

# Responsible Integration of Large Language Models (LLMs) in Navy Operational Plan Generation

**Simon Kapiamba, Hesham Fouad, Ira S. Moskowitz**

Information Technology Division---5580

Naval Research Laboratory

Washington, DC 20375

simon.t.kapiamba.civ@us.navy.mil, hesham.y.fouad.civ@us.navy.mil, ira.s.moskowitz.civ@us.navy.mil

## Abstract

This paper outlines an approach for assessing and quantifying the risks associated with integrating Large Language Models (LLMs) in generating naval operational plans. It aims to explore the potential benefits and challenges of LLMs in this context and to suggest a methodology for a comprehensive risk assessment framework.

primarily focused on their technical capabilities and applications in civilian contexts. However, the unique challenges of military operations, particularly in naval environments, necessitate a specialized approach to risk assessment. This research seeks to bridge this gap by developing a methodology grounded in the specific requirements and complexities of naval operational planning.

## Background

Large Language Models (LLMs) are advanced artificial intelligence systems trained on vast datasets to process and generate text, enabling them to perform tasks ranging from simple question answering to complex content creation. LLMs, such as ChatGPT and Bard, have shown remarkable capabilities in understanding, interpreting, and generating human language. Their potential use in naval operations could offer significant strategic advantages, such as enhanced decision support, scenario analysis, and operational planning efficiency. However, the deployment of these advanced AI technologies in sensitive military contexts raises critical questions about risks, including potential biases, operational security concerns, and the reliability of AI-generated strategies.

The significance of this research lies in its focus on the integration of LLMs in naval operational plan generation. By examining the potential risks and developing a framework for the assessment of these risks, this research aims to contribute to the safe and strategic use of AI technologies in these naval operational contexts, while keeping technological advancements aligned with operational safety and effectiveness.

A review of the existing literature reveals a gap in comprehensive risk assessment frameworks specifically tailored for LLMs in military applications. Studies on LLMs have

## Methodology

The theoretical framework for this research is based on Bayesian networks, which offer a structured approach to modeling the complex relationships between various risk factors associated with LLM deployment. Bayesian networks are particularly suited for this purpose due to their ability to handle uncertainty and probabilistic relationships, making them ideal for assessing the multifaceted risks of LLM integration in naval operations.

Incorporating insights from the methodology outlined by Lauria and Duchessi (2007), this research adopts a similar structured approach to constructing Bayesian networks rooted in empirical data collection and analysis. Our methodology begins with the development and administration of a questionnaire whose responses will serve as the initial data source for identifying the most probable relationships between variables related to naval operational plan risks. This process sets up the subsequent parameterization of the conditional distributions for each node within our Bayesian network. Through this approach, we aim to build a robust Bayesian network model that accurately reflects the complex interdependencies and probabilistic nature of risks associated with LLM generated operational plans, further refining our model as we analyze more operational plans and integrate additional empirical data.

The methodology for this research involves several key steps:

- Questionnaire Development and Administration
- Data Collection and Analysis
- Bayesian Network Modeling
- Risk Categorization and Assessment

### Questionnaire Development and Administration

A detailed questionnaire will be designed to gather information from a wide range of stakeholders, such as naval personnel, AI experts, and military strategists. The questionnaire will contain a wide variety of naval operational plans, some LLM generated and some by operational planners, to identify potential avenues of risk in these plans.

### Data Collection and Analysis

Responses from the questionnaire will be collected and analyzed to identify concern areas on naval operational plans. This analysis will form the basis for identifying key risk factors associated with operational plans along with those specific to LLM generated plans.

### Bayesian Network Modeling

The identified risk factors will be modeled using Bayesian networks to understand their interdependencies and the probabilistic nature of these risks. This model will serve as a dynamic tool for assessing and visualizing the complex risk landscape of these generated operational plans.

### Risk Categorization and Assessment

Risks identified through the questionnaire and Bayesian network model will be categorized and assessed in terms of their potential impact on naval operations. This step involves a thorough evaluation of each risk factor, considering both its likelihood and severity.

## Analysis

The analysis phase of the research will involve a comparative study of hypothetical plans generated by LLMs and those created by human strategists. This comparison aims to highlight the strengths, limitations, and potential risks associated with LLM-generated plans. Additionally, the research will explore strategies for mitigating identified risks, such as incorporating checks and balances, enhancing transparency, and ensuring the continuous validation of LLM outputs.

The research is expected to yield several key outcomes:

- A comprehensive list of risk factors associated with the use of LLMs in naval operational planning, derived from stakeholder feedback and Bayesian network analysis.

- A structured framework for assessing the identified risks, providing a basis for decision-makers to evaluate and mitigate potential challenges in integrating LLMs into operational planning.
- Findings from the comparative analysis of LLM-generated and human-created plans, offering valuable insights into the capabilities and limitations of LLMs in operational planning contexts.
- Recommendations for mitigating the identified risks, ensuring that the integration of LLMs into actual operations is conducted in a manner that maximizes benefits while minimizing potential drawbacks.

## Preliminary Results

Here we have used ChatGPT to generate a sample fictional operational plan based on the following commander's intent: **"My intent is to support the Armed Forces of the Philippines in Humanitarian Assistance and Disaster Response (HA/DR) operations in the affected areas of the Central Philippines. We will provide all available assistance to alleviate human suffering and restore normalcy"** (Fig 1).

The LLM-generated operational plan for fictional HA/DR operations in the Central Philippines showcases the model's capability to structure a comprehensive response strategy. However, it also highlights several potential risks associated with LLM-generated plans.

The plan outlines setting up a forward-operating base on Panay Island without considering the island's current capacity to support such operations or the potential impact on local communities. This oversight could strain local resources or disrupt ongoing recovery efforts by local authorities.

The use of drones for immediate aerial surveys assumes that areas most in need can be quickly and accurately identified. However, this approach may not account for rapidly changing weather conditions or new emergencies arising post-assessment, potentially leading to misallocation of resources.

The reliance on naval helicopters for air drops in areas inaccessible by road could be problematic if there are unforeseen technical issues or if the helicopters are needed simultaneously for multiple urgent tasks. This highlights a potential overreliance on specific assets without considering alternative or backup methods for delivery.

The recovery phase suggests assisting in the restoration of critical infrastructure without detailing the scope of involvement or the criteria for beginning the withdrawal of naval assets. This lack of specificity could lead to premature withdrawal or extended engagements beyond the naval operation's capacity, affecting the overall effectiveness of the recovery efforts.

1. **Deployment Phase:**
  - Deploy naval assets, including two logistics ships for carrying relief supplies and a hospital ship for medical assistance, to the Central Philippines within 48 hours of operation activation.
  - Establish a forward-operating base on Panay Island to coordinate logistics and relief efforts effectively.
2. **Assessment Phase:**
  - Conduct an immediate aerial survey using drones to assess the extent of damage and identify critical areas requiring urgent assistance.
  - Establish communication with local authorities and the Armed Forces of the Philippines to gather intelligence and coordinate efforts.
3. **Relief Distribution Phase:**
  - Prioritize delivery of food, clean water, and medical supplies to the most affected areas identified in the assessment phase.
  - Utilize naval helicopters for air drops in areas inaccessible by road.
4. **Medical Assistance Phase:**
  - Set up a field hospital using the hospital ship's medical personnel and facilities to provide emergency medical care to the injured.
  - Coordinate with local health authorities to identify additional medical needs and direct resources accordingly.
5. **Recovery Phase:**
  - Assist in the restoration of critical infrastructure, including roads, bridges, and communication lines, to facilitate the recovery process.
  - Begin gradual withdrawal of naval assets as local authorities and international relief organizations take over long-term recovery efforts.
6. **Debriefing and Analysis Phase:**
  - Conduct a debriefing session with all stakeholders to evaluate the operation's effectiveness and identify lessons learned for future HA/DR missions.

Figure 1: ChatGPT generated operational plan

The examination of this LLM-generated operational plan not only illustrates the potential of AI in enhancing operational planning capabilities but also makes evident the necessity of our research effort in developing comprehensive risk models for LLM-generated naval operational plans. Identifying specific risk factors inherent in the AI's approach—such as assumptions about local infrastructure capacity, reliance on technology, and vagueness in the plan's execution and completion criteria—demonstrates the need for a framework that can evaluate these risks effectively so that they can be accounted for when these tools are used. Our research aims to address these gaps by using methodologies to assess, categorize, and manage the risks associated with deploying LLMs in complex operational environments. By integrating these risk models, we can better ensure that the innovative capabilities of LLMs are leveraged responsibly, enhancing the effectiveness and reliability of operational planning.

Palantir AIP (Artificial Intelligence Platform) is a cutting-edge approach to integrating artificial intelligence systems, including many capabilities that rely on LLMs, into operational environments. While leveraging these advanced tools allows organizations to harness the massive capabilities of AI systems, the use of LLMs in platforms such as Palantir AIP makes clear the critical need for comprehensive research into all of the potential associated risk factors. As

these AI-driven systems are deployed to assist with and automate more and more significant aspects of operational planning, the particular nuances of AI-generated content become important areas of concern.

Looking forward to the deployment of LLMs in naval operations exposes another avenue of risk, particularly in terms of the infrastructure required to support such technologies. The heavy computational and data transmission demands of LLMs, along with the need for real-time data processing and seamless communication for executing AI-driven operational plans, require a network infrastructure that is not only resilient but also highly adaptable. This is where the potential integration of Software Defined Networks (SDNs) becomes relevant. SDNs, known for their agility and configurability, represent an advanced networking framework that could potentially support the requirements of LLM deployments. However, the introduction of SDNs also brings forth questions regarding the readiness of existing naval network infrastructures to adapt to such advanced technologies, highlighting another area for consideration as we advance towards incorporating AI capabilities in naval operations.

The anticipated findings of this research have significant implications for the future of naval operations from the strategic integration of LLMs. By providing a comprehensive risk assessment framework, this research aims to facilitate informed decision-making and responsible use of AI technologies in naval operations to seed the ground work for its responsible use in wider military contexts. Moreover, the research addresses critical gaps in the literature, contributing to a deeper understanding of the challenges and opportunities presented by LLM integration in naval operations.

This research outlines a comprehensive approach to assessing the risks associated with integrating LLMs into naval operational planning. By developing a structured risk assessment framework and exploring the potential benefits and challenges of LLM deployment, this research aims to contribute to the responsible and effective use of AI technologies in military operations. Future research directions include the refinement of the risk assessment framework based on empirical findings, the exploration of additional applications of LLMs in wider military contexts, and the development of guidelines for the ethical use of AI in sensitive operational environments.

## Acknowledgements

We thank Bill Lawless for his assistance and guidance with this paper.

## References

- Du, Y., Li, S., Torralba, A., Tenenbaum, J.B. and Mordatch, I., 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325*.
- Esmailzadeh, Y., 2023. Potential Risks of ChatGPT: Implications for Counterterrorism and International Security. *International Journal of Multicultural and Multireligious Understanding (IJMMU) Vol, 10*.
- Ferrara, E., 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.
- Lauría, E.J. and Duchessi, P.J., 2007. A methodology for developing Bayesian networks: An application to information technology (IT) implementation. *European Journal of operational research*, 179(1), pp.234-252.
- Kenett, Y.N., Levi, E., Anaki, D. and Faust, M., 2017. The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(9), p.1470.
- Subramanian, D., Bhattachrajya, D., Torrado, R.R., Kephart, J., Chenthamarakshan, V. and Rios, J., 2017, December. A cognitive assistant for risk identification and modeling. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 1570-1579). IEEE.