SYM-AM-25-359

Excerpt from the Proceedings of the Twenty-Second Annual Acquisition Research Symposium and Innovation Summit

Volume III

Simplifying the Complex: A Conversational Approach to Configuring Military Simulators

Published: May 5, 2025

The research presented in this report was supported by the Acquisition Research Program at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

Simplifying the Complex: A Conversational Approach to Configuring Military Simulators

Protima Banerjee—has held technical and leadership positions designing and developing high performance data and computing systems in the DoD and commercial sectors for almost 30 years. Her contributions are part of critical programs provided to the U.S. Navy, U.S. Coast Guard and associated foreign military sales. Banerjee is a Technology Fellow at ASRC Federal and an Adjunct Professor of Computer Science at Rowan University. Her research interests include conversational interfaces, improving AI-driven search answer validation through semantic methods, optimizing large-scale data processing, advancing human-AI collaboration, and developing resilient AI models for mission-critical applications. [pbanerjee@asrcfederal.com]

Abstract

Complex military software systems, such as simulators and decision support tools, often necessitate extensive user training to master configuration tasks. This research proposes a novel approach to streamline user interactions with these systems by leveraging the capabilities of large language models in conjunction with semantic information structured in knowledge graphs. By employing a conversational interface, inexperienced users can interact with complex systems using natural language, significantly reducing the learning curve and operational overhead.

Introduction

Military software systems, such as simulators and decision support tools, are integral to modern defense operations. These systems are often highly complex, involving many intricate configuration parameters that require significant training to setup correctly. Operators must possess a combination of deep domain knowledge and technical expertise; this type of skilled individual is difficult to attract, train and retain. Consequently, the learning curve associated with these systems can create operational bottlenecks, ultimately hindering mission agility and effectiveness.

This research introduces a novel approach to address these challenges by leveraging the capabilities of large language models (LLMs) in conjunction with semantic information represented in knowledge graphs (KGs). LLMs, known for their proficiency in natural language understanding, enable users to interact with complex systems using conversational language, drastically simplifying the configuration and operation of these systems. By pairing LLMs with the semantic and syntactic data representation capabilities of KGs, this approach offers a more intuitive and accessible interface. This conversational interface can enable domain experts to interact with simulation systems in a manner that feels natural and intuitive, substantially reducing the time and effort required for training while minimizing the potential for user error.

The remainder of this paper walks through our problem statement, background and related work in LLMs and KGs, describes our technical approach, experiment setup, and concludes with a discussion of our results and future work.

Problem Statement

Traditional user interfaces for military simulation systems are often based on menus, forms, and commands and are not always intuitive. These interfaces are designed with expert users in mind, but they may alienate or frustrate novice users and domain experts who are unfamiliar with the intricacies of the underlying system. As a result, simulation system operators often require significant training and a long onboarding process. This not only affects system performance but can also contribute to increased operational overhead.

In response to this challenge, there is an urgent need for more efficient and accessible interaction interface. In recent years, LLMs have emerged as a promising approach to user interaction for military systems (Mikhailov, 2023). This technology enables operators to communicate with complex software using everyday language and has been shown to be effective for tasks such as course of action generation. However, even with the advancements in modern LLMs, several key issues must be addressed before such conversational interfaces can be reliably used to interact with real-world military systems:

1. **Natural Language to Structured Data Translation:** One of the most critical challenges is translating human-readable natural language inputs into structured, machine-readable data. This involves accurately interpreting user queries, which can vary greatly in terms of phrasing, domain-specific terminology, and syntax, and converting them into a format that the underlying system can process—typically a structured data format like JSON. For complex military systems, the translation must not only be accurate but also capable of handling both general and domain-specific contexts, such as military terminology, operational constraints, and system requirements. Achieving this goal would make natural language interaction a viable alternative to traditional interfaces, reducing the cognitive load on users and eliminating the need for detailed system expertise.

2. **Data Integrity and Completeness:** Once a user query is translated into structured data, ensuring that the data is complete, consistent, and free from errors is critical. Inaccuracies or omissions in system configurations can have serious consequences, particularly in mission-critical environments. Prior to deployment, an interface must be able to identify and address common issues in user input, such as incomplete or contradictory instructions, and ensure that the generated data meets the necessary standards for use in military systems. This includes validating the integrity of the data against the system's operational rules and constraints, as well as providing feedback to the user when input conflicts or inconsistencies are detected.

This research explores the above topics and proposes a novel architecture that leverages LLMs in conjunction with semantic information in the form of a structured KGs.

Background and Related Work

LLM Use Cases

LLMs have emerged as a powerful tool for processing and analyzing vast amounts of information, improving decision-making, and enhancing human-machine interaction (Caballero, 2024). In this section we briefly describe a selection of the military-relevant use cases in which LLMs have shown promise.

Intelligence Analysis

One of the earliest domains to explore the use of LLMs in the military has been intelligence gathering and analysis as shown by Logan (2024) and Nitzl (2024). The volume of data generated from various sources—including satellite imagery, signals intelligence (SIGINT), open-source intelligence (OSINT), and classified reports—has outpaced traditional analytical methods. LLMs have been deployed to automate and enhance intelligence analysis by extracting key insights, summarizing reports, and identifying patterns or anomalies that may indicate threats.

Military Decision-Aids

LLMs have also been explored to aid military-decision making. This process is inherently complex, requiring commanders to synthesize information from multiple domains—land, air, sea, space, and cyber. LLMs improve the efficiency of current systems by helping military planners analyze battlefield conditions, generate courses of action (COAs), and evaluate mission risks (Goecks, 2024). These AI-driven systems can serve as an assistant to commanders by generating real-time reports, summarizing intelligence briefings, and suggesting potential responses based on historical data and current operational factors.

Additionally, LLMs usage for battlefield management systems is another area of active research (Connolly, 2024). These systems process sensor data, intelligence reports, and battlefield communications, allowing commanders to access critical information through natural language queries. By combining LLMs with knowledge graphs and structured data sources, military operators can retrieve highly relevant and contextual information without the need for extensive manual searching.

Cybersecurity Operations

The modern battlefield extends into cyberspace, where cyber warfare and digital threats pose significant challenges. LLMs have been increasingly employed in cybersecurity operations for automated offense and defense mechanisms (Anurin, 2024). In this domain, LLMs can be used to analyze vast amounts of cyber threat intelligence, detecting patterns of malicious activity and predicting potential vulnerabilities. Additionally, LLMs have been integrated into cybersecurity chatbots and virtual assistants to help analysts rapidly assess and respond to cyber incidents (Shafee, 2024).

Human Computer Teaming

Another promising domain area for LLM use in the military is human-computer teaming. The military has increasingly relied on autonomous systems, including unmanned aerial vehicles (UAVs), robotic ground units, and AI-driven mission control assistants. Effective communication between human operators and these autonomous systems is essential for mission success. LLMs have been explored as a means to enhance human-machine interaction by providing more intuitive and natural language interfaces (Javaid, 2024).

Logistics and Supply Chain Management

Efficient logistics and supply chain management are crucial for sustaining military operations. LLMs could be utilized to optimize logistics planning, streamline supply chain coordination, and predict equipment maintenance needs as shown in Aghaei (2025) and Olena (2024). By analyzing historical data and real-time logistical information, these models help military logisticians identify potential bottlenecks, improve inventory management, and ensure timely delivery of critical supplies.

One application of LLMs in logistics involves predictive maintenance (Lukens, 2023). AI-driven models analyze sensor data and maintenance records to forecast potential mechanical failures, allowing for proactive maintenance scheduling. This capability reduces downtime and enhances the overall readiness of military assets.

Regulatory Compliance.

Finally, LLM use has been explored for automation of compliance processes (Makovec, 2024). By reasoning through and automating some or all of the compliance process, LLMs have the potential to help reduce administrative workloads and improve overall efficiency in compliance operations.

Limitations of LLMs

Despite significant progress operationalizing LLM use, existing approaches face several limitations (Biswas, 2023):

- Contextual Understanding: LLMs struggle with domain-specific language understanding and operational contexts.
- Data Validation: Most systems lack robust mechanisms to identify and resolve inconsistencies or omissions in the generated data.

KGs

A KG is a structured representation of information that captures relationships between entities in a way that mimics human understanding (Hogan, 2021). Unlike traditional databases that store information in isolated tables, KGs use a network of interconnected nodes and edges to represent data as a web of relationships. This enables more intuitive data retrieval, contextual reasoning, and advanced analytics. KGs power applications like search engines, recommendation systems, and AI-driven assistants by enabling machines to understand and infer meaning from complex data (Peng, 2023). They are built using ontologies, making them particularly valuable for domains like the military and have the potential to play a crucial role in powering intelligent applications.

Technical Approach

Overview

The integration of LLMs and KGs offers a promising pathway to overcome the above limitations. By combining the natural language capabilities of LLMs with the semantic structuring power of KGs, it is possible to create natural language system interfaces that are both intuitive and reliable. These interfaces can facilitate the translation of user inputs into structured data, while ensuring data integrity through validation mechanisms. An overview of our technical approach and architecture is presented in Figure 1.

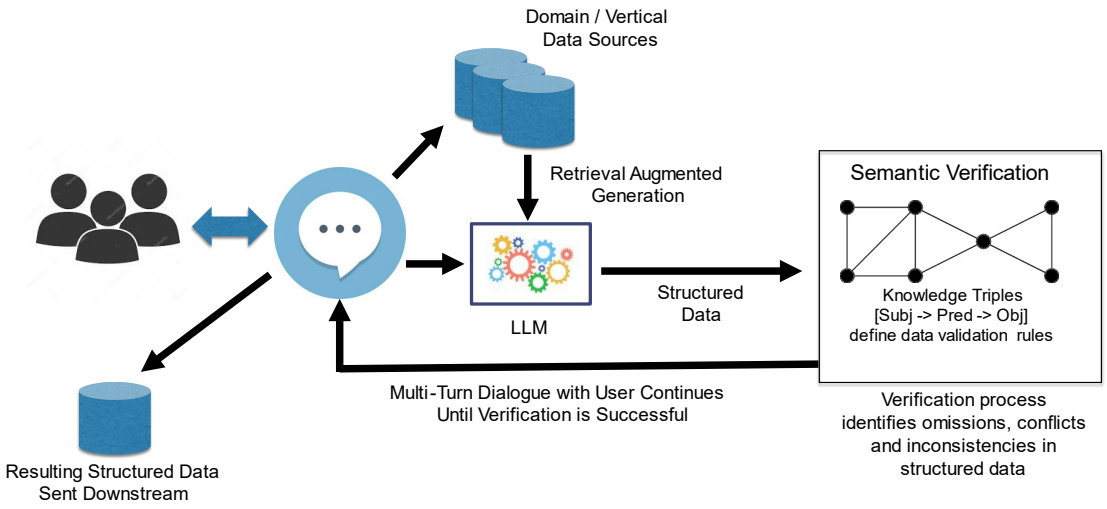

Figure 1. Technical Approach and System Architecture

The system design consists of several key layers working together to ensure accuracy, consistency, and usability of the final output. The process begins with a user interaction component presented as a chat interface. This interface manages communication and forwards

user queries to the LLM for processing. The processing layer leverages retrieval-augmented generation (RAG; Gao, 2023), allowing the LLM to access relevant external knowledge sources, such as structured databases or KGs, to generate a well-formed structured data output.

Once generated, the structured data passes through to a semantic verification component, which checks for data inconsistencies, omissions, and conflicts. This layer applies a combination of rule-based logic and reasoning techniques to ensure the generated data aligns with known facts and domain constraints. If issues are detected, the system triggers a feedback loop, presenting the user with clear questions to direct the necessary adjustments. The user can then refine their input iteratively until all inconsistencies are resolved, ensuring high data integrity every for every single instance of generated data.

After validation, the finalized structured data is sent to the the military simulator to configure the simulation scenario. It is noted here that while we are focused on a miltary simulation system as the downstream data processing engine in this case, this architecture could be applicable to any system that requires complex structured data inputs.

Semantic Verification

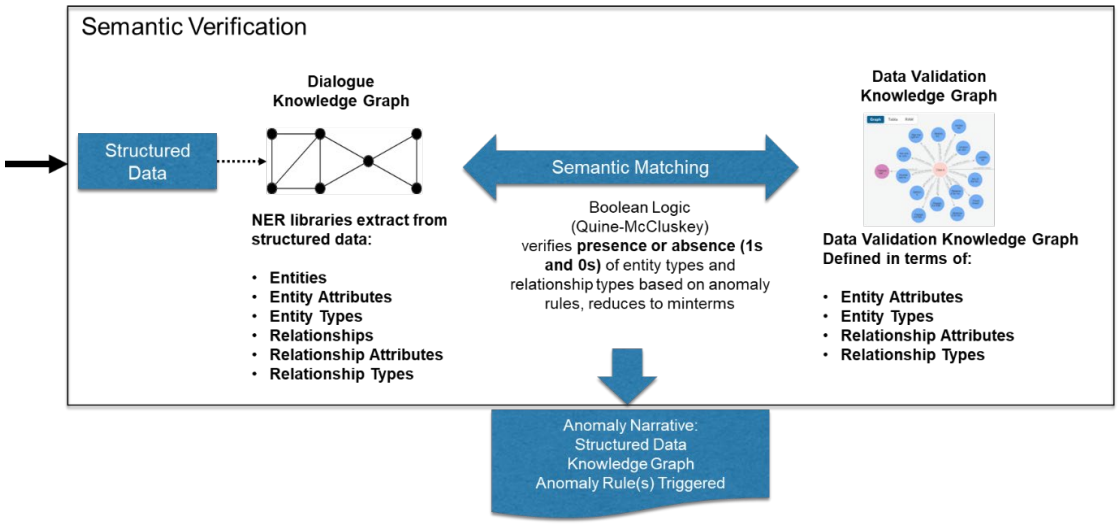

Figure 2. Design of the Semantic Verification Component

The semantic verification component shown in Figure 2 plays a critical role in ensuring the accuracy and consistency of structured data by leveraging knowledge graphs and Boolean logic algorithms. This component operates in multiple stages to validate structured data effectively.

First, the component constructs a KG from the structured data input. It utilizes Named Entity Recognition (NER) libraries to extract key elements, including entities, entity attributes, relationships, relationship attributes, and relationship types. This transformation ensures that the structured data is represented in a graph-based format suitable for comparison.

Next, the generated KG is compared against a predefined validation KG. This reference graph consists of known validation triples, defined in terms of entity attributes, entity types, relationship attributes, and relationship types. The goal of this step is to detect inconsistencies by examining the alignment between the structured data and the established validation rules.

To facilitate this matching process, the system employs Boolean logic algorithms, such as Quine-McCluskey, to systematically verify the presence or absence of required entity and

relationship types. Each entity type and relationship type is mapped to a binary representation (1s and 0s), indicating whether they exist within the structured data. Using minimization techniques, the algorithm reduces these binary values to a minimal set of essential conditions (minterms) that highlight the key discrepancies.

Finally, the component generates an “anomaly narrative” for the user, detailing the specific validation rule(s) that were triggered due to inconsistencies. This narrative provides actionable insights, guiding the user to refine their input iteratively until the structured data fully aligns with the expected KG. By enforcing structured validation through formal logic, this approach ensures data integrity before passing the refined output to downstream systems.

One of the key advantages of this method is its ability to provide users with clear, actionable explanations of discrepancies. Instead of vague error messages, users receive precise feedback on which validation rules were triggered, allowing them to refine their input iteratively. This structured feedback loop ensures that only high-quality, validated data proceeds to downstream systems, reducing errors and improving decision-making.

The approach is also scalable, as KGs enable a structured representation of complex relationships, and the use of Boolean logic matching optimizes computational processing. By outputting only the minterms of the Boolean logic evaluation, the system reduces redundant validation checks, ensuring that the foundational inconsistencies are flagged. This efficiency is particularly beneficial in large-scale applications where structured data verification must be performed rapidly.

Technology Stack

Our implementation technology stack is shown in Figure 3.

	Category	Description
Knowledge Graph Toolset	Evaluation	Amazon Relational Data Store (RDS) for storing experiment results, Langchain Evaluation Framework
	Knowledge Graph Correlation	Cypher stored procedures running in Neo4j implement graph -matching algorithms, (Developed code)
	Graph Database	Neo4j (Containerized graph database, running on AWS compute)
	Knowledge Graph Generation	Langchain LLMGraphTransformer (Generates knowledge graphs from text, augmented with custom developed code for detection of complex relationships between entities)
LLM Toolset	LLM framework	Langchain (Langchain agents implement narrative management - addition and summarization)
	LLM (Hosted)	**gpt-3.5-turbo** accessed via OpenAI API, running remotely
	LLM (On-premise)	**Llama, Gemma, etc.** accessed via the Ollama framework, **running locally on AWS compute**
Infrastructure	Python	Python 3.10
	Hardware	Amazon Web Services (AWS) GovCloud (1 NVIDIA GPUs and vCPUs)

Figure 3. Implementation Technology Stack

Evaluation

Representative Dataset: Automated Identification System

Automated Identification System (AIS) data was selected as an ideal test case for our system prototype due to its structured yet moderately complex nature. This openly available

maritime dataset (Kress, 2023) contains standardized information transmitted by commercial vessels and Coast Guard ships for collision avoidance.

AIS data is well-documented and includes essential mandatory fields such as ship identification number (MMSI), latitude, longitude, speed, and course. Additionally, several AIS fields require accuracy checks—for example, position and course must be correctly expressed in degrees, speed in knots, and timestamps in UTC. The need for consistency checks, such as ensuring that a ship's course and heading align and that its status corresponds logically with speed and heading, makes this dataset particularly relevant for detecting anomalies or inconsistencies. AIS domain knowledge is readily available from open sources such as NOAA AccessAIS (NOAA, 2024) and can be used to automatically or semi-automatically populate a domain knowledge rules as seen in Figure 4. The specific data set we used in these experiments consisted of 1000 AIS messages from vessel traffic around the port of New Orleans on March 31, 2022.

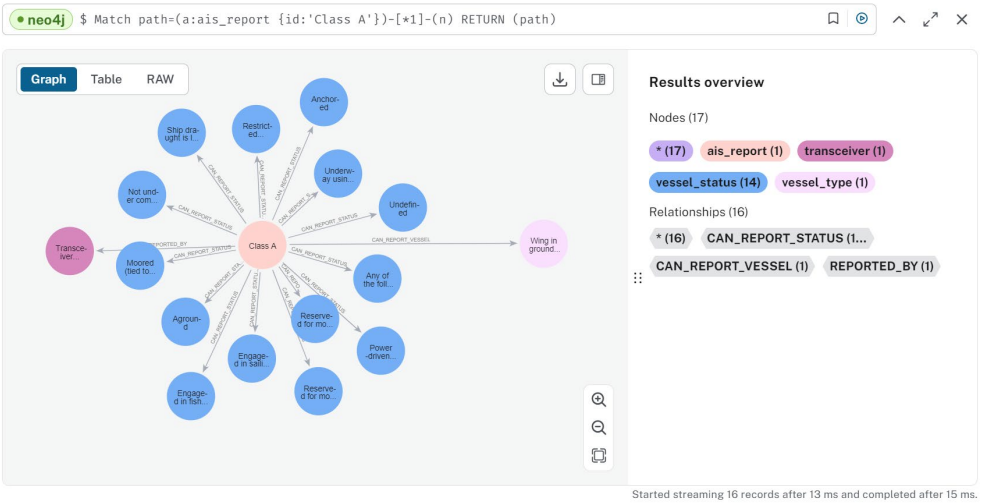

Figure 4. AIS Domain Rules Captured as a KG in Neo4J

Experiment Design

To evaluate the reliability of our technical approach, we designed a multi-step experimental framework, as shown in Figure 5.

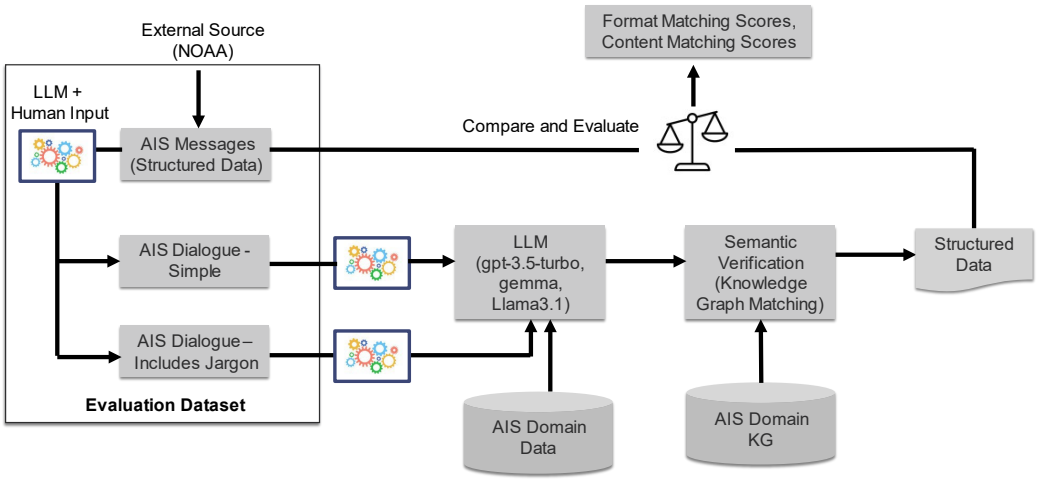

Figure 1. Experiment Framework and Setup

We began by establishing a source of truth dataset derived from structured AIS messages. We transcribed this structured data into natural language dialogue using a combination of LLM processing and human input. We created two distinct dialogue styles:

- Simple Dialogue—A straightforward transformation of AIS data into natural language, maintaining clarity and minimal linguistic variation.
- Colloquial and Jargon-Based Dialogue—A more complex transformation incorporating maritime jargon, conversational elements, and informal phrasing to mimic real-world human communication.

Next, we took the generated dialogues and processed them through various LLMs to reconstruct structured AIS data. We also incorporated our semantic verification process, ensuring that the extracted information adhered to expected AIS data structure formats and consistency rules. By varying the LLMs used in this process, we examined differences in their ability to infer structured data from both simple and jargon-heavy dialogue inputs.

Finally, we compared the reconstructed structured data to the original source of truth AIS dataset. The evaluation process generated format and content scores, assessing how accurately the LLM-driven reconstruction aligned with the original structured information. Format scores measured adherence to expected data structures (e.g., proper formatting in JSON, adherence to schema), while content scores quantified semantic accuracy, ensuring that key details such as course, heading, and ship status were correctly transcribed.

Our goal in this experiment was to determine how reliably our technical approach using LLMs in conjunction with a KG could convert human-generated dialogue into structured data. By analyzing performance across different dialogue styles and LLM models, we aimed to identify potential challenges and opportunities in using AI to extract structured data from human communication in the maritime data domains.

Experiment Results and Discussion

The results of our experiments are presented in Figure 6.

LLM	Dialogue Type	Semantic Verification	Format Score (%)	Content Score (%) Vessel Type Identification	Content Score (%) Status Identification
Gpt-3.5-turbo	Simple	No	100%	100%	94%
Gpt-3.5-turbo	Simple	Yes	100%	100%	99%
Gpt-3.5-turbo	**Jargon**	**No**	100%	24%	65%
Gpt-3.5-turbo	**Jargon**	**Yes**	100%	73%	79%
Llama3-8b	Simple	No	94%	100%	99%
Llama3-8b	Simple	Yes	95%	100%	99%
Llama3-8b	Jargon	No	94%	34.5%	40%
Llama3-8b	Jargon	Yes	95%	48%	38%
Gemma-7b	Simple	No	100%	100%	99%
Gemma-7b	Simple	Yes	100%	100%	99%
Gemma-7b	**Jargon**	**No**	100%	3%	16%
Gemma-7b	**Jargon**	**Yes**	100%	49%	39%

Figure 6. Experiment Results

Our experiments demonstrated that 100% format similarity, including strict adherence to JSON schema conformance, is easily achievable using state-of-the-art LLMs and LLM frameworks. Regardless of the complexity of the input dialogue, LLMs consistently produced structured outputs that matched the expected format.

However, content similarity presented more significant challenges. When the input dialogue was simple and direct, LLMs successfully extracted and translated the information into the correct structured data fields. In contrast, when jargon and colloquial language were introduced, accuracy dropped substantially. This was particularly evident in two key AIS fields: vessel type and vessel status. Both fields are encoded as numerical values in structured AIS data but would be described in natural language when reported by humans. For example, a vessel type code "31" could correspond to the words "Tug," "Tugboat," "Towing vessel," "ship assist vessel," etc. in dialogue, and an LLM must correctly map such descriptions back to their respective code values. Another example is status code "0," which could correspond to the words "underway," "at sea," "cruising," "sailing," etc. Accuracy in this mapping varied depending on the LLM, but we found that incorporating a KG significantly improved accuracy for two of the models tested (GPT-3.5-Turbo and Gemma:7B). The knowledge graph provided a structured reference, reducing ambiguity and improving the alignment between natural language descriptions and standardized AIS codes.

Several additional systematic errors were observed in the reconstructed structured data. LLMs exhibited a 5% error rate in latitude and longitude rounding, which could introduce small but meaningful inaccuracies in precise geospatial applications. They also struggled with date conversions, with a 10%–15% error rate in formatting timestamps correctly into ISO 8601. Even simple unit conversions, such as feet to meters, resulted in a 5% error rate, highlighting a consistent challenge in numerical data transformations.

To achieve 100% content accuracy, our findings indicate that additional validation and consistency checks must be incorporated into the processing framework. These include enforcing strict unit conversion rules, leveraging semantic matching techniques for natural language descriptions, and integrating knowledge graphs to improve structured data reconstruction when jargon and ambiguous terminology are present.

Conclusions and Future Work

This research addresses the challenges of interacting with complex military software systems by proposing a novel approach to conversational user interfaces. Our experiments demonstrate that while format similarity in structured data reconstruction is easily achievable with state-of-the-art LLMs and LLM frameworks, content accuracy remains a challenge, especially when processing natural language with jargon and domain-specific terminology. We found that KGs significantly improve accuracy in mapping ambiguous natural language descriptions to structured code values, particularly for fields like vessel type and vessel status. However, issues such as geospatial rounding errors, incorrect date formatting, and inconsistent unit conversions highlight the need for additional validation mechanisms to ensure high-fidelity structured data extraction.

As a next step, we aim to refine our approach by integrating verification techniques for numerical transformations, expanding the knowledge graph to cover more maritime-specific terminology, and fine-tuning LLMs with domain-specific training data. Further research will explore hybrid AI architectures that combine LLMs, KGs, and (potentially) deterministic validation mechanisms to achieve near-perfect content accuracy. These improvements will enhance the reliability of AI-driven structured data extraction, making it more applicable to real-world military simulators and other structured data domains.

References

Aghaei, R. K. (2025). *The potential of large language models in supply chain management: advancing decision-making, efficiency, and innovation*. ArXiv preprint arXiv:2501.15411.

Anurin, A. N. (2024). *Catastrophic cyber capabilities benchmark (3CB): Robustly evaluating LLM agent cyber offense capabilities*. ArXiv preprint arXiv:2410.09114.

Biswas, S. (2023). Prospective role of chat GPT in the military: According to chatGPT. *Quios*.

Caballero, W. N. (2024). *On large language models in national security applications*. ArXiv preprint arXiv:2407.03453.

Connolly, B. J. (2024). Battlefield information and tactics engine (BITE): A multimodal large language model approach for battlespace management. *SPIE Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, *13051*.

Gao, Y. Y. (2023). *Retrieval-augmented generation for large language models: A survey*. ArXiv preprint arXiv:2312.10997.

Goecks, V. G. (2024). COA-GPT: Generative pre-trained transformers for accelerated course of action development in military operations. *International Conference on Military Communication and Information Systems (ICMCIS)*.

Hogan, A. E. (2021). Knowledge graphs. *ACM Computing Surveys (Csur) 54*(4), 1–37.

Javaid, S. F. (2024). Large language models for UAVs: Current state and pathways to the future. *IEEE Open Journal of Vehicular Technology*.

Kress, M. M. (2023). *AIS data: An overview of free sources*.

Logan, S. (2024). Tell me what you don't know: Large language models and the pathologies of intelligence analysis. *Australian Journal of International Affairs*, 220–228.

Lukens, S. (2023). Evaluating the performance of chatgpt in the automation of maintenance recommendations for prognostics and health management. *Annual Conference of the PHM Society*, *15*(1), 1–18.

Makovec, B. R. (2024). *Preparing AI for compliance: Initial steps of a framework for teaching LLMs to reason about compliance*.

Mikhailov, D. I. (2023). *Optimizing national security strategies through LLM-driven artificial intelligence integration*. ArXiv preprint arXiv:2305.13927.

Nitzl, C. C. (2024). *The use of artificial intelligence in military intelligence: An experimental investigation of added value in the analysis process*. ArXiv. preprint arXiv:2412.03610.

NOAA. (2024). *AccessAIS*. https://coast.noaa.gov/digitalcoast/tools/ais.html

Olena, K. (2024). *Application of LLMs for a chatbot system in the logistics industry.*

Peng, C. X. (2023). Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review, 56*(11), 13071–13102.

Shafee, S. A. (2024). Evaluation of LLM-based chatbots for OSINT-based cyber threat awareness. *Expert Systems with Applications*.

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net