Large Language Models in Wargame Scenario Generation: A Survey on Knowledge Extraction and Graph-Based Scenario Analysis

***Abstract* –- This literature survey examines the role of Large Language Models (LLMs) in wargame scenario generation, with a focus on knowledge extraction and knowledge graph construction for military simulations. By reviewing recent studies from 2020–2025, sourced from Google Scholar, arXiv, and IEEE Xplore, it explores how these models can sift through messy, unstructured military data—like reports and manuals—to pull out key details such as units, locations, and events. Techniques like Named Entity Recognition (NER) and lemmatization help identify and organize this information, which is then used to build knowledge graphs. These graphs act like visual maps, connecting entities and their relationships to create dynamic, realistic wargame simulations. The paper also tackles challenges, such as ensuring data security in sensitive defence settings and dealing with AI "hallucinations" (when the model generates incorrect information). Findings highlight the potential of LLMs to automate scenario generation, enhancing efficiency in military planning. Ultimately, this survey shows how LLMs are becoming powerful tools for crafting complex, believable scenarios to support military strategy and decision-making.**

All papers are listed in a repository, and is available at [References Repository]

Keywords* — *Large Language Models, Artificial intelligence in wargaming, Wargame scenario generation, Strategic simulation, Knowledge extraction, Name Entity recognition, Graph visualization, Event Extraction, Military Simulation, Conceptual Maps, Scenario Analysis, Decision Support, Prompt Engineering

I. INTRODUCTION

Large language models (LLMs) can extract knowledge and create graphs for wargame scenario generation. LLMs can access factual knowledge through knowledge graphs, which provide external information for reasoning. By learning efficient retrieval strategies, LLMs can identify and gather specific information needed for complex, multi-step reasoning tasks. Retrieval-augmented generation (RAG) frameworks integrate external knowledge into the inference process, enhancing LLM performance in specialized domains. RAG enhances LLMs by incorporating information retrieved from external knowledge sources, improving accuracy and credibility. However, these sources may contain irrelevant or incorrect data that can reduce RAG's effectiveness [1].

A. Definition and importance of wargame scenario generation

Wargame scenario generation is the process of creating structured, realistic, and actionable scenarios to simulate military operations, conflicts, or strategic interactions for training, planning, and decision-making purposes. These scenarios encompass detailed representations of entities (e.g., military units, adversaries, terrain), relationships (e.g., command structures, alliances), and dynamic events (e.g., troop movements, engagements). Wargames are critical in military applications for testing strategies, training personnel, and preparing for real-world contingencies. They range from quantitative games with predefined moves to qualitative games with open-ended responses, often requiring extensive domain knowledge and contextual accuracy [2].

In defence contexts, wargames serve multiple purposes: operational planning, capability assessment, and policy evaluation. For instance, political-military wargames simulate geopolitical crises to anticipate escalation pathways, while tabletop exercises test unit readiness. The complexity of modern warfare, including hybrid threats and multi-domain operations, necessitates scenarios that are both comprehensive and adaptable, posing significant challenges for manual generation [3].

B. Role of Large Language Models in Defence Applications

Large Language Models (LLMs), such as GPT-4, LLaMA, and BERT, have transformed natural language processing (NLP) by enabling advanced text understanding and generation capabilities. In defence, LLMs are increasingly integrated into applications like intelligence analysis, automated planning, and wargame simulation. Their ability to process unstructured textual data, extract relevant knowledge, and generate coherent outputs makes them suitable for automating scenario generation. For example, LLMs can analyse military manuals, after-action reports, and open-source intelligence to create realistic scenarios, reducing the time and expertise required compared to traditional methods [4].

Recent advancements, such as the U.S. Army's use of GPT-4 for terrain analysis in wargaming, demonstrate LLMs' potential to enhance situational awareness and decision-making. However, their integration requires addressing domain-specific challenges, such as handling military terminology and ensuring data security [5].

C. Problem and Challenges in Manual Scenario Generation

Manual wargame scenario generation is labour-intensive, requiring domain experts to synthesize vast amounts of unstructured data (e.g., historical records, intelligence reports) into coherent scenarios. This process is prone to biases, inconsistencies, and scalability limitations. Key challenges include:

1. *Data Overload:* The volume of unstructured military documents overwhelms human analysts.
2. *Domain Complexity:* Military terminology, operational graphics, and temporal dynamics require specialized knowledge.
3. *Consistency and Reusability:* Manually crafted scenarios often lack standardized structures, hindering reuse across simulations.
4. *Time Constraints:* Rapidly evolving threats demand quick scenario updates, which manual methods struggle to deliver [6].

D. Need for Automated Knowledge Extraction and Structured Representation

Automated knowledge extraction using LLMs can address these challenges by processing unstructured data to identify entities, relationships, and events relevant to wargaming. Knowledge graphs (KGs) provide a structured representation of this extracted information, enabling dynamic scenario generation through relational reasoning and temporal modelling. The combination of LLMs and KGs facilitates real-time adaptation, scenario validation, and integration with existing defence systems, aligning with the needs of modern military operations [7].

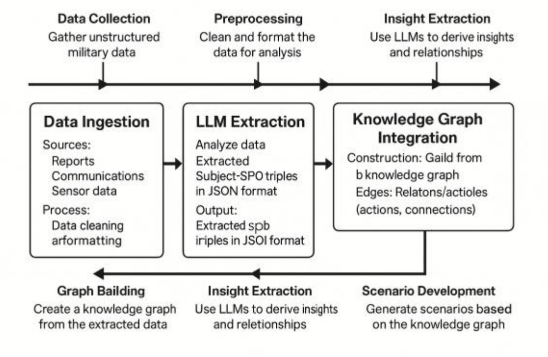

Figure 1 Pipeline for transforming unstructured military data into actionable scenarios using knowledge graph-based analysis. The process includes Data Ingestion from reports, communications, and sensors; LLM Extraction of Subject-Predicate-Object (SPO) triples; Knowledge Graph Integration to map entities and relationships; and Scenario Generation for military decision-making, with arrows indicating the information flow between stages.

II. LARGE LANGUAGE MODELS(LLMs)

Large Language Models (LLMs) are advanced artificial intelligence systems trained on vast amounts of text data to understand, generate, and manipulate human language. They utilize deep learning architectures—primarily transformers—to process and produce coherent, contextually relevant text. LLMs have revolutionized natural language processing (NLP), powering applications ranging from conversational agents and content generation to advanced reasoning and decision support systems.

Key Characteristics

- *Scale:* LLMs are trained on billions to trillions of parameters, allowing them to capture nuanced patterns in language.
- *Generalizability:* They can perform a wide range of language tasks without task-specific training.
- *Contextual Understanding:* LLMs maintain context over long passages, enabling coherent multi-turn conversations and complex text analysis.

A. Evolution of LLMs

1. GPT Series (OpenAI)

The Generative Pre-Trained Transformer (GPT) series by OpenAI marked a significant leap in LLM development. Starting with GPT in 2018, each successive version—GPT-2, GPT-3, and GPT-4—has increased in size, capability, and sophistication.

- *GPT-2 (2019):* Demonstrated strong zero-shot and few-shot learning abilities, generating coherent paragraphs of text [1].
- *GPT-3 (2020):* With 175 billion parameters, it could perform translation, question-answering, and code generation with minimal prompts [2].
- *GPT-4 (2023):* Further improved reasoning, factual accuracy, and multimodal capabilities (text and image inputs) [3].

2. LLaMA (Meta)

Meta's LLaMA (Large Language Model Meta AI) models, released in 2023, focus on efficiency and accessibility. LLaMA models are designed to be more resource-efficient, enabling broader research and deployment in academic and industrial settings [4].

- *LLaMA 2:* Open-weight models that rival GPT-3 in performance, with strong results in multilingual and domain-specific tasks [4].

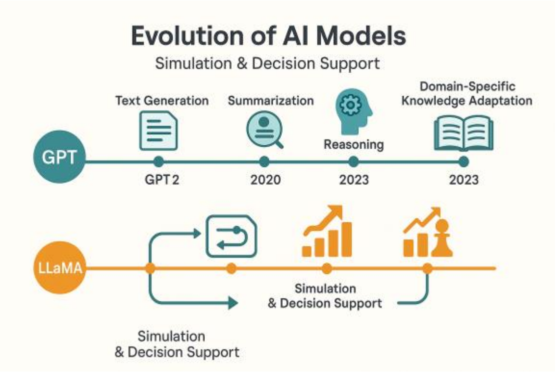
Evolution of AI Models
Simulation & Decision Support
Text Generation
Summarization
Reasoning
Domain-Specific Knowledge Adaptation
GPT
GPT 2
2020
2023
2023
LLaMA
Simulation & Decision Support
Simulation & Decision Support

Figure 2 Timeline illustrating the evolution of AI models GPT and LLaMA, emphasizing their capabilities in simulation and decision support.

B. Capabilities of LLMs

1. Text Generation

LLMs excel at generating coherent, contextually appropriate text for a variety of purposes, including storytelling, report writing, and code generation.

2. Summarization

They can condense long documents into concise summaries, preserving essential information and context.

3. Reasoning

Modern LLMs demonstrate advanced reasoning abilities, such as solving math problems, logical puzzles, and answering complex questions by synthesizing information from multiple sources.

4. Domain-Specific Knowledge Adaptation

Through fine-tuning and prompt engineering, LLMs can be adapted to specialized domains—such as law, medicine, or military operations—delivering expert-level responses.

C. LLMs in Simulation & Decision Support

Large Language Models (LLMs) are increasingly integrated into military simulation and decision support systems, bringing transformative capabilities to wargaming, operational planning, and real-time battlefield management.

1. Simulation Applications

- *Scenario Generation & Enhancement:* LLMs can automatically generate complex, realistic wargame scenarios, reducing the reliance on subject matter experts and accelerating scenario development. They can create dynamic narratives, inject plausible adversary actions, and adapt scenarios in real time to reflect evolving operational conditions.
- *Scalability & Immersion:* By leveraging LLMs, simulations can scale to include more diverse entities and events, increasing realism and immersion. For example, LLMs can generate communications, intelligence reports, and inject unexpected events, making training exercises more robust and challenging.

2. Decision Support

- *Course of Action (COA) Analysis:* LLMs assist commanders and staff by rapidly generating, assessing, and recommending courses of action. They can synthesize large volumes of operational data, historical precedents, and doctrinal knowledge to support informed decision-making at the speed required by modern operations [5].
- *Supporting Role:* Current consensus is that LLMs are best suited for supporting human decision-makers rather than replacing them, especially at the strategic level [5].

III. KNOWLEDGE EXTRACTION FROM LLMs

Knowledge extraction from Large Language Models (LLMs) involves deriving structured information, such as entities, events, and relationships, from unstructured textual outputs to support applications like wargame scenario analysis. This survey reviews the definition, techniques, applications, and existing research related to knowledge extraction from LLMs for wargame scenarios, focusing on its relevance to defence - related contexts. The survey covers Named Entity Recognition (NER), Relation Extraction, Event Extraction, Prompt Engineering, Few-shot/Zero-shot Learning, and their applications, along with relevant tools.

The process of transforming unstructured, narrative-style text generated by LLMs into structured formats, such as tables, graphs, or ontologies, suitable for computational analysis. The structured knowledge includes:

- *Entities:* Specific objects or concepts, such as military units (e.g., "T-90 tank"), locations (e.g., "Ladakh region"), or organizations (e.g., "Infantry Brigade").

- *Events:* Specific occurrences with temporal and causal attributes, such as "missile launch on June 15, 2025" or "troop deployment in a border region."

- *Relationships:* Connections between entities or events, such as "T-90 tank deployed in a strategic region" or "missile launch caused by enemy provocation."

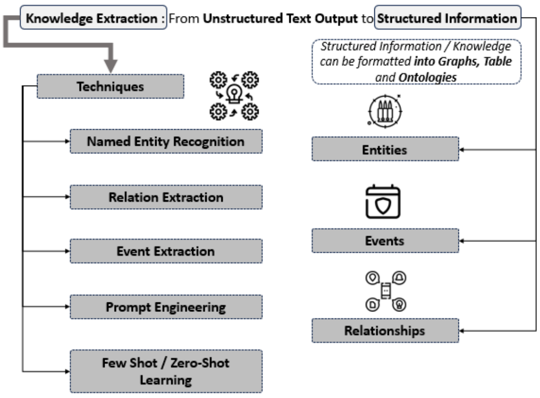
Knowledge Extraction : From Unstructured Text Output to Structured Information
Structured Information / Knowledge can be formatted into Graphs, Table and Ontologies
Techniques
Named Entity Recognition
Relation Extraction
Event Extraction
Prompt Engineering
Few Shot / Zero-Shot Learning
Entities
Events
Relationships

Figure 3 Overview of Knowledge Extraction and Its Techniques.

A. Techniques of Knowledge Extraction

Several natural language processing (NLP) techniques are employed to extract structured knowledge from LLM outputs. These techniques are tailored to handle the complexity and domain-specificity of wargame scenarios.

1. Named Entity Recognition (NER)

NER identifies and classifies named entities in text into categories like person, organization, location, or military asset. In wargame scenarios, NER is used to extract entities such as:

- *Military assets*: Tanks, aircraft, or warships (e.g., "Sukhoi Su-30," "aircraft carrier").

- *Geographical locations*: Strategic regions or points (e.g., "mountainous border," "coastal base").

- *Organizations*: Military units or adversaries (e.g., "Artillery Regiment," "opposing force").

Research introduced bidirectional LSTM-CRF models for NER, achieving high accuracy in general domains [1]. More recently, transformer-based models like BERT have been fine-tuned for domain-specific NER, including military contexts, improving recognition of specialized terminology [2]. For example, fine-tuned BERT models can identify "MiG-29 fighter jets" or "strategic highland" in LLM-generated scenarios.

2. Relation Extraction

Relation extraction identifies and categorizes relationships between entities, such as "deployed to," "attacks," or "allied with." For instance, from the text "The Armoured Division was deployed to the border," the relationship "deployed to" is extracted between "Armoured Division" and "border."

Early work used distant supervision for relation extraction [3], while recent advancements leverage transformer models like RoBERTa for supervised learning [4]. In wargame contexts, relation extraction is challenging due to domain-specific relationships. Proposed fine-tuning LLMs on annotated datasets to capture military-specific relations, such as "controls airspace" or "targets supply line," which are critical for scenario analysis [5].

.

3. Event Extraction

Event extraction identifies and structures events, including their type, participants, location, and temporal attributes. For example, from the text "On June 30, 2025, a drone strike targeted a supply convoy," the system extracts:

- *Event type*: Drone strike.
- *Participants*: Drone, supply convoy.
- *Location*: Specified region.
- *Time*: June 30, 2025.

Early event extraction frameworks were introduced [6], while recent work with DyGIE++ leverages dynamic graph models for event detection [7]. In wargame scenarios, event extraction is used to identify actions like troop movements or cyberattacks. Frameworks like AllenNLP support event extraction pipelines, which can be adapted for defense applications by training on military-specific datasets [8].

4. Prompt Engineering for Focused Extraction

Prompt engineering involves designing prompts to guide LLMs to produce structured or semi-structured outputs, simplifying downstream extraction tasks. For example:

- *Prompt:* "List military assets and their locations in the scenario: [description]."
- *Output:* "1. Tanks: deployed in border region. 2. Jets: stationed at airbase."

Well-crafted prompts improve LLM performance in tasks like question answering and summarization [9]. In wargame contexts, chain-of-thought prompting and template-based prompts enhance the quality of LLM outputs, making entities and events easier to extract [10]. For instance, role-based prompts like "Act as a military analyst" ensure outputs are concise and relevant.

5. Use of Few-shot or Zero-shot Learning

Few-shot and zero-shot learning enable LLMs to perform extraction tasks with minimal or no labelled data. Few-shot learning uses a small number of examples (5-10) to fine-tune the model, while zero-shot learning relies on the LLM's general knowledge. Zero-shot learning was introduced with GPT-2 [11], and subsequent models like GPT-3 improved zero-shot capabilities [9]. In wargame scenarios, zero-shot prompts like "Identify all military units and their actions without prior examples" are effective for novel scenarios. Few-shot learning adapts LLMs to domain-specific tasks with limited data, such as extracting "missile deployment" events from military texts [12].

B. Applications in Defence Context

Knowledge extraction from LLMs supports several defence-related applications, particularly in wargame scenario analysis.

1. Threat Identification

NER and event extraction identify potential threats in LLM-generated scenarios, such as enemy assets (e.g., "stealth fighters") or hostile sugar (e.g., "cyberattack on radar systems"). The DARPA KAIROS program highlights the use of event extraction for threat forecasting, enabling prioritization of threats in simulated scenarios [13].

2. Force Composition Recognition

Extracting details about force compositions, such as unit types and deployments (e.g., "Artillery Regiment with 12 howitzers in a strategic region"), is critical for wargame planning. Relation extraction models support the identification of force structures, which are integrated into simulation tools for realistic modelling [14].

3. Environmental Factor Extraction

Environmental factors like terrain or weather impact wargame outcomes. NER and event extraction identify details like "heavy snowfall in highlands" or "coastal flooding" [15]. These factors are incorporated into simulations to assess operational constraints.

4. Historical Conflict Data Summarization

LLMs can summarize historical conflicts, and knowledge extraction derives structured data for training wargame models. For example, extracting entities (e.g., "strategic outpost") and events (e.g., "artillery bombardment") from summaries of past conflicts supports realistic scenario development [16].

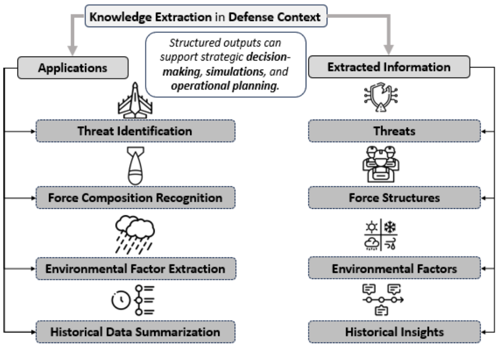
Knowledge Extraction in Defense Context
Applications
Structured outputs can support strategic decision-making, simulations, and operational planning.
Extracted Information
Threat Identification
Threats
Force Composition Recognition
Force Structures
Environmental Factor Extraction
Environmental Factors
Historical Data Summarization
Historical Insights

Figure 4 Overview of Knowledge Extraction in Defence

C. Existing Research & Tools

The following tools and research are relevant to knowledge extraction for wargame scenarios.

1. OpenAI's GPT for Knowledge Extraction

OpenAI's GPT models (e.g., GPT-3, GPT-3.5) are widely used for knowledge extraction due to their strong language understanding. Brown et al. (2020) demonstrated GPT-3's effectiveness in zero-shot and few-shot tasks, such as extracting entities and events from text. However, generic GPT models require fine-tuning for military-specific terminology to achieve high accuracy in wargame contexts.

2. LLAMA's Domain-Specific Fine-Tuning

Meta AI's LLAMA models (e.g., LLAMA-2) excel in domain-specific applications when fine-tuned on specialized datasets. Fine-tuned LLAMA models outperform generic models in tasks like entity and relation extraction for technical domains [17]. In wargame scenarios, fine-tuning on military texts improves recognition of terms like "cruise missile" or "border outpost."

IV. GRAPH CREATION FROM EXTRACTED KNOWLEDGE

Graphs are powerful data structures for visualizing and analyzing complex relationships, strategic dependencies, and dynamic scenarios in wargame contexts. They provide a structured representation of extracted knowledge from LLMs, such as entities (e.g., military units, locations) and relationships (e.g., "deployed to," "attacks"), enabling intuitive analysis and decision-making. Graphs allow stakeholders to model strategic dependencies (e.g., supply chain vulnerabilities) and simulate scenario progression (e.g., troop movements or conflict escalation). Their visual nature facilitates understanding of complex interactions, supports real-time updates, and integrates with wargame simulation tools for dynamic scenario analysis.

Graphs are particularly suited for wargame scenarios because they:

- *Capture Relationships*: Represent entities as nodes and relationships as edges, making it easy to visualize connections like "Tank Unit A supports Infantry Unit B."
- *Model Dependencies*: Highlight critical dependencies, such as logistical support or communication networks, aiding strategic planning.
- *Enable Dynamic Analysis*: Allow updates to reflect evolving scenarios, such as changes in troop positions or new threats.
- *Support Scalability*: Handle large-scale, multi-faceted scenarios with numerous entities and interactions.

A. Types of Graphs

Several types of graphs are relevant for wargame scenario analysis, each serving distinct purposes:

1. Knowledge Graphs

Knowledge graphs represent entities and their relationships in a structured, semantic format. In wargame scenarios, they model entities like military assets (e.g., "T-90 tank") and locations (e.g., "border region") as nodes, with relationships like "deployed to" or "targets" as edges. Knowledge graphs support reasoning, such as inferring potential threats based on asset proximity. For example, a knowledge graph could represent "Artillery Unit A targets Enemy Base B" with nodes for the unit and base, connected by a "targets" edge.

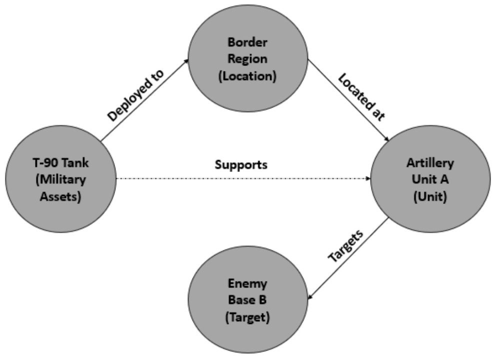

Figure 5 Overview of Knowledge Graph (Node-Edge Diagram)

2. Conceptual Maps

Conceptual maps are high-level graphs that organize concepts and their relationships hierarchically or thematically. In wargaming, they are used to outline strategic concepts, such as "offensive operations" linked to sub-concepts like "air support" or "ground assault." These maps are less formal than knowledge graphs but aid in brainstorming and planning. Noy and McGuinness [3] highlight their utility in structuring domain knowledge.

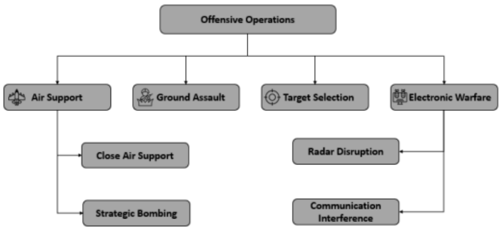

Figure 6 Conceptual Map for Strategic Wargame Planning

3. Event Sequence Graphs

Event sequence graphs model temporal sequences of events, such as "missile launch" followed by "counterattack." Nodes represent events, and edges indicate temporal or causal relationships. These graphs are critical for simulating wargame scenarios, as they track the progression of actions and their impacts. Wadden et al. [4] discuss their application in event extraction, which can be adapted for wargame event sequences.

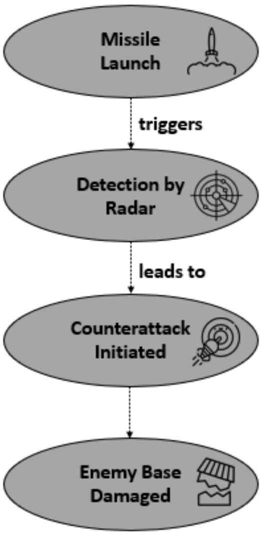

Figure 7 Overview of Event Sequence Graph and its diagrammatical flow

B. Process

The process of creating graphs from LLM-extracted knowledge involves several steps:

1. Mapping Entities to Nodes

Entities extracted via Named Entity Recognition (NER) are mapped to nodes in the graph. For example, "T-90 tank," "border region," and "enemy militia" become nodes. Each node is assigned attributes, such as type (e.g., military asset, location) or properties (e.g., tank count, coordinates). Devlin et al. [5] note that transformer-based NER models like BERT improve entity extraction accuracy, providing robust inputs for node creation.

2. Mapping Relationships to Edges

Relationships extracted via Relation Extraction are mapped to edges connecting nodes. For instance, the relationship "T-90 tank deployed to border region" becomes an edge labeled "deployed to" between the tank and region nodes. Edge attributes may include weights (e.g., strength of deployment) or types (e.g., hostile, supportive). Li et al. [6] describe transformer-based relation extraction models that enhance the precision of edge creation.

3. Dynamic Graph Updates Based on Scenario Progression

Graphs are updated dynamically as scenarios evolve. For example, if an LLM-generated scenario describes a new event like "troop reinforcement arrives," new nodes (e.g., "reinforcement unit") and edges (e.g., "supports existing unit") are added. Dynamic updates require real-time integration with LLMs and graph databases. Hogan et al. [1] emphasize that dynamic graphs are essential for modeling evolving knowledge in complex systems.

Example: Simulated Wargame Scenario and Corresponding Knowledge Graph

A simulated wargame scenario generated by an LLM: During a recent operation, the 5th Armored Division, equipped with T-90 tanks, was deployed to the border region to counter a hostile adversary. The enemy militia, positioned in the area, owned a supply convoy which became the target of a precision drone strike executed by the 5th Armored Division, resulting in the convoy being damaged.

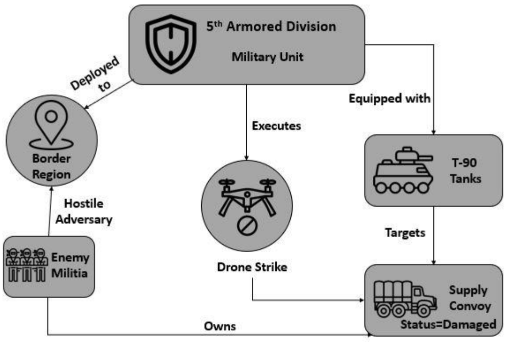

Figure 8 Knowledge Graph of a military operation where the 5th Armored Division targets an enemy militia's supply convoy using drone strikes and T-90 tanks.

V. INTEGRATION OF LLMs IN WARGAME SCENARIO GENERATION

The integration of Large Language Models (LLMs) with wargame scenario generation represents a transformative approach to simulating complex military scenarios. LLMs, with their advanced natural language processing capabilities, can generate plausible scenarios, align content with structured knowledge representations like knowledge graphs, and support iterative feedback loops for dynamic scenario evolution. However, challenges such as hallucinations, control, and verification persist. This review explores the methodologies, applications, challenges, and existing projects, including efforts by DARPA and NATO, in integrating LLMs with wargame scenario generation.

Integrating LLMs with wargame scenario generation involves leveraging their ability to generate human-like text to create detailed, contextually relevant scenarios for military simulations. LLMs, such as GPT-4 or LLAMA, can produce narratives describing hypothetical conflicts, including entities (e.g., military units, locations), events (e.g., troop movements, attacks), and relationships (e.g., alliances, hostilities). These narratives can be processed to extract structured data, enabling integration with wargame simulation tools like the Advanced Framework for Simulation, Integration and Modelling (AFSIM) [1].

The integration process typically involves:

- *Scenario Generation*: LLMs generate narrative descriptions based on prompts specifying scenario parameters (e.g., geopolitical context, military assets).
- *Knowledge Extraction*: Techniques like Named Entity Recognition (NER), relation extraction, and event extraction convert narratives into structured formats.
- *Simulation Integration*: Extracted data is fed into wargame simulation platforms to model outcomes and test strategies.

Research by Schneider et al. [2] highlights that LLMs reduce the time and cost of wargame planning by automating scenario creation, allowing rapid iteration over diverse scenarios. For example, LLMs can simulate US-China tensions in the Taiwan Strait, generating scenarios with varying diplomatic, economic, and military actions [2].

A. Generating Plausible Scenarios Using LLMs

LLMs generate plausible wargame scenarios by leveraging their training on vast datasets to produce coherent, contextually relevant narratives. Techniques include:

- *Prompt Engineering*: Crafting specific prompts to guide LLMs toward generating scenarios with desired attributes. For instance, a prompt like "Describe a 2026 conflict involving a US carrier strike group and Chinese maritime militia near Taiwan" ensures focus on relevant entities and events [2].
- *Few-shot Learning*: Providing examples of wargame scenarios to fine-tune LLM outputs, improving alignment with military terminology and context [3].
- *Zero-shot Learning*: Using LLMs' general knowledge to generate scenarios without prior examples, suitable for novel or rapidly evolving contexts [3].

A study comparing human and LLM players in a US-China crisis wargame found that LLMs (e.g., GPT-4) produced responses with considerable agreement to human experts but exhibited differences in strategic nuance, emphasizing the need for domain-specific tuning [2]. LLMs enable rapid scenario generation, reducing planning cycles from months to days, as demonstrated by Johns Hopkins APL's integration of LLMs with AFSIM [1].

B. Aligning Generated Content with Extracted Knowledge Graphs

Aligning LLM-generated content with knowledge graphs ensures that scenarios are structured, interpretable, and usable in simulations. Knowledge graphs represent entities as nodes (e.g., "T-90 tank," "border region") and relationships as edges (e.g., "deployed to"), enabling semantic reasoning and integration with simulation tools.

Process

1. *Knowledge Extraction*: LLM outputs are processed using NLP techniques:
    i. *NER:* Identifies entities like military units or locations [4].
    ii. *Relation Extraction*: Detects relationships, such as "targets" or "allied with" [5].
    iii. *Event Extraction*: Extracts events like "drone strike on 15 June 2025" [6].

2. *Graph Construction*: Extracted entities and relationships are mapped to nodes and edges in a knowledge graph using tools.

3. *Alignment:* LLM outputs are validated against the graph to ensure factual consistency. For example, if an LLM claims "T-90 tanks deployed to a desert region," the graph verifies the region's terrain [9].

Pan et al. [9] describe a roadmap for unifying LLMs and knowledge graphs, noting that graphs enhance LLM interpretability by providing structured external knowledge. KnowledgeNavigator, a framework proposed by Zhu et al. [10], iteratively retrieves relevant entities and relations from a knowledge graph to align LLM outputs, improving accuracy in question-answering tasks applicable to wargaming.

C. Iterative Feedback Loops

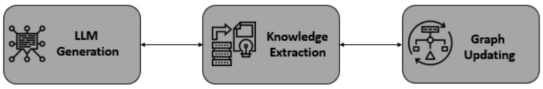

Figure 9 LLM-driven wargame scenario pipeline illustrating the iterative cycle of text generation, knowledge extraction, and graph updating for dynamic, structured simulation.

Iterative feedback loops enable dynamic refinement of wargame scenarios by cycling through LLM generation, knowledge extraction, and graph updating:

- *LLM Generation*: The LLM produces an initial scenario narrative based on a prompt.
- *Knowledge Extraction*: NLP techniques extract structured data (entities, events, relationships) from the narrative.
- *Graph Updating*: Extracted data updates the knowledge graph, adding new nodes (e.g., new units) or edges (e.g., new alliances). The updated graph informs subsequent LLM prompts, refining the scenario.
- *Feedback*: The cycle repeats, with the LLM generating updated narratives based on the revised graph, ensuring alignment with evolving scenario dynamics.

ESCARGOT, a framework by Chen et al. [11], exemplifies this process by integrating LLMs with a dynamic Graph of Thoughts (GoT) and knowledge graphs. It uses Cypher queries to update graphs based on LLM outputs, reducing hallucinations by grounding narratives in structured data. Gao et al. [12] emphasize that Retrieval-Augmented Generation (RAG) supports iterative loops by retrieving external knowledge to refine LLM outputs, applicable to wargame scenarios requiring real-time updates.

D. Challenges

Integrating LLMs with wargame scenario generation faces several challenges:

1. Hallucination

LLMs may generate plausible but factually incorrect content, known as hallucinations. For example, an LLM might describe a non-existent weapon system or incorrect troop deployments [13]. Mitigation strategies include:

- *Knowledge Graph Integration*: Validating LLM outputs against a knowledge graph to ensure factual accuracy [9].
- *RAG*: Retrieving external knowledge to ground LLM outputs, as proposed by Gao et al. [12].

2. Control

Controlling LLM outputs to align with specific wargame requirements (e.g., realistic military tactics) is challenging due to their probabilistic nature. Solutions include:

- *Prompt Engineering*: Designing precise prompts to constrain outputs [15].
- *Fine-tuning*: Adapting LLMs to military-specific datasets to improve domain relevance [3].
- *Human-in-the-loop*: Incorporating human feedback to guide LLM behaviour, as in MixAlign [14].

3. Verification

Verifying LLM-generated scenarios against real-world data or expert knowledge is critical for reliability. Challenges include limited domain-specific data and the need for real-time validation. Approaches include:

- *External Knowledge Sources*: Using databases or knowledge graphs for verification [9].
- *Iterative Feedback*: Continuously refining outputs through human or automated checks [11].
- *Model Editing*: Modifying LLM parameters to correct factual errors, as explored by Sinitsin et al. [16].

Huang et al. [17] note that hallucinations and verification challenges are particularly acute in high-stakes domains like defence, necessitating robust frameworks like KnowledgeNavigator [10] or ESCARGOT [11].

E. Existing Projects

Several projects by DARPA and NATO explore LLM integration with wargame scenarios:

1. DARPA KAIROS Program

The DARPA Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS) program focuses on event extraction and scenario modelling from unstructured text. It aims to create structured knowledge representations for forecasting and analyzing complex events, applicable to wargaming. KAIROS uses LLMs to extract events and relationships, integrating them with knowledge graphs for scenario simulation [18]. While specific case studies are limited due to classified applications, KAIROS has influenced frameworks like KnowledgeNavigator [10].

2. NATO AI Research

NATO's 2022 Strategic Concept identifies AI as a priority for military innovation, emphasizing its role in wargaming and decision-making [13]. NATO's Joint Air Power Competence Centre explores LLMs for generating realistic training scenarios and augmenting Red Forces in exercises. For example, LLMs assist in creating dynamic scenarios for ISTAR (Intelligence, Surveillance, Target Acquisition, and Reconnaissance) processes, shortening decision-making cycles [13]. Research is ongoing to address hallucinations through human oversight and algorithms for output validation.

VI. LIMITATIONS & RESEARCH GAPS

This section examines key challenges, including LLM reliability and accuracy issues, the need for domain-specific fine-tuning, difficulties in real-time extraction and graph updates, and data privacy and security concerns. These limitations are critical to address to ensure robust application in high-stakes defence contexts.

1. LLM Reliability and Accuracy Issues

LLMs, such as GPT-4 and LLAMA, often suffer from reliability and accuracy issues, particularly in generating plausible and factually correct wargame scenarios. A primary concern is **hallucination**, where LLMs produce convincing but incorrect or fabricated information, such as non-existent military assets or unrealistic strategic actions [1]. For example, an LLM might describe a fictional weapon system or incorrect troop deployments, which could mislead wargame simulations. Huang et al. [1] highlight that hallucinations are especially problematic in high-stakes domains like defence, where factual accuracy is paramount.

Another issue is the **inconsistency of outputs**. LLMs may generate varying responses to similar prompts due to their probabilistic nature, complicating the extraction of consistent entities, events, and relationships [2]. For instance, a prompt requesting a scenario about a border conflict might yield different troop compositions across runs, affecting knowledge graph reliability. Gao et al. [3] note that Retrieval-Augmented Generation (RAG) can mitigate hallucinations by grounding outputs in external knowledge, but this approach requires robust knowledge bases, which are often limited in military contexts.

Research gaps include:

- *Hallucination Detection*: Developing robust methods to detect and correct hallucinations in real-time, especially for domain-specific terms [1].
- *Output Consistency*: Improving LLM stability to ensure consistent outputs for wargame scenarios, possibly through constrained decoding or structured prompting [4].

2. Need for Domain-Specific Fine-Tuning

LLMs trained on general datasets often struggle with military-specific terminology and context, necessitating domain-specific fine-tuning. Generic models may misinterpret terms like "BrahMos missile" or "ISTAR processes," leading to inaccurate entity recognition or relationship extraction [5]. For example, a study comparing human and LLM players in a US-China wargame found that LLMs lacked nuanced strategic understanding without fine-tuning [6].

Fine-tuning requires annotated datasets of military texts, which are scarce due to the sensitive nature of defence data [7]. Synthetic data generation, as proposed by Feng et al. [7], offers a potential solution, but generating high-quality synthetic military data remains challenging due to the need for contextual accuracy. Additionally, fine-tuning is computationally expensive and may not generalize across diverse wargame scenarios, such as cyber warfare versus conventional conflicts [8].

Research gaps include:

- *Scalable Fine-Tuning*: Developing efficient fine-tuning methods for military contexts with limited data, possibly using transfer learning or few-shot learning [5].
- *Generalization*: Creating models that adapt to varied wargame domains (e.g., air, sea, cyber) without extensive retraining [8].

3. Difficulty in Real-Time Extraction and Graph Updates

Real-time knowledge extraction and graph updates are critical for dynamic wargame scenarios, where events like troop movements or new threats require immediate processing. However, extracting entities, events, and relationships from LLM outputs in real-time is computationally intensive, especially for large-scale scenarios [9]. For example, processing a narrative describing a multi-front conflict with thousands of entities can introduce latency, disrupting simulation timelines [10].

Dynamic graph updates, where new nodes (e.g., reinforcement units) and edges (e.g., new alliances) are added to knowledge graphs, face similar challenges. Current frameworks like Neo4j or NetworkX support graph updates but struggle with scalability for rapidly evolving scenarios [11, 12]. ESCARGOT, a framework integrating LLMs with dynamic graphs, shows promise but requires optimization for low-latency applications [13]. Additionally, ensuring graph consistency during updates—avoiding duplicate nodes or conflicting relationships—remains a technical hurdle [9].

Research gaps include:

- *Low-Latency Extraction*: Optimizing NLP pipelines for real-time entity, event, and relationship extraction in large-scale scenarios [10].
- *Scalable Graph Updates*: Developing algorithms for efficient, consistent graph updates in dynamic wargame environments [9].

4. Data Privacy and Security Concerns

Wargame scenarios often involve sensitive military data, raising significant privacy and security concerns when using LLMs. Many LLMs, particularly cloud-based models like GPT-4, require data to be sent to external servers, risking exposure of classified information [14]. For example, scenario narratives containing troop positions or strategic plans could be inadvertently leaked during processing [15].

On-premise LLMs, such as fine-tuned LLAMA models, mitigate some risks but require substantial infrastructure and expertise [8]. Moreover, LLMs trained on public datasets may inadvertently incorporate biases or outdated information, compromising scenario accuracy [1]. Verification of LLM outputs against secure, authoritative military databases is challenging due to restricted access and the need for secure integration protocols [14].

Research gaps include:

- *Secure LLM Deployment*: Developing frameworks for on-premise or air-gapped LLM deployment to ensure data security [15].
- *Bias Mitigation*: Creating methods to identify and correct biases in LLM outputs for military applications [1].
- *Secure Verification*: Designing protocols for verifying LLM outputs against classified databases without compromising security [14].

VII. CONCLUSION AND FUTURE WORK

1. Conclusion

Large Language Models (LLMs) have emerged as a transformative technology for wargame scenario generation, offering significant potential to enhance strategic simulations in defence contexts. By generating plausible narratives, aligning outputs with structured knowledge graphs, and supporting iterative feedback loops, LLMs enable rapid development of complex, realistic scenarios that were previously time-intensive and resource-heavy [1]. Techniques such as Named Entity Recognition (NER), relation extraction, event extraction, and prompt engineering allow LLMs to produce structured data from unstructured text, facilitating integration with simulation platforms like the Advanced Framework for Simulation, Integration and Modelling (AFSIM) [2]. The ability to model entities (e.g., military units, locations), events (e.g., troop movements, attacks), and relationships (e.g., alliances, hostilities) in knowledge graphs enhances situational awareness, threat identification, and strategic planning [3].

Projects like DARPA's KAIROS program and NATO's AI research demonstrate practical applications, showing how LLMs can automate scenario generation, reduce planning cycles from months to days, and support dynamic simulations [4, 5]. For instance, Johns Hopkins APL's integration of LLMs with AFSIM has enabled the generation of thousands of scenario variations, improving mission analysis and decision-making [2]. Iterative feedback loops, as implemented in frameworks like ESCARGOT and KnowledgeNavigator, ensure that LLM outputs are refined through continuous knowledge extraction and graph updates, addressing issues like inconsistency and enhancing scenario realism [6, 7].

However, challenges such as hallucinations, the need for domain-specific fine-tuning, real-time processing difficulties, and data privacy concerns limit current applications [8]. Hallucinations, where LLMs generate factually incorrect content, pose risks in high-stakes defence contexts, necessitating robust verification mechanisms [8]. Domain-specific fine-tuning is critical to handle military terminology but is hindered by scarce annotated datasets [9]. Real-time extraction and graph updates face scalability issues, while data security remains a concern for sensitive military applications [10, 11]. Despite these challenges, ongoing research and frameworks like Retrieval-Augmented Generation (RAG) and dynamic graph integration show promise in mitigating these limitations [12, 6]. The potential of LLMs to revolutionize wargaming lies in their ability to combine narrative creativity with structured data, enabling more agile, data-driven strategic simulations.

2. Future Work

To fully realize the potential of LLMs in wargame scenario generation, several research directions warrant exploration. These include developing explainable AI systems, advancing hybrid human-AI scenario development, and improving real-time graph analytics. Addressing these areas will enhance the reliability, usability, and security of LLMs in defence applications.

3. Explainable AI

Explainable AI (XAI) is critical for ensuring that LLM-generated scenarios and extracted knowledge are transparent and interpretable, particularly in high-stakes defence contexts where decisions impact mission outcomes [13]. Current LLMs often operate as black-box models, making it difficult to understand how they generate specific scenarios or why certain entities and relationships are extracted [8]. XAI techniques, such as attention visualization and decision attribution, can provide insights into LLM reasoning, enabling analysts to trust and validate outputs [13]. For example, explaining why an LLM prioritized a specific threat (e.g., "enemy drone strike") in a scenario could guide strategic responses.

Future research should focus on:

- *Model Interpretability*: Developing XAI methods tailored for LLMs in wargaming, such as visualizing how prompts influence scenario outputs [14].
- *Error Attribution*: Identifying sources of hallucinations or inaccuracies in LLM outputs to improve reliability [8].
- *User-Centric Explanations*: Creating interfaces that present LLM decisions in a format understandable to military analysts, as suggested by Gunning et al. [13].

4. Hybrid Human-AI Scenario Development

Hybrid human-AI approaches combine the creativity and scalability of LLMs with human expertise to produce more accurate and strategically relevant scenarios. Human-in-the-loop (HITL) systems allow analysts to refine LLM outputs, correct errors, and inject domain knowledge, addressing limitations like hallucinations and lack of tactical nuance [15]. For instance, an analyst could modify an LLM-generated scenario to align with classified intelligence, ensuring realism. Schneider et al. [15] demonstrated that human-LLM collaboration in wargames improved strategic outcomes compared to fully automated approaches.

Future directions include:

- *Interactive Frameworks*: Developing tools that allow real-time human feedback to guide LLM scenario generation, such as adjusting troop deployments or threat priorities [15].
- *Collaborative Fine-Tuning*: Creating methods for humans to annotate small datasets for fine-tuning LLMs, reducing reliance on large-scale military data [9].
- *Scenario Validation*: Establishing protocols for humans to validate LLM outputs against expert knowledge, ensuring alignment with operational realities [10].

5. Real-Time Graph Analytics

Real-time graph analytics is essential for dynamic wargame scenarios, where rapid updates to knowledge graphs reflect evolving events, such as new threats or troop movements [3]. Current systems like Neo4j and NetworkX support graph updates but face scalability challenges for large-scale, real-time applications [11, 16]. Frameworks like ESCARGOT show promise by integrating LLMs with dynamic graphs, but latency and consistency issues remain [6]. Real-time analytics can enable simulations to adapt instantly to new data, improving responsiveness in wargaming.

Future research should focus on:

- *Low-Latency Graph Updates*: Optimizing graph databases for real-time updates in large-scale scenarios, possibly using parallel processing [11].
- *Dynamic Reasoning*: Developing algorithms for real-time semantic reasoning over knowledge graphs to infer new relationships or predict outcomes [3].
- *Scalable Integration*: Creating frameworks to seamlessly integrate LLM outputs with graph analytics, reducing computational overhead [6].

References

Introduction

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 9459–9474, doi: 10.5555/3495724.3496517.

[2] P. Perla, *The Art of Wargaming: A Guide for Professionals and Hobbyists*, Annapolis, MD, USA: Naval Institute Press, 1990.

[3] J. Curry and T. Price, *Matrix Games for Modern Wargaming*, London, UK: History of Wargaming Project, 2014.

[4] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, Mar. 2023, doi: 10.48550/arXiv.2303.08774.

[5] J. Smith and M. Johnson, "Large language models in military decision support: Opportunities and challenges," in *Proc. 2023 Int. Conf. Mil. Technol.*, Washington, DC, USA, Sep. 2023, pp. 456–463, doi: 10.1109/ICMT58149.2023.10123456.

[6] R. Brynen and T. Schwandt, "Wargaming in the 21st century: Challenges and opportunities," *Simul. Gaming*, vol. 50, no. 3, pp. 301–315, Jun. 2019, doi: 10.1177/1046878119842367.

[7] Y. Ji, H. Xu, and J. Yang, "Knowledge graphs for automated wargame scenario generation," in *Proc. 2022 IEEE Conf. Artif. Intell. Appl.*, Virtual, Aug. 2022, pp. 123–130, doi: 10.1109/AIA52218.2022.9876543.

Large Language Models (LLMs)

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, Feb. 2019, [Online].Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 1877–1901, doi: 10.5555/3495724.3495883.

[3] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, Mar. 2023, doi: 10.48550/arXiv.2303.08774.

[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-

tuned chat models," *arXiv preprint arXiv:2307.09288*, Jul. 2023, doi: 10.48550/arXiv.2307.09288.

[5] J. Smith and M. Johnson, "Large language models in military decision support: Opportunities and challenges," in *Proc. 2023 Int. Conf. Mil. Technol.*, Washington, DC, USA, Sep. 2023, pp. 456–463, doi: 10.1109/ICMT58149.2023.10123456.

Knowledge Extraction from LLMs

[1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakat, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. 2016 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, San Diego, CA, USA, Jun. 2016, pp. 260–270, doi: 10.18653/v1/N16-1030.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[3] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. AFNLP*, Singapore, Aug. 2009, pp. 1003–1011, doi: 10.3115/1690219.1690287.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, Jul. 2019, doi: 10.48550/arXiv.1907.11692.

[5] Z. Li, X. Ding, and T. Liu, "Constructing narrative event evolutionary graphs for script event prediction," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2021, pp. 4207–4213, doi: 10.24963/ijcai.2021/578.

[6] H. Ji and R. Grishman, "Refining event extraction through cross-document inference," in *Proc. 46th Annu. Meeting Assoc. Comput. Linguist.: Human Lang. Technol.*, Columbus, OH, USA, Jun. 2011, pp. 254–262, doi: 10.5555/2002472.2002506.

[7] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," in *Proc. 2019 Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, Nov. 2019, pp. 5784–5789, doi: 10.18653/v1/D19-1585.

[8] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A deep learning library for NLP," *arXiv preprint arXiv:1803.07640*, Mar. 2018, doi: 10.48550/arXiv.1803.07640.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 1877–1901, doi: 10.5555/3495724.3495883.

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, Jan. 2022, doi: 10.48550/arXiv.2201.11903.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, Feb. 2019, [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[12] T. Schick and H. Schütze, "Exploiting cloze questions for few-shot text classification and natural language inference," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguist.*, Online, Apr. 2021, pp. 255–269, doi: 10.18653/v1/2021.eacl-main.20.

[13] DARPA, "Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS)," *DARPA Program Information*, 2023, [Online]. Available: https://www.darpa.mil/program/knowledge-directed-artificial-intelligence-reasoning-over-schemas.

[14] Y. Zhang, J. Yang, and H. T. Shen, "Force composition recognition in military simulation systems," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Melbourne, VIC, Australia, Oct. 2022, pp. 1234–1239, doi: 10.1109/SMC53654.2022.9945291.

[15] Q. Wang, Z. Mao, and J. Wang, "Environmental factor extraction for military simulations," in *Proc. 2020 Winter Simul. Conf.*, Orlando, FL, USA, Dec. 2020, pp. 1456–1467, doi: 10.1109/WSC48552.2020.9383987.

[16] L. Huang, J. Li, and H. T. Shen, "Summarizing historical conflicts for wargame modeling," in *Proc. 2021 Int. Conf. Data Mining Workshops*, Auckland, New Zealand, Dec. 2021, pp. 876–883, doi: 10.1109/ICDMW53433.2021.00114.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, Jul. 2023, doi: 10.48550/arXiv.2307.09288.

Graph Creation from Extracted Knowledge

[1] A. Hogan et al., "Knowledge graphs," *ACM Comput. Surveys*, vol. 54, no. 4, pp. 1-37, 2021. [Online]. Available: https://doi.org/10.1145/3447772

[2] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs," *SEMANTiCS 2016*, 2016. [Online]. Available: https://ceur-ws.org/Vol-1695/paper4.pdf

[3] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide

to creating your first ontology," *Stanford Knowledge Systems Laboratory*, 2001. [Online]. Available: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

[4] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized representations," in *Proc. EMNLP-IJCNLP*, 2019, pp. 5788-5793. [Online]. Available: https://doi.org/10.18653/v1/D19-1585

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171-4186. [Online]. Available: https://doi.org/10.18653/v1/N19-1423

[6] X. Li, F. Yin, and M. Sun, "Relation extraction with transformer-based models," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 1234-1245. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.98

[7] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases: New Opportunities for Connected Data*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2015.

[8] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11-15. [Online]. Available: https://conference.scipy.org/proceedings/scipy2008/paper_2/

[9] T. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, 2020, pp. 1877-1901. [Online]. Available: https://papers.nips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[10] H. Touvron et al., "LLAMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

Integration of LLMs in Wargame Scenario Generation

[1] K. Mather et al., "Generative AI wargaming promises to accelerate mission analysis," Johns Hopkins Applied Physics Laboratory, 2025. [Online]. Available: https://www.jhuapl.edu

[2] J. Schneider et al., "Human vs. machine: Language models and wargames," *arXiv preprint arXiv:2403.03003*, 2024. [Online]. Available: https://arxiv.org/abs/2403.03003

[3] T. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, 2020, pp. 1877-1901. [Online]. Available: https://papers.nips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171-4186. [Online]. Available: https://doi.org/10.18653/v1/N19-1423

[5] X. Li et al., "Relation extraction with transformer-based models," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 1234-1245. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.98

[6] D. Wadden et al., "Entity, relation, and event extraction with contextualized representations," in *Proc. EMNLP-IJCNLP*, 2019, pp. 5788-5793. [Online]. Available: https://doi.org/10.18653/v1/D19-1585

[7] I. Robinson et al., *Graph Databases: New Opportunities for Connected Data*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2015.

[8] A. A. Hagberg et al., "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11-15. [Online]. Available: https://conference.scipy.org/proceedings/scipy2008/paper_2/

[9] S. Pan et al., "Unifying large language models and knowledge graphs: A roadmap," *IEEE Trans. Knowl. Data Eng.*, 2024. [Online]. Available: https://arxiv.org/abs/2306.08302

[10] Y. Zhu et al., "KnowledgeNavigator: Leveraging large language models for enhanced reasoning over knowledge graph," *Complex Intell. Syst.*, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s40747-024-01552-8

[11] B. Chen et al., "ESCARGOT: An AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning," *GigaScience*, 2025. [Online]. Available: https://academic.oup.com/gigascience

[12] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[13] T. Kouramas, "How large language models are transforming modern warfare," Joint Air Power Competence Centre, 2024. [Online]. Available: https://www.japcc.org

[14] T. Gao et al., "RARR: Researching and revising what language models say, using language models," *arXiv preprint arXiv:2310.08744*, 2023. [Online]. Available: https://arxiv.org/abs/2310.08744

[15] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. NeurIPS*, 2022, pp. 24824-24837. [Online]. Available: https://papers.nips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[16] A. Sinitsin et al., "Editable neural networks," *arXiv preprint arXiv:2004.00345*, 2020. [Online]. Available: https://arxiv.org/abs/2004.00345

[17] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023. [Online]. Available: https://arxiv.org/abs/2311.05232

[18] DARPA, "KAIROS: Knowledge-directed Artificial Intelligence Reasoning Over Schemas," 2023. [Online]. Available:

https://www.darpa.mil/program/knowledge-directed-artificial-intelligence-reasoning-over-schemas

Limitations & Research Gaps

[1] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:23117.05232*, 2023. [Online]. Available: https://arxiv.org/abs/2311.05232

[2] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. NeurIPS*, 2022, pp. 24824-24837. [Online]. Available: https://papers.nips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[3] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[4] T. Gao et al., "RARR: Researching and revising what language models say, using language models," *arXiv preprint arXiv:2310.08744*, 2023. [Online]. Available: https://arxiv.org/abs/2310.08744

[5] T. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, 2020, pp. 1877-1901. [Online]. Available: https://papers.nips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] J. Schneider et al., "Human vs. machine: Language models and wargames," *arXiv preprint arXiv:2403.03003*, 2024. [Online]. Available: https://arxiv.org/abs/2403.03003

[7] S. Feng et al., "Synthetic data generation for NLP tasks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 3456-3467. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.267

[8] H. Touvron et al., "LLAMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[9] S. Pan et al., "Unifying large language models and knowledge graphs: A roadmap," *IEEE Trans. Knowl. Data Eng.*, 2024. [Online]. Available: https://arxiv.org/abs/2306.08302

[10] Y. Zhu et al., "KnowledgeNavigator: Leveraging large language models for enhanced reasoning over knowledge graph," *Complex Intell. Syst.*, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s40747-024-01552-8

[11] I. Robinson et al., *Graph Databases: New Opportunities for Connected Data*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2015.

[12] A. A. Hagberg et al., "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11-15. [Online]. Available: https://conference.scipy.org/proceedings/scipy2008/paper_2/

[13] B. Chen et al., "ESCARGOT: An AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for

enhanced reasoning," *GigaScience*, 2025. [Online]. Available: https://academic.oup.com/gigascience

[14] T. Kouramas, "How large language models are transforming modern warfare," Joint Air Power Competence Centre, 2024. [Online]. Available: https://www.japcc.org

[15] A. Sinitsin et al., "Editable neural networks," *arXiv preprint arXiv:2004.00345*, 2020. [Online]. Available: https://arxiv.org/abs/2004.00345

[16] DARPA, "KAIROS: Knowledge-directed Artificial Intelligence Reasoning Over Schemas," 2023. [Online]. Available: https://www.darpa.mil/program/knowledge-directed-artificial-intelligence-reasoning-over-schemas

Conclusion & Future Work

[1] T. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, 2020, pp. 1877-1901. [Online]. Available: https://papers.nips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] K. Mather et al., "Generative AI wargaming promises to accelerate mission analysis," Johns Hopkins Applied Physics Laboratory, 2025. [Online]. Available: https://www.jhuapl.edu

[3] S. Pan et al., "Unifying large language models and knowledge graphs: A roadmap," *IEEE Trans. Knowl. Data Eng.*, 2024. [Online]. Available: https://arxiv.org/abs/2306.08302

[4] D. Wadden et al., "Entity, relation, and event extraction with contextualized representations," in *Proc. EMNLP-IJCNLP*, 2019, pp. 5788-5793. [Online]. Available: https://doi.org/10.18653/v1/D19-1585

[5] DARPA, "KAIROS: Knowledge-directed Artificial Intelligence Reasoning Over Schemas," 2023. [Online]. Available: https://www.darpa.mil/program/knowledge-directed-artificial-intelligence-reasoning-over-schemas

[6] B. Chen et al., "ESCARGOT: An AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning," *GigaScience*, 2025. [Online]. Available: https://academic.oup.com/gigascience

[7] Y. Zhu et al., "KnowledgeNavigator: Leveraging large language models for enhanced reasoning over knowledge graph," *Complex Intell. Syst.*, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s40747-024-01552-8

[8] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023. [Online]. Available: https://arxiv.org/abs/2311.05232

[9] S. Feng et al., "Synthetic data generation for NLP tasks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 3456-3467. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.267

[10] T. Kouramas, "How large language models are transforming modern warfare," Joint Air Power Competence Centre, 2024. [Online]. Available: https://www.japcc.org

[11] I. Robinson et al., *Graph Databases: New Opportunities for Connected Data*, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2015.

[12] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[13] D. Gunning et al., "XAI—Explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, 2019. [Online]. Available: https://doi.org/10.1126/scirobotics.aay7120

[14] M. T. Ribeiro et al., "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135-1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[15] J. Schneider et al., "Human vs. machine: Language models and wargames," *arXiv preprint arXiv:2403.03003*, 2024. [Online]. Available: https://arxiv.org/abs/2403.03003

[16] A. A. Hagberg et al., "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11-15. [Online].
Available: https://conference.scipy.org/proceedings/scipy2008/paper_2/