPT to complete the statement "The quick brown fox jumps over the lazy",

a While technically the tokenization occurs on syllables or smaller aspects of a word, what follows is still generally correct and serves as a high-level example of what the black box is doing 'under the hood.'

COL(R) Nicholas Clark, Ph.D., is an Associate Professor in the Department of Mathematics at the University of St. Thomas (Minnesota). Prior to joining St. Thomas, COL(R) Clark served as an Associate Professor at the United States Military Academy at West Point from 2016 to 2024, where he founded and led the Applied Statistics and Data Science Program. In 2021, he created a curriculum in data literacy that is now the widest adopted program in the U.S. Army. Prior to his academic appointments, COL(R) Clark served as an intelligence officer for multiple units with U.S. Special Operations Command (SOCOM).

it would state that the word with the highest probability is 'dog' and give us this word as the answer. However, we note that 'under the hood,' the algorithm is evaluating all possible words and assigning each a probability. For instance, if instead we ask the algorithm to provide the top-five most likely words and associated probabilities, we would get:

Rank	*Word*	*Estimated Probability*
1	dog	~85%
2	cat	~5%
3	boy	~3%
4	man	~2%
5	cow	~1%

However, we are often looking for more than a single word response. This is where the idea of auto-regression comes into play. If, instead, we wanted the algorithm to provide the next three words that would come after the statement "The quick brown fox jumps over the lazy", we would get:

Rank	*Next Three Words*	*Estimated Probability*
1	dog. The	~65%
2	dog and ran	~10%
3	dog without stopping	~7%
4	dog, then	~6%
5	cat. The	~3%

Here, the algorithm first predicts the first word; the second word, then, is conditional upon the first word that the algorithm provided. That is, first the algorithm says that the most likely word is 'dog', then it pretends that we passed in the prompt, "provide the next two words that come after the statement 'The quick brown fox jumps over the lazy dog'", and it determined that the most likely next word would be "." It then repeats this and says that since we are starting a sentence, we would next expect to get "The" for the final word.

To see how these probabilities are calculated, we consider the first prompt, "Complete the statement 'the quick brown fox jumps over the lazy'". To provide a probabilistic output, the model has to know what 'right' looks like. To do this, the algorithms are trained on Common Crawl; books (fiction and nonfiction); Wikipedia; WebText (Reddit, forums, etc.); technical content, manuals, and examples from real-world use.

That is, the model essentially looks across these examples and sees what most people would put as the next word in this prompt. Here, essentially the model shows that around 85% of the times people type the words "the quick brown fox jumps over the lazy" the next word is 'dog'.

So, what can go wrong? Let's say we type in "The quick brown fox jumps over the lazy cat" into a Google search. Google's AI tool will state:

> *The phrase "The quick brown fox jumps over the lazy cat" is a common English pangram, meaning it includes all letters of the English alphabet. It is often used for testing typing and fonts.*

The problem is, this is just wrong. The statement is not a pangram; however, the algorithm stated that it was because overwhelmingly most of the time people type out "The quick brown fox jumps over the lazy", they are using the pangram. Therefore, the strength of those words overwhelms the word 'cat', essentially ignoring the fact that we prompted it with 'cat' rather than 'dog.'

To see what else can go wrong, consider the prompt "Provide the next 10 words after the statement, 'the quick brown fox jumps over the lazy'". We would get "dog and ran swiftly across the green grassy field." The issue is, without having user knowledge on what we expected, we would have no idea whether this was correct. The algorithm, though, does not provide any warning that it is much more certain that the one-word completion is 'dog' than it is the 10-word completion is 'dog and ran swiftly across the green grassy field.'

One final potential issue is in the training data itself. Often, when developers are dissatisfied with the output from the algorithm, they will up-weight, or down-weight, certain datasets. For instance, recently xAI felt that its algorithm was providing responses that were too 'politically correct.'[4] The developers, subsequently, gave more weight to training data that was not seen as 'politically correct.' This resulted in their algorithm overweighing conspiracy theories and provided vile, antisemitic responses.[5] The ability to fine-tune is a double-edged sword: It enables customization but also opens the door to dangerous distortions.

LLMs in CT Operational Planning

One potential use for LLMs in CT would be for an operational planner to use the tool to generate courses of action. One immediate concern, as discussed above, would be that the algorithm would provide nonsense. However, there are other reasons that this may not be beneficial for the CT community. Perhaps the largest concern is embodied in the quote by President Dwight Eisenhower, "Plans are worthless, but planning is everything."[6] The use of generative AI for operational planning may, in fact, make our planners worse by removing the real benefits of the planning process and limit the CT forces' ability to respond dynamically to branches and sequels.

To understand the risk, we must look at the recent study by Kosmyna et al. on what happens to the human brain when individuals use AI assistance such as ChatGPT for writing papers. The study took three groups and asked them to write essays using ChatGPT, Google search, and nothing at all. They then examined the brain activity of the users and found that those that used ChatGPT had the lowest brain engagement and consistently underperformed the other groups. Perhaps most disturbingly, over the course of the study, ChatGPT users got lazier with each subsequent essay, often resorting to copy-and-paste.[7]

While special operations forces are extremely selective, perhaps less appreciated is the growth of the operator or support personnel while they are assigned to a special operations unit. One of the tenants of special operations is that competent SOF cannot be created after an emergency.[8] This is due to the training and growth that is required by individuals after they have been selected. Reliance on generative AI may impede this growth and limit the intellectual development of both operator and support personnel.

Operational plans in CT, in particular, require creativity that likely would not be produced through generative AI. A hallmark of special operations is that they are granted greater license to innovate during ongoing operations.[9] While it certainly is possible

to create an LLM that integrates domain specific knowledge into an algorithm through fine-tuning existing tools,[10] a lesser appreciated aspect of human planning in special operations is that a planner knows when and where to be creative and when to rely on conventional military methodology. Further, recent research has shown that the creativity employed by generative AI is predictable rather than truly being innovative.[11]

Where generative AI may be of use in operational planning is through automating the routine tasks of order production. For example, within a CT operations order (OPORD), there are typically paragraphs that an operations officer may find themselves cutting and pasting from previous OPORDs. Paragraph completion and other tools that rely on generative AI may be of use in these instances, however the wholesale adaptation of LLMs inside of operations planning is likely to impede both individual and unit growth and also lead to adverse outcomes.

LLMs in Intelligence Operations

The military domain that is often seen as rife for improvement by the use of generative AI is intelligence. Articles such as the joint report by UNICRI and UNCCT on "Countering Terrorism Online with Artificial Intelligence" seem to highlight a multitude of ways that AI can assist in CT.[12] However, as the article mentions, here the term AI is misleading and, in fact, most of the algorithms discussed are widely known and rely on structured data that often is missing in CT operations. Where AI has the most potential in intelligence is automating the processing of raw information into structured data, commonly referred to as data engineering. Generative AI may assist intelligence analysts who have a background in programming, however this, too, may be problematic.

Articles on using AI in intelligence operations often cover everything from basic regression models to more advanced topics such as neural networks or generative adversary networks. However, it is important to note that these are not examples of generative AI. While algorithms such as those that underlie ChatGPT are based off neural networks, saying a neural network is a form of generative AI would be like saying an engine is a form of a car.

One of the difficulties, though, in using more advanced analytical techniques in intelligence operations is that they rely on having high-quality, curated datasets. If there are issues with the data, then we cannot create algorithms to fix this. When data is messy or observed imperfectly, then an advanced algorithm may just be providing a false level of certainty.

As an example, we recently created an algorithm to assist in automating the process of creating a gridded reference graphic, or GRG.[13] This relied on using a set of satellite images to predict where buildings and key road intersections would be. In training the algorithm, we discovered that depending on where in the world we were observing, the model would require very different weights. That is, if we relied on data from Europe, the model would perform horribly in North Africa. The difficulty, then, became in creating such a robust set of training data that the algorithms could be useful in multiple areas. However, the algorithm used here was not generative AI. Where an LLM, perhaps, could be useful would be in the coding up of the convolutional neural network (CNN)[14] that we used in this instance. This, though, is far from certain. Recent research suggests that in some instances, generative AI may actually slow down developers.[15]

One final caution for CT intelligence analysts tempted to use generative AI as part of their work process is that the algorithms do not quantify uncertainty. That is, while traditional statistics allow researchers to yield a range of plausible values, generative AI typically provides a single output, or multiple outputs, but does not quantify how certain they are in their response. This is problematic for intelligence analysts who typically get asked how certain they are in their analysis and are asked to provide likelihood assessments for a variety of outcomes.

"Instead of focusing on generative AI, the CT community would benefit from focusing on learning and applying statistical tools to data and to ensure that good data governance exists in order to standardize aspects of data engineering."

If Not Generative AI, Then What?

In general, LLMs excel in three broad categories: quickly creating coding demonstrations,[b] translating between different coding languages, and explaining and critiquing coding. However, these are not the areas where intelligence or operations experts within CT typically need help. In fact, creating demonstrations that are not necessarily scalable often are distractions from the day-to-day work that these professionals need to accomplish. Rather, where intelligence professionals need to better leverage data is in creating predictive analytics and quantifying uncertainty in their predictions. Therefore, instead of focusing on generative AI, the CT community would benefit from focusing on learning and applying statistical tools to data and to ensure that good data governance exists in order to standardize aspects of data engineering.

Of the multitude of potential quantitative methods that CT professionals might focus on, the community would benefit from an increased awareness of Bayesian methodologies. Unlike purely data-driven models, Bayesian approaches allow analysts to formally incorporate prior knowledge, whether derived from field experience, historical data, or expert judgment, into probabilistic frameworks. This capacity to combine prior beliefs with new evidence enables more nuanced and interpretable assessments of uncertainty, especially in complex, evolving, and highly fluid environments where data may be sparse or noisy. Such probabilistic reasoning should not be treated as a niche tool but rather integrated into the broader analytic training of intelligence professionals. Teaching Bayesian inference alongside more traditional statistical and algorithmic methods would empower analysts to make more transparent and defensible judgments about future risks.

Operations experts and leaders inside the CT community would also benefit from increased instruction in probabilistic reasoning including Bayesian techniques. A basic lack of understanding of the meaning of probabilities often results in command teams asking

b For example, writing Python code to automate a task or creating a dashboard to allow users to interface with data.

intelligence professionals to either quantify that which cannot be quantified or to adjust the data in order to meet prespecified conclusions. A stronger understanding of how probabilities can be used to weigh data with subject matter expertise would allow planners to better quantify risk and also help to focus reconnaissance tasks on refining uncertainty rather than on areas of highest threat. For example, we recently demonstrated that dynamically re-tasking unmanned aerial surveillance (UAS) to areas of higher uncertainty rather than to areas of highest threat, or flying in a predetermined pattern, allowed users to more quickly map out the entire region of nuclear contamination.[16]

Still, regardless of the algorithm or framework employed, one fundamental truth remains: High-quality, consistent data is indispensable. Even the most sophisticated methodologies will falter—potentially catastrophically—when applied to flawed or incomplete data. Thus, investments in data infrastructure and governance are not ancillary but central to any successful analytic strategy. Everyone within an organization should understand how data is structured, shared, and stored. Data standardization is instrumental for any organization to successfully operate, and far too often it is a lack of clearly enforced rules that limit an organization's ability to successfully incorporate data into their decision-making processes. Recent research has shown that even after hiring AI experts (and paying them well), organizations fail to gain insight from data as a result of not having a data-driven culture—a key component of which is a common understanding of data standards and organizational goals.[17]

Conclusion

Generative AI holds clear appeal for the CT community. But beneath the surface lies a host of unresolved concerns: opaque reasoning, unreliable creativity, biased training, and the erosion of essential human competencies. The promise of AI must be weighed against its risks, however—not just technical ones, but cognitive and operational as well. Prior to any use of AI, organizations should upskill their population in probabilistic reasoning and basic data governance. It is only after these are properly understood that an organization can fully recognize whether tools such as generative AI are appropriate for their formation.

A final concern for the CT community is that AI development is dominated by a handful of U.S. tech corporations: Microsoft (partnered with OpenAI), Google (DeepMind, Gemini), Amazon (Bedrock, CodeWhisperer), Meta (Llama), and a few smaller players. These firms leverage their access to data, compute infrastructure, and talent pipelines to consolidate market share, extract rents, and shape public policy. In order to use their tools, they likely will request access to the most sensitive CT data possessed by the U.S. government and seem unlikely to share the underlying architecture for their algorithms. This will create a situation where the U.S. government will continue to need to purchase products and services for algorithmic maintenance in order to use tools that the corporations hope will become integral components of both the intelligence and operational planning cycles. While this issue is not unique to AI technology, this is particularly pronounced with generative AI as once organizations have access to CT data, the relationship between the U.S. government and corporations will become one-sided as the U.S. government will have to rely on the computational resources provided by the corporations. Future data acquisition that the government may attempt to use as leverage in contract discussions will be of little value once the initial tranche is provided to train the models. To potentially mitigate this, acquisition professionals will need to ensure that contracts stipulate that models are firewalled off from the corporations and prevent the models from learning from the unique sources of data that have been ingested. However, this still requires an increase in education grounded in basic data analytic skills that are largely missing from military curricula.

In an environment where adaptability, innovation, and judgment can determine life or death, overreliance on generative AI may do more harm than good. Instead, this article advocates for an increase in education in Bayesian reasoning and data principles. In many cases, generative AI is a distraction and should only be used within a disciplined, use-case-driven approach, one that leverages narrow efficiencies while preserving the uniquely human strengths that remain irreplaceable in counterterrorism work. CTC

Citations

1 Allison Morrow and Lisa Eadicicco, "Grok antisemitic outbursts reflect a problem with AI chatbots," CNN, July 10, 2025.
2 Allyson Park, "Just In: SOCOM Using AI to Speed Up Acquisition Workflows," National Defense Magazine, May 6, 2025.
3 Pranjal Kumar, "Large language models (LLMs): Survey, technical frameworks, and future challenges," *Artificial Intelligence Review* 57:10 (2024).
4 Morrow and Eadicicco.
5 Ibid.
6 "Remarks at the National Defense Executive Reserve Conference," U.S. Presidency Project, n.d.
7 Nataliya Kosmyna et al., "Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task," arXiv.org, June 10, 2025.
8 Bryan D. Brown, "U.S. Special Operations Command – Meeting the Challenges

of the 21st Century," *Joint Forces Quarterly* 40 (2006).
9 Robert G. Spulak, "Innovate or Die: Innovation and Technology for Special Operations," *JSOU Report* 10:7 (2010).
10 Zirui Song et al., "Injecting domain-specific knowledge into large language models: A comprehensive survey," arXiv.org, *2025.*
11 Mark A. Runco, "AI can only produce artificial creativity," *Journal of Creativity* 33:3 (2023).
12 See "Countering Terrorism Online with Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia," United Nations Interregional Crime and Justice Research Institute and United Nations Office of Counter-Terrorism, 2021.
13 Samuel Humphries, Trevor Parker, Bryan Jonas, Bryan Adams, and Nicholas J. Clark, "A dual U-Net algorithm for automating feature extraction from satellite imagery," *Journal of Defense Modeling and Simulation* 18:3 (2021): pp. 193-205.
14 "Convolutional neural networks," IBM, n.d.
15 Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein, "Measuring the impact of early-2025 AI on experienced open-source developer productivity," arxiv.org, July 12, 2025.
16 Daniel Echeveste, Andrew Lee, and Nicholas Clark, "Using Spatial Uncertainty to Dynamically Determine UAS Flight Paths," *Journal of Intelligent & Robotic Systems* 101:76 (2021).
17 Thomas H. Davenport and DJ Patil, "Is data scientist still the sexiest job of the 21st century?" Harvard Business Review, July 15, 2022.