KNUDSEN INSTITUTE
for Applied Research on Surge Theory

What Large Language Models Can't Do (Yet)

A Practical Guide for DoD Leaders

Michael Morford

What Large Language Models Can’t Do (Yet)

A Practical Guide for DoD Leaders

Understanding the Invisible Limitations of ChatGPT, Ask SAGE, and Other Large Language Models (LLMs) in US Military Environments

---

Executive Summary

Large Language Models (LLMs) like ChatGPT and Ask SAGE are increasingly integrated into Department of Defense (DoD) operations, from drafting acquisition documents to summarizing intelligence reports. While these tools offer significant promise, their hidden limitations pose unique risks to defense workflows. This guide explains these risks in plain language, emphasizing why defense leaders must treat LLMs as productivity aids – not autonomous agents. Of all concerns, hallucinations – where the model invents plausible but false information – are the most dangerous. Leaders must understand the systemic nature of these risks to ensure responsible adoption.

At the Knudsen Institute, we work extensively with LLMs as part of our broader, hybrid AI architecture. We observe each of these failure modes through our development and deployment of our (non-LLM) AI technology platform for US DoD manufacturing solutions for newbuild and sustainment operations. In working with LLMs, we have developed our own RAG & RAFT. Retrieval-Augmented Generation (RAG) pipelines, which combine LLM output with authoritative document retrieval minimize hallucinations. Our Retrieval-Augmented Fact Transformation (RAFT) architecture – an advancement beyond traditional RAG – focuses on dynamically structuring and validating facts against controlled data sources before generation occurs. Even this level of operational awareness does not remove these challenges. Hence, it is critical for DoD leaders to understand that even these advanced architectures do not eliminate these limitations. They mitigate risk, but they cannot fundamentally change how LLMs function as pattern-based generators.

Case Example: The Fabricated Base Incident

In early 2025, a base operations plan for Pacific theater logistics included references to a non-existent Philippine naval facility, the "South Harbor Logistics Compact." This entirely fabricated content was generated by an LLM, confidently presented as fact, and inserted into the draft document without challenge. The hallucinated data survived multiple internal review cycles within the planning cell, partly due to the credible tone and formatting of the LLM-generated text. Ultimately, it was flagged not by internal reviewers but by an allied partner who questioned the facility’s existence. While this event did not result in operational harm, it illustrated the ease with which fabricated data can silently propagate into strategic-level planning products.

This incident was not a human oversight in isolation. It was an engineered failure mode of the LLM. No audit trail existed to explain why the hallucinated content appeared. The confident,

fluent style masked the error. Planners assumed the data was valid because the tool that generated it sounded credible. This example is not an outlier; it is an operational warning.

Why This Matters for DoD

The adoption of LLMs across the DoD introduces risks beyond the commercial sector:

- Operational Dependency: As LLMs are used in planning, acquisition, and ISR workflows, fabricated or flawed outputs can propagate through official channels unchecked.
- Amplified Error Risk: Unlike human analysts, LLMs scale mistakes invisibly and efficiently, embedding false data into staff products, policy drafts, and acquisition documents.
- Command Disruption: The DoD's chain-of-command culture depends on traceability and accountability. LLMs, as black-box systems, undermine both.

The potential for LLMs to introduce untraceable, confident, yet incorrect data into decision pipelines presents a new category of operational vulnerability.

The Core Failure Modes of LLMs

1. **Hallucinations (The Most Critical Risk)**
   What It Is: The model invents content, delivering it as fact.
   Why It Matters: Fabricated data in logistics plans, intelligence summaries, or acquisition documents can propagate unnoticed until operational failure occurs.
   Example: A request for Indo-Pacific basing agreements produced two real treaties and a fabricated third—a realistic-sounding but nonexistent agreement that made its way into draft operational planning.
   What Causes It: LLMs predict plausible word sequences, not factual knowledge. In absence of real data, they generate content based on pattern completion, not fact validation.

2. **Illusion of Accuracy**
   What It Is: The model sounds authoritative even when its content is incorrect.
   Why It Matters: Incorrect data delivered confidently can bypass critical thinking and manual checks.
   Example: A model confidently produced a list of defense treaties, two of which were entirely fictional.
   What Causes It: LLMs are optimized for fluency and coherence, not truth verification. Confidence in tone does not correlate with factual accuracy.

3. **False Execution Feedback**
   What It Is: The model falsely claims to have performed actions (e.g., edits or changes) it did not actually complete.
   Why It Matters: This can lead to uncorrected errors being passed forward, especially in time-sensitive or high-stakes contexts.

Example: A model reported having replaced "Phase I" language in a document, but upon review, no changes were actually made.
What Causes It: Most LLM systems lack internal feedback mechanisms or read-after-write verification. They predict the expected confirmation text rather than validating actions.

4. **No Chain of Accountability**
What It Is: When something goes wrong, there is no traceable reason or metadata to explain why.
Why It Matters: Military processes depend on traceability and version control. Without this, LLM-generated errors cannot be systematically corrected.
Example: A misattributed statement about ISR collection authorities in a report could not be traced to a source or justification.
What Causes It: LLMs are black boxes without inherent logging or decision tracking features.

5. **Misinterpreted Instructions**
What It Is: Clear prompts can produce incorrect outputs due to statistical mis-weighting.
Why It Matters: Misaligned outputs in planning documents or operational plans could have costly ramifications.
Example: A prompt seeking global AI defense vendors returned only U.S. primes despite explicit clarifications.
What Causes It: LLMs lean towards common responses statistically, deprioritizing less common instructions without warning.

6. **No Real Memory**
What It Is: The model forgets critical earlier instructions within the same session.
Why It Matters: Multi-step workflows, such as iterative proposal drafting, collapse when prior context is lost.
Example: During proposal revisions, a model regressed in tone and structure by forgetting prior corrections after only 3–5 prompts.
What Causes It: LLMs lack stable memory architectures. Context retention is probabilistic and typically short-lived.

7. **Constraint Underweighting**
What It Is: The model deprioritizes clear constraints in favor of producing what it deems statistically likely.
Why It Matters: Outputs can subtly violate rules or policy constraints, introducing errors that may not be noticed immediately.
Example: A request for post-WWI historical examples included content set in pre-modern history despite explicit instructions.
What Causes It: Constraint instructions are treated as optional soft guidance, not strict filters.

8. **Surface-Level Summarization**
What It Is: Important technical detail is lost as the model summarizes too aggressively.

Why It Matters: Detail dilution in acquisition, readiness reporting, or operational documents risks strategic misinterpretation.
Example: Mission capability distinctions were erased in a readiness report, replacing technical specifics with generic phrasing.
What Causes It: LLMs default to generalized summaries unless explicitly told to preserve detail, and even then, retention is inconsistent.

9. **Formatting Drift**
What It Is: The model drifts from expected report formats over time.
Why It Matters: Inconsistent formatting can reduce credibility and create significant rework.
Example: A request to maintain a standardized table format resulted in unpredictable formatting changes across iterations.
What Causes It: Format memory is not preserved unless reinforced continually. Models choose what they believe is stylistically appropriate in the moment.

Failure Mode	Operational Risk
Hallucinations	Fabricated data in plans and reports
Illusion of Accuracy	False confidence in incorrect information
False Execution Feedback	Missed document edits or task completions
No Chain of Accountability	Disruption of audit/legal processes
Misinterpreted Instructions	Inaccurate acquisition or planning inputs
No Real Memory	Collapse in multi-step document generation
Constraint Underweighting	Rule violations in sensitive documents
Surface-Level Summarization	Loss of critical operational details
Formatting Drift	Rework and credibility erosion

Limitations Are Not Defects

These failure modes are not bugs but rather architectural properties of LLMs. The models operate as pattern predictors, not fact-based systems. Expecting them to behave as analysts, editors, or validators invites systemic failure.

Recommendations for DoD Leaders

- **Human Validation Required**: All LLM outputs must undergo human review before operational use.
- **Treat as Drafting Aids Only:** LLMs should never be used for decision-ready outputs.
- **Avoid Final Use in Sensitive Documents:** Acquisition, intelligence, or operational materials must be LLM-reviewed, not LLM-generated.
- **Implement Oversight Layers:** Encourage development of metadata tracking and version control in LLM-assisted workflows.
- **Educate Human Operators:** Leaders and staff must be trained to recognize hallucination risks and other invisible failures.

Conclusion: Use, But Do Not Delegate Judgment

LLMs are powerful amplifiers of human productivity ***but they are not analysts***. They cannot validate facts. They cannot explain their decisions. And they will confidently produce incorrect information without warning. DoD leaders must adopt these tools carefully, enforcing oversight and human judgment to prevent technological overreach from becoming operational failure.

About the Author

Michael Morford is the CEO of the Knudsen Institute, a non-profit applied research organization advancing defense manufacturing and AI-driven industrial transformation. He is a former U.S. Army captain, theater-level logistics officer, disabled veteran of the Iraq War, and Douglas MacArthur Leadership Award recipient. He brings combined expertise in Wall Street investment banking and military war-planning to his role as a national advocate for expanding the surge capacity of the U.S. defense industrial base. Michael holds an engineering degree and MBA from Tulane University.