Leveraging Large Language Models for Enhanced Wargaming in Multi-Domain Operations

Dominic Weller
Max Meltschack
Dominik Schwindling
Bundeswehr Office for Defence Planning
Lilienthalstr. 12, 82024 Taufkirchen
GERMANY

Dominic1Weller@bundeswehr.org, MaxMeltschack@bundeswehr.org.
DominikSchwindling@bundeswehr.org

ABSTRACT

Wargaming is a pivotal tool in military training, strategic planning and operational readiness. However, traditional and digital wargames face challenges in scalability, immersion and regarding the availability of subject matter experts during development and execution. In this paper we try to combine wargaming with Large Language Models (LLMs), a disruptive technology which garnered significant attention as it is capable to generate human-like text based on vast datasets, making them powerful tools for natural language processing tasks in order to address the challenges of wargaming. For this, we present a prototype initially developed during the NATO TIDE Hackathon 2024, demonstrating the practical integration of LLMs in quantitative wargames in the context of multi-domain operations. We highlight the applicability and potential of LLM in automating content generation, acting as interactive facilitator and information provider and finally in enhancing immersion of digital wargames by action masking. However, we also underscore limitations, especially when it comes to simulating complex adversarial behaviour. In our conclusion we emphasize the need for well-defined small-scale implementations of LLM in order to leverage their practical benefits.

1.0 INTRODUCTION

Wargaming has long been a crucial tool for military training, strategic planning, and operational readiness [1]. It allows military personnel to simulate complex scenarios, examine novel concepts, and understand potential outcomes without the risks and costs associated with real-world exercises [2]. As early as 1824, wargames were used in the sense of the Prussian Kriegsspiel to prepare military leaders for the challenges of modern warfare [3]. While the focus at the time was on the use of novel topographical maps, contemporary wargames can be used to bring the complexity of multi-domain operations (MDO) into the consciousness of modern military leaders [4]. In order to generate a decisive advantage on the battlefield of the future, they must be empowered to orchestrate their activities across all operational domains, also with non-military stakeholders, beyond the principles of previous joint operations. Digital wargames represent a cost-effective measure to simulate high-pace multi-domain scenarios, in which soldiers are forced to make quick decisions and to deal with increasing complexity. [5]

However, mapping networked dependencies and realistic scenarios of a progressively digitized (military) world is becoming increasingly difficult for conventional wargames and poses challenges for developers.
It is within this context that the transformative potential of Large Language Models (LLMs) becomes particularly relevant. LLMs, such as OpenAI's GPT-4, have garnered significant attention and acclaim in recent years for their ability to generate human-like text based on vast datasets, making them powerful tools for natural language processing tasks [6]. These models have emerged as disruptive and

transformative technologies with widespread applications across various domains. More sophisticated models are even able to handle multimodal inputs and outputs, such as images, text, or audio, enhancing their utility [7][8][9].

The implications of LLMs on military matters are already recognizable and they are expected to even become more relevant in the future [10][11]. However, the enthusiasm surrounding LLMs carries the risk of neglecting limitations and dangers arising from unconsidered usage. One significant concern is their susceptibility to bias, as they learn from large datasets that may contain prejudiced or unbalanced information, potentially leading to biased outputs [12]. Additionally, LLMs can be used to generate misinformation, posing threats to information integrity and public trust [13][14]. Their deployment also raises ethical questions regarding privacy, as they may inadvertently reveal sensitive information from their training data [15][14]. The computational resources required to train and operate LLMs are substantial, leading to high costs and environmental impacts due to energy consumption [16][17][18][19][14]. Moreover, there are concerns about over-reliance on LLMs, which might reduce human critical thinking and decision-making skills in certain contexts [20].

Some of these aspects, including the benefits of potential practical utility as well as the limitations, will be discussed in relation to wargames, illustrated with an example. Since Wargaming and LLM offer significant advantages on their own, our aim is now to create synergies through their combination. These circumstances can also be seen as an incentive to consider the combination of LLMs and wargames.

Our research question focuses on how LLMs can enhance wargaming for multi-domain operations by addressing specific challenges such as scalability, immersion, and the scarcity of subject matter experts. A prototype initially developed during the NATO TIDE Hackathon 2024 in Amsterdam demonstrates possible practical applications of this potential integration in varying scopes and complexities, providing a basis for further exploration, development, and ideas in this field.

2.0 METHODOLOGY

We will first outline the wargame we developed and its corresponding game mechanics, initially excluding the use of LLMs. Next, we will describe the implementation and architecture of the software. After this, an initial literature review to highlight similar approaches and better contextualize our demonstrator is performed.

2.1 Concept of the Wargame

The developed wargame draws upon foundational literature in the field, including the NATO Wargaming Handbook [5] and the Wargaming Handbook of the German Armed Forces [22] reflecting key elements of wargaming, such as opposing forces, scripted injects, and the competition for and prioritization of scarce resources. Our wargame immerses a single player in a scenario set in the Baltic States with a focus on multi-domain operations unfolding across military (land, air, sea forces), political, economic, cyber, and space domains. The player takes on the role of the defender, while an aggressor challenges him and tries to capture several strategic locations. For this purpose, the aggressor is able to take up to three actions in each of the ten rounds. In the same manner, the defender also can take three actions per round to influence the MDO environment in order to defend himself and to deescalate the situation.

A single action is represented by deploying a game card which implies a specific impact on the scores of the player. Each game card is assigned to one of three categories (Military, Cyber-Hybrid, Political-Economic) and therefore affects different scores.

The first score is the escalation score which tracks the tension level between the two factions. This score is mainly influenced by playing cards of the military and political domain and can be increased (e.g. deployment of troops in the border region) or decreased (e.g. diplomatic efforts). The effectiveness of political and economic actions is influenced by the morale score, which indicates the public sentiment toward the actions of the corresponding faction. Moreover, the communication score affects the capability to coordinate military troops, which has influence on the striking power of military actions. This score can be lowered, for example, as a result of attacks in the cyber domain, or increased as a result of the establishment of redundant communication channels. The communication score serves in particular to sensitize players to the consequences of a partial or complete failure of the communication link on their command and control capabilities in an increasingly digitized world. Finally, the resource score gives information about the availability of resources which are needed to take actions in order to sustain strategic initiatives and capabilities throughout the game.

Whenever the aggressor decides that they have sufficiently influenced or prepared both their own and the opponent's situation, they can choose to initiate an attack on one of the seven strategic locations by choosing a military card. This triggers a simulated battle influenced by the scores and strategic conditions like the strength of the troops at the corresponding location.

To achieve overall victory, the defender must maintain specific conditions by the end of round 10 but also throughout the game. Therefore, the player has to ensure that the communication and morale score remains above thirty percent, that the resource score remains above zero and at least four of the seven strategic locations are in the defender's control. Failure to meet these conditions or allowing the escalation score to reach level ten results in a mutual loss, signifying an inability to contain the conflict and prevent escalation.

2.2 Implementation

The implementation of our demonstrator involved several components, leveraging the REST architecture to ensure modularity and scalability. For the backend we utilized Python and the Django framework providing a robust and flexible platform for managing game logic and data. The frontend was developed using plain CSS, JavaScript, and HTML.

Regarding the LLM, we accessed the NATO Software Factory's infrastructure, which provided access to ChatGPT via the Azure OpenAI Service. We experimented with both ChatGPT 3.5 Turbo and ChatGPT 4.0, ultimately selecting ChatGPT 3.5 Turbo due to its consistent performance and fewer restrictions on military-related queries. Later, models on local hardware were also implemented and used (Mistral 7b, Llama 7b, Hermes 7b, Falcon 7b). However, the findings described here refer, unless noted otherwise, to the use of ChatGPT 3.5 Turbo, the most powerful and capable model we utilized. Additionally, there is the possibility to reference larger custom datasets through the embedding of this data, or to gather specific information from the internet using a web crawler or by providing documents in PDF format, all of which can be parsed and utilized by the LLM. In summary, we will use the developed minimal viable product to gain direct experience with the combination of LLM and wargames. Furthermore, this will enable us to derive insights and develop further considerations for future applications. An overview of the architecture and of the LLM applications within the wargame can be found in Figure 2-1.

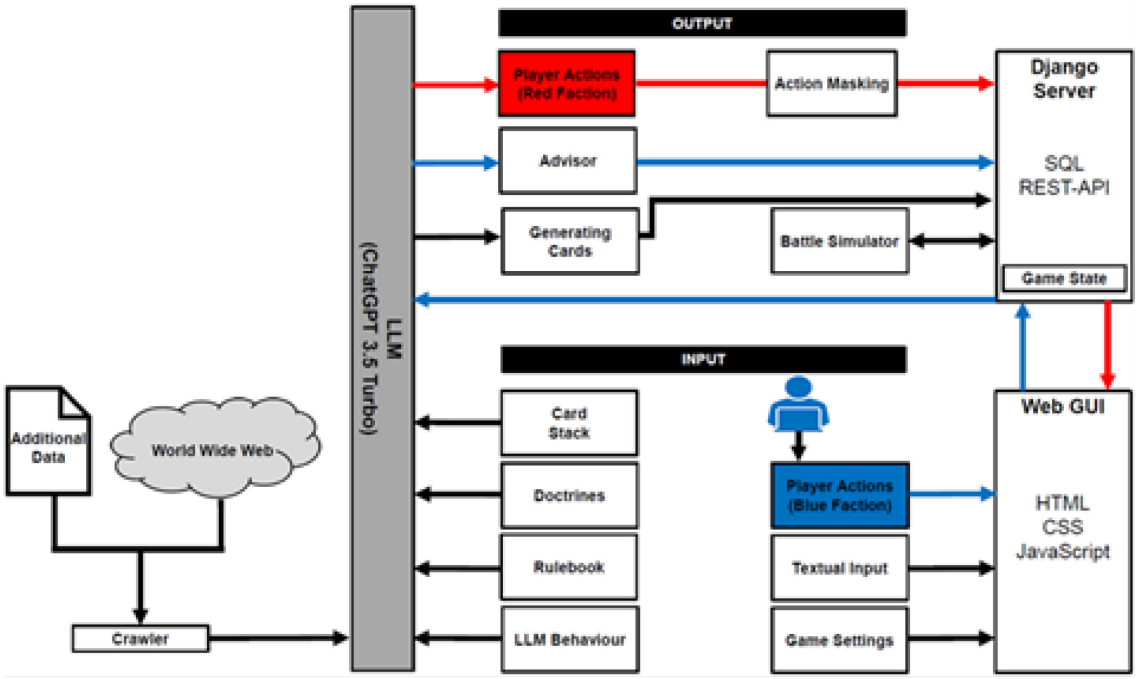

Figure 2-1: Visualization of the architecture and LLM-Use-Cases within the wargame.

2.3 Related Literature

To position our demonstrator within the existing body of knowledge, we conducted a review of related literature. Our objective was to compare our approach with established methodologies and pinpoint the distinctive contributions of our demonstrator.

The potential applications of artificial intelligence (AI) in wargaming are numerous [23]. LLMs represent a significant advancement in the context of wargaming, marking a pivotal shift when contrasted with earlier approaches. However, persistent concerns regarding the deployment of AI in wargames include issues of explainability and the necessity for cautious implementation. [23][24]

In practical terms, AI's role in wargaming spans from generating scenarios and providing decision support to players, to representing opposing forces [25]. Prior efforts have utilized AI virtual assistants for establishing legal and regulatory frameworks [26], and for simulating the actions of participating nations [27]. These efforts predominantly rely on LLM agents. The latter approach represents a multi-agent concept which foregoes human interaction and solely relies on human interaction. It is argued that multi-agent LLM wargames could redefine conflict resolution strategies and contribute insights into human history to mitigate future conflicts [27]. Additionally, proponents like [28] extol the transformative potential of LLMs in wargaming, emphasizing their ability to consistently deliver superior performance across a spectrum of data inputs. [8] and [29] state that in LLM human behaviour is already implicitly encoded due to the data they are trained on. Following, there is potential to use (multi-agent) LLM for the simulation of behaviour that mimics human abilities in strategy reasoning in order to function as a decision support tool and to be used as a kind of policy exploration [30][31][32].

In contrast to these highly optimistic assessments, [33] advocate for integrating LLMs into the decision step of the OODA loop (observe-orient-decide-act), albeit with caution against overestimating their capabilities. Rather, they recommend deploying LLMs only as supportive tools under human oversight. [34] emphasizes the significant differences of the working wise of LLM compared to human decision-making processes.

From a technical perspective, research is actively advancing both the general capabilities of LLMs [35] and their specific applications in wargaming contexts [34]. This includes efforts to optimize LLM performance and tackle communication challenges inherent in human-machine interactions [34]. Additionally, advancements in AI are enhancing player engagement by improving the behaviours of nonplayer characters represented by AI [36], highlighting the importance of effective coordination and communication, also between non-player agents [37][27]. Other works utilize AI, particularly deep learning, to address the challenge of incomplete information in wargames by introducing a set of wargame replays and demonstrating enhanced capabilities for location prediction in a specific wargame [38].

The Chinese People's Liberation Army has been working on the implementation of AI in wargames since 2017, in order to improve the quality of wargames in training, as well as a test of how AI systems can support human command and decision-making in the future [21][39]. The United States Armed Forces States are also increasingly focusing on the use of AI in wargames and decision making processes [40][41][42][43].

According to our research, there are currently no approaches explicitly combining MDO in quantitative wargames with the extensive use of LLM.

3.0 RESULTS

In this section we describe the use cases of the LLM within our wargame in an aggregated form. The results are then discussed in the following section in more detail.

3.1 Scalability

In addition to the detailed creation of game mechanics and the definition of the underlying mathematical models, the extensive and time-consuming work of game designers and developers consists in particular of the development of realistic game content [44]. Beyond the basic game structure, the success of individual wargames largely depends on the scenarios and vignettes, which are designed to immerse players in the game world. A time-consuming factor in the development of wargames is repetitive work such as the creation of game maps or various scenarios. Often, this content follows some specific patterns (e.g. game cards) to achieve a degree of similarity, reproducibility or scalability. These pre-defined structures and patterns make it possible to address the stated problem systematically by using LLM.

In our case, to enhance the scalability of the wargame during the development phase and to illustrate the use case in a basic form using our demonstrator, we utilized an LLM to generate additional game cards that represent the possible actions of the factions. Existing cards were provided to the LLM as examples and served as foundation. The LLM then created new cards, populating properties like the scenario description in free-text, the card category (e.g. military), costs (resource points), impact on the escalation score and the value of the category-specific score(s). For military cards this corresponds to the impact score, the location where the troops are supposed to be deployed and the decision whether to initiate an attack. The requirements for card creation can be specified within a prompt. For instance, we formulated requirements such that the cards thematically reference real hybrid attacks from recent years. As a result, the generated cards included scenarios such as the 2015 cyberattack on the German Bundestag, representing the theft of confidential emails and documents, thereby leading to a loss of morale points [45]. Cards generated through this method underwent an additional step to assure their quality. Prior to their integration into the game, these cards were reviewed by the game developers, and adjusted as necessary to ensure balance regarding point values and effects. In this context, balance refers to the careful adjustment of game mechanics so that no single card disproportionately affects the gameplay. This involves ensuring that point values and effects are designed to maintain a fair and engaging experience by preventing any card from being overly dominant or underperforming, thereby supporting a well-rounded

and strategic game environment. The function to generate additional game content, such as new cards, is not directly integrated into the wargame, but can be used during the set-up phase before the game begins.

3.2 Immersion

In order to offer players a realistic environment and to allow them to fully immerse themselves in the scenarios, several points must be considered when designing a game. Accordingly, the design of the game must be playable and interesting so that players overcome their initial scepticism and (unconsciously) expose themselves to the learning process. For this, it is not necessary to perfectly represent reality, which can only be partially achieved through models. As long as the player feels they are encountering a real world, immersion is achieved.

Aspects that promote immersion include comprehensible rules, realistic and detailed scenarios, or appealing graphics in the case of computer-aided wargames, especially for amateur players [39].

A feature we present to enhance immersion in the game is the depiction of asymmetric information regarding the opponent's moves. This is our attempt to depict the flow of information about potential troop deployments, or even hybrid actions and cyber-attacks in the real world, where such subversive actions of the enemy often only can be perceived implicitly without being explicitly recognizable or attributable. For example, the effects of malicious actions, such as a power outage, are typically reported in the media, while the cause of the failure often remains speculative. Therefore, instead of directly revealing the actions played by the opposing faction by overtly displaying the game cards or scores, we employ the LLM and prompt it to assume the role of a journalist in the conflict region. This journalist writes brief articles with concise headlines after each round, describing the actions of the opposing side indirectly. The articles contextualize the played actions within the broader scope of world events observed by the journalist. The corresponding article is then shown in a chat window.

To achieve this, the LLM is provided with information about the current scores of both factions and the cards played in the most recent turns, enabling contextualization and better interpretation. Additionally, we enhanced the prompt by requesting references to historical events when possible, as well as an assessment and evaluation of the current and foreseeable developments.

This approach gives the player an implicit understanding of the actions taken by the opponent, significantly enhancing the immersion.

3.3 Scarcity of Subject Matter Experts

The expertise required for designing, developing and carrying out wargames is usually by involving subject matter experts (SME). Depending on the scope and number of specialist areas involved, the necessary steps and actions within a wargame are very labour-intensive regarding time and costs. In particular, the availability of experts from the individual specialist areas is not always guaranteed, for example due to scheduling difficulties. Because of the tense political situation in the world, the availability of experts in the fields of Russian studies or sinology is currently only guaranteed to a limited extent, although these are precisely two important fields in the development of current wargames. In addition to the experts responsible for the conceptual content and the design of the game, there is always a lack of knowledgeable specialists to facilitate wargames (white cell). These facilitators are essential in complex (analytical) wargames to guide players, keep an eye on the rules, assess unforeseen game situations and to answer questions. This is a starting point for us to model game expertise both in design and operation using LLMs as expert systems. In this use case, we present two more complex approaches aimed at either replacing subject matter experts or at least supporting the work of SMEs.

3.3.1 Opposing Forces

We represent the opponent comprehensively through an LLM. It dynamically and autonomously selects actions from the previously presented categories. The selection process of the LLM is influenced by globally observable scores (escalation score) as well as by the individual scores of the opposing faction, which corresponds to the player that the LLM represents. Additionally, the actions depend on the last actions taken by the real player. Therefore, the cards played are explicitly provided to the LLM. Naturally, the LLM is also supplied with the fundamental game rules as input. Ultimately, the output returned by the LLM for each game round consists of three played cards.

This functionality is underpinned by a directive for the LLM to justify each move based on the available information. Both the move and its corresponding explanation are stored in a database, which can be particularly useful for subsequent analysis.

Furthermore, at the beginning of the game, it is possible to describe specific characteristics or desired behavioural traits in the settings (e.g. “aggressive” or “cooperative”). This allows for free-text formulation as well as for the adjustment of a parameter commonly understood as influencing creativity.

3.3.2 Advisor

Although we do not represent the blue side and therefore the defensive faction with an LLM, we still aim to present and test a use case for the application of it. Thus, the player has the option to engage in a dialogue with an advisor, represented by an LLM, via a chat window. The advisor, for instance, can offer guidance on content-related aspects and suggest subsequent actions. This functionality is supplemented by specific and previously unavailable information (see 2.2) allowing the advisor to access the NATO Operations Assessment Handbook for example, thereby providing a foundation for coherent analyses. The described function as a decision support system thus reflects the reality in which military decisions are never made alone, but only after the presentation of various possibilities by military staff. By using this approach it is possible to get a quick overview of important documents that are relevant for a specific decision in the abundant mass of available documents.

Shifting focus to the mechanics of game design, it's essential that the rules of the game are understandable for the players and that the game mechanics are clearly recognizable. This principle is sometimes only met in classic board games after supervised practice rounds. Board games always require a continuous balancing of the principle between playability and accuracy, a trade-off which computer-based games can partially overcome [46]. In addition to board and figure games, digital wargames as the third type of wargame have the advantage of being able to scale complex scenarios compared to the other two classes [44]. Such complex scenarios may lead to very extensive rulebooks easily containing up to 30 densely typed pages of rules and tables. Even experts can be put off by those rulebooks and amateurs are usually accustomed to short and simple rule sets, so the rulebooks for wargames can sometimes be overwhelming for them [44], especially if these regulations reflect the context of the MDO environment, in which complex relationships are imminent.

For this reason, complex games require a skilled facilitator to answer the player's questions and to accelerate the learning process. We have implemented a virtual facilitator using an LLM in order to make the game more efficient without a human facilitator. Therefore, we also use the LLM for the defensive factions as a methodological guidance regarding the game.

4.0 DISCUSSION

While our approach shares foundational aspects and features with previous work like [37], [26] or [34], a significant divergence lies in their focus on qualitative, open-ended wargames compared to our emphasis on

quantitative wargames with discrete action and state spaces in an MDO setting. Their methodology allows for more exploratory and narrative-driven scenarios, whereas ours is designed for more structured and measurable outcomes. This observation will be reflected in some of the aspects discussed below.

We find that generating game content using LLM is feasible, providing real added value. It is assumed that the created game cards are reviewed, ideally by a subject matter expert, before being integrated into the game. This means that the game cards undergo a detailed evaluation to assess their quality and suitability, ensuring they meet the necessary standards and align with the game's objectives before being included in the final version (validation). Therefore, they should serve as a foundation that can be appropriately expanded or modified by SMEs. The applicability of this use case also depends on the scope, level of detail and complexity of the content to be created. For example, it required significantly more iterations to generate additional maps according to our ideas than it did previously regarding the game cards. We ultimately expect a more efficient and faster creation and modification of wargames as a result. The flexible creation of new content offers the opportunity to prepare decision-makers for the changing complex and uncertain environment of conflict scenarios that they have to face within the current global dynamic [41]. Our approach is rather simplistic and demonstrates basic feasibility. An investigation into more extensive or even automated applications is still pending.

The action-masking we suggested and implemented is also a way to enhance the immersion and de-gamification of a wargame. It is a very simple yet effective means and it is conceivable that this function could optionally complement the display or the omission of specific scores. Of course, it is also possible to rely exclusively on the action masking for tracing the development of the game's situation and to disclaim displaying scores completely. The depiction of these complex environments is only possible to a limited extent with the conventional tools of classic board games with their decades-old game mechanics [41]. According to our research, this is a new method for increasing immersion in wargames.

Moreover, the LLM serves as an interactive reference for specific information or methodologies, establishing links to relevant content. While mere information provision is possible, replacing human expertise based on extensive experience regarding complex situations is not achievable at the moment. A detailed discussion on this aspect can be found in subsequent sections. However, it should be noted that the training of AI systems has so far been based on databases of colloquial language and hardly any data sets on military jargon have been incorporated which implies a linguistic restriction. In this context, the German Armed Forces are currently conducting a study on how AI, and particularly LLMs, can accelerate and improve the command process of the German land forces. The study's considerations range from simple information provision to the automatic recognition of situational pictures, ultimately leading to action recommendations and is therefore similar to the approach which was showcased within our demonstrator.

Using an LLM as a facilitator for establishing an understanding of the game mechanics also seems sensible and feasible to us. Players can ask step-by-step questions about the appropriate actions required by the game, thereby learning in a methodologically and didactically meaningful way. This proposal harmonizes with existing approaches in the field of educational science. In this context, AI-powered virtual agents are described as being able to accelerate the transfer of knowledge through self-regulated and personalized learning. [47]

The interactivity of an LLM, compared to conventional digital tutorials, allows for specific uncertainties to be explained in more detail, in different ways, or through examples. This is possible because the LLM, for example, has insight into the internal state machine of the program or game, and therefore knows the player's current in-game situation as well as actions taken or possible next steps.

Ultimately, this approach is expected to yield significant efficiency gains with regard to the required number of facilitators as well as the time needed for a game to be understood and actually played, enhancing the playability of a wargame, including its initial playability. Another advantage of using a digital advisor

powered by LLMs is that it mitigates players' potential reluctance to repeatedly ask the same questions during a game, a common issue with human facilitators. Players often hesitate to seek further clarification from facilitators, fearing it may reflect poorly on their understanding [44]. This hesitation can lead to incorrect or unintended actions, which may negatively impact the game's outcome due to a lack of proper understanding. Such information-based knowledge dissemination appears particularly valuable in the field of educational wargames. The combination of defined rules with flexible verbalisation using LLM is noteworthy.

Because we employed the LLM to autonomously represent the opposing faction, it can be argued that the demonstrator now functions more as a virtual, computer-based simulation rather than a traditional wargame. The distinction between wargaming and simulation is not entirely straightforward. The boundaries between these two categories become increasingly blurred with the integration of AI and the enhanced depiction of human-like behaviour. We acknowledge this challenge and its implications. [1][48][49][44][50]

Before we go into more detail, there is also the need to briefly differentiate what exactly is to be achieved with this functionality. When it comes to replicating human behaviour, a distinction must be made between behaviours and capabilities of AI which are capable of giving the impression of a human being and therefore only creating believable human-like behaviour [51] and actually achieving human capabilities. The former usually relates to individual use cases or tasks [31]. Machine learning does not necessarily have to be used for this. More basic methods such as symbolic AI can also be used, as it has often been the case in video games in recent years and as it can be traced in the work of [52]. The latter corresponds to the abilities attributed to so-called strong or conscious AI. There is a consensus that this is not possible yet [53][54]. However, creating believable human-like behaviour is very possible and LLM can play a significant role in this [55][8][56][57]. LLM and machine learning methods are currently being used for this purpose, with the hope of creating agents that enable a wide range of generic behaviours and interaction possibilities without being limited to specific applications, thus intending to bridge the path from believable behaviour to strong AI [55]. In the scope of our demonstrator we refer to the believable behaviour.

We found that the actions and behaviour of the LLM remained simplistic in this use case. It became evident that, for the AI to perform convincingly even to a minimal extent, an extraordinarily precise description of the task and behaviour, as well as a well-thought-out and well-defined game design, are necessary. This impression is particularly strong when we examine the LLM's reasoning for its chosen actions and the strategies developed from them. For example, in many instances, it disproportionately favored military and aggressive actions, even when the situation or prompts called for more nuanced or defensive strategies. However, it is important to note that the LLM's behavior may not be solely due to its own limitations. The design of the wargame itself might have contributed to this tendency. Certain game mechanics, scenarios, or reward structures may unintentionally influence the LLM towards favoring military actions, suggesting that a balanced game design plays a significant role in shaping the LLMs decisions.

We also suspect that this negative impression arises because our game is a quantitative wargame, featuring a discrete action and state space, which imposes strong structural constraints on possible actions and behaviours. Other AI approaches, such as reinforcement learning, are likely to be better suited to representing the opposing faction, considering the characteristics of reinforcement learning and of our setting [58][59][60]. Such approaches were recently explored by [41] and [40]. Since reinforcement learning is capable of impressive and sometimes superhuman performance, as it can be seen, for example, with AlphaStar in StarCraft II, AlphaGo in Go or AlphaZero in chess [61][62][63], some opinions expect similarly convincing results from LLMs in the future [14].

As described in a previous section, LLMs are rather employed in the context of qualitative and open-ended wargames. The corresponding works report comparatively more optimistic assessments of the LLM usage, even though not entirely without doubt. This observation underscores our belief that employing LLMs to simulate substantively valuable and qualitatively robust opponents represents a use case that cannot be

readily assured or effectively realized within the scope of our demonstrator. At the beginning we mentioned that LLMs can exhibit biases in the way they work and the responses they generate due to the underlying data on which they were trained. Following this, we seek to examine whether certain LLMs, due to their data foundations and training processes, are even capable of representing specific opponents and, by extension, different values and worldviews. Can it be stated that some LLMs are influenced by Western or Eastern perspectives, thereby affecting the quality of their opponent representations? This question remains unresolved and warrants further investigation.

Nevertheless, we would also like to shed light on conceivable positive aspects of the use of AI for the purpose of presenting an opposing side. Although player strive to thoroughly consider the opponent's thinking, intentions, and strategies in the event of a conflict, they are invariably influenced by their own social and societal biases. This fits to Thomas Schelling's theorem, in that he argues that, despite rigorous analyses and thorough deliberations, we could never draw up a list of things that would not occur to us [1]. Although there is still no AI that can be proven to digitally represent the thinking of a human or our opponent, it opens the possibility for alternative and potentially unconventional ways of thinking. AI-powered wargaming can therefore contribute to transcending the boundaries of conventional thinking.

The discussion about using AI or LLM as opponent representation leads to a demand or problem statement frequently encountered in similar works and represents a general issue with the use of LLMs: the need for an appropriate methodology for validation and verification. This is particularly problematic due to the stochastic nature of LLMs. For example, it is questionable how we can ensure that the LLM actually understood the game rules in order to be able to answer correctly to game related questions. Another example is that we became aware of issues during our application, as the LLM initially and irregularly failed to produce outputs in the required structure. Ensuring that the LLM exclusively produced the required output in terms of structure, was only possible with a correspondingly designed architecture and through appropriate error handling. More specifically, this means that the tasks assigned to the LLM must be designed to be highly detailed, unambiguous, and consequently straightforward. This has significant implications for practical and technical usability. A similar point is made in [55] concerning the creation of believable behavior. As for the content of the output, we relied on individual manual validation. Anyway, these challenges are known and are being actively addressed and resolved. For example, it is now natively possible for ChatGPT-4 to generate JSON formats.

Finally, we can also state that the extensive use of LLM on local or conventional hardware is limited in its scope and performance, even though it is basically suitable for the deployment and usage of LLM. Hardware requirements are still a limiting factor at the moment. At least when such extensive and intensive use of LLMs occurs, as is the case with us. Especially when working with classified data, proprietary solutions are indispensable or established solutions cannot be used without further ado. Although we did not explicitly investigate the behaviour and performance of the various locally implemented models, we were able to make some observations during development. In addition to the known differences in computation times, with Mistral being the fastest local model, we noticed, for example, that some models often generated content that could not be used directly. This was due to some models responding in a conversational format rather than generating solely the requested article or card when instructed to do so. It is also quite apparent that using LLMs in many use cases significantly slows down the game speed and extends the duration of a round due to the time required for generating responses. Nevertheless, LLMs enable rapid development and prototyping of AI containing applications compared to implementing rule-based approaches or other commonly used AI methodologies.

5.0 CONCLUSION

There is a pressing need to initiate small-scale, productive use cases that make effective and efficient use of LLMs in the boundaries of well-defined tasks. Rather than solely focusing on extensive applications, starting

with bounded and specific implementations can pave the way for meaningful advancements in utilizing LLMs responsibly and effectively across various fields. This approach not only mitigates the risks associated with broader deployments but also establishes a foundation for refining and scaling up these technologies in a sustainable manner. LLMs should and can take on supportive roles, but they should not represent complex, game-determining functionalities.

Our contribution to the discourse lies in identifying previously overlooked use cases and in applying existent use cases within quantitative MDO wargames. Amongst other things, we identify straightforward and reliable approaches that prioritize productivity over complexity, aligning with practical implementation needs. The exploration into integrating LLMs into wargaming for MDO reveals significant potential but also challenges. LLMs offer clear benefits in enhancing immersion and enabling features like action masking, which improve the overall gameplay experience, operational realism and player engagement. A further advantage of LLMs lies in their ability to automate content generation, thereby broadening the scope and scalability of wargaming simulations.

Despite these advancements, the integration of LLMs as AI-driven adversaries requires careful consideration of their accuracy and quality and it is limited by its simplicity.

Compared to other existing approaches, we evaluate the use of LLMs quite conservatively and cautiously. This is partly due to the fact that we are examining a quantitative wargame within the MDO context, which imposes greater constraints on the use of LLMs, preventing the full realization of the LLM's inherent strengths.

Frequently cited comparisons between AI capabilities that surpass human abilities and the potential or actual capabilities of LLMs in strategic development and problem-solving for military, political, or strategic issues using wargaming tools, in our opinion, lack foundation and are exaggerated.

In our demonstrator, we were only able to roughly outline and delineate the individual use cases. Consequently, there remains significant potential to investigate the individual examples in greater detail to gather further insights.

We also observed that the use of LLMs in educational wargames has not yet been extensively explored. We see further potential, particularly concerning the advisor function and regarding a better learning process. It is also of interest to us to determine to what extent a LLM is capable of generating more complex game content and scenarios that can be considered balanced and well-designed, without requiring substantial modifications by a subject matter expert. Further, in our demonstrator we did not consider the use of AI and LLM in the analysis and evaluation of analytical wargames [4]. Additionally, there should be a more intensive focus in the future on integrating various types of documents and improving the architecture.

It is also conceivable that the behaviour of various LLM models within our demonstrator is examined more closely and compared with each other as it was already done in [34] in a different context.

6.0 REFERENCES

[1] C. Turnitsa, C. Blais, and A. Tolk, Simulation and Wargaming. John Wiley & Sons, 2022. [Online]. Available: https://play.google.com/store/books/details?id=qk9WEAAAQBAJ

[2] “Wargaming Handbook.” Development, Concepts and Doctrine Centre, Ministry of Defence Shrivenham, Aug. 2017. [Online]. Available: https://www.gov.uk/government/publications/defence-wargaming-handbook

[3] J. Wintjes, “A school for war – A brief history of the Prussiankriegsspiel,” Simulation and Wargaming. Wiley, pp. 23–64, Dec. 22, 2021. doi: 10.1002/9781119604815.ch2.

[4] J. Wintjes and S. Pielström, Pluie de Balles: A “gamified” Kriegsspiel. BoD – Books on Demand, 2023. [Online]. Available: https://play.google.com/store/books/details?id=0fPqEAAAQBAJ

[5] “NATO Wargaming Handbook.” Nato Allied Command Transformation, Sep. 2023. [Online]. Available: https://paxsims.files.wordpress.com/2023/09/nato-wargaming-handbook-202309.pdf

[6] OpenAI et al., “GPT-4 Technical Report,” arXiv [cs.CL], Mar. 15, 2023. [Online]. Available: http://arxiv.org/abs/2303.08774

[7] L. Floridi and M. Chiriatti, “GPT-3: Its nature, scope, limits, and consequences,” Minds Mach. (Dordr.), vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[8] T. B. Brown et al., “Language Models are Few-Shot Learners,” arXiv [cs.CL], May 28, 2020. [Online]. Available: http://arxiv.org/abs/2005.14165

[9] W. Liang et al., “Mapping the increasing use of LLMs in scientific papers,” arXiv [cs.CL], Apr. 01, 2024. [Online]. Available: http://arxiv.org/abs/2404.01268

[10] A. B. Rashid, A. K. Kausik, A. Al Hassan Sunny, and M. H. Bappy, “Artificial intelligence in the military: An overview of the capabilities, applications, and challenges,” Int. J. Intell. Syst., vol. 2023, pp. 1–31, Nov. 2023, doi: 10.1155/2023/8676366.

[11] Bundeswehr (Army Concepts and Capabilities Development Centre), “Artificial Intelligence in Land Forces - A position paper developed by the German Army Concepts and Capabilities Development Centre.” 2019. [Online]. Available: https://www.bundeswehr.de/resource/blob/156026/79046a24322feb96b2d8cce168315249/download-positionspapier-englische-version-data.pdf

[12] I. O. Gallegos et al., “Bias and fairness in large language models: A survey,” arXiv [cs.CL], Sep. 01, 2023. [Online]. Available: http://arxiv.org/abs/2309.00770

[13] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, “Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions,” in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM, Apr. 2023. doi: 10.1145/3544548.3581318.

[14] R. Bommasani et al., “On the Opportunities and Risks of Foundation Models,” arXiv [cs.LG], Aug. 16, 2021. [Online]. Available: http://arxiv.org/abs/2108.07258

[15] B. Yan et al., “On protecting the data privacy of large language models (LLMs): A survey,” arXiv [cs.CR], Mar. 08, 2024. [Online]. Available: http://arxiv.org/abs/2403.05156

[16] A. S. Luccioni, Y. Jernite, and E. Strubell, “Power hungry processing: Watts driving the cost of AI deployment?,” arXiv [cs.LG], Nov. 28, 2023. [Online]. Available: http://arxiv.org/abs/2311.16863

[17] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, “Recalibrating global data center energy-use estimates,” Science, vol. 367, no. 6481, pp. 984–986, Feb. 2020, doi: 10.1126/science.aba3758.

[18] J. Stojkovic, E. Choukse, C. Zhang, I. Goiri, and J. Torrellas, "Towards greener LLMs: Bringing energy-efficiency to the forefront of LLM inference," arXiv [cs.AI], Mar. 29, 2024. [Online]. Available: http://arxiv.org/abs/2403.20306

[19] U. Gupta et al., "Chasing carbon: The elusive environmental footprint of computing," in 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA), IEEE, Feb. 2021. doi: 10.1109/hpca51647.2021.00076.

[20] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots," in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA: ACM, Mar. 2021. doi: 10.1145/3442188.3445922.

[21] D. Cheng, "The People's Liberation Army on Wargaming," War on the Rocks. Accessed: Jul. 28, 2024. [Online]. Available: https://warontherocks.com/2015/02/the-peoples-liberation-army-on-wargaming/

[22] "Wargaming Handbuch der Bundeswehr." Doktrinzentrum der Bundeswehr, 2024. [Online]. Available: https://paxsims.wordpress.com/wp-content/uploads/2024/05/1716363823022.pdf

[23] J. Goodman, S. Risi, and S. Lucas, "AI and Wargaming," arXiv [cs.AI], Sep. 18, 2020. [Online]. Available: http://arxiv.org/abs/2009.08922

[24] P. K. Davis and P. Bracken, "Artificial intelligence for wargaming and modeling," J. Def. Model. Simul. Appl. Methodol. Technol., p. 154851292110731, Feb. 2022, doi: 10.1177/15485129211073126.

[25] J. Keller, "DARPA SCEPTER project seeks to develop battle planning for complex military engagements at machine speed," Mil. Aerosp. Electron., Jan. 2022, Accessed: Jul. 09, 2024. [Online]. Available: https://idstch.com/technology/ict/darpa-scepter/

[26] A. Knack and R. Powell, "Artificial Intelligence in Wargaming: An evidence-based assessment of AI applications," Jun. 2023, [Online]. Available: https://cetas.turing.ac.uk/publications/artificial-intelligence-wargaming

[27] W. Hua et al., "War and Peace (WarAgent): Large Language Model-based Multi-Agent Simulation of World Wars," arXiv [cs.AI], Nov. 28, 2023. [Online]. Available: http://arxiv.org/abs/2311.17227

[28] Y. Chen and S. Chu, "Large Language Models in Wargaming: Methodology Application and Robustness," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Jun. 2024, Accessed: Jul. 11, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024W/AdvML/html/Chen_Large_Language_Models_in_Wargaming_Methodology_Application_and_Robustness_CVPRW_2024_paper.html

[29] R. E. Guingrich and M. S. A. Graziano, "Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction," Front. Psychol., vol. 15, p. 1322781, Mar. 2024, doi: 10.3389/fpsyg.2024.1322781.

[30] K. Sreedhar and L. Chilton, "Simulating Human Strategic Behavior: Comparing Single and Multi-agent LLMs," arXiv [cs.HC], Feb. 13, 2024. [Online]. Available: http://arxiv.org/abs/2402.08189

[31] M. Barthet, A. Khalifa, A. Liapis, and G. N. Yannakakis, "Generative Personas That Behave and Experience Like Humans," arXiv [cs.AI], Aug. 26, 2022. [Online]. Available: http://arxiv.org/abs/2209.00459

[32] A. Fuchs, A. Passarella, and M. Conti, "Modeling Human Behavior Part I -- Learning and Belief Approaches," arXiv [cs.AI], May 13, 2022. [Online]. Available: http://arxiv.org/abs/2205.06485

[33] A. Chochtoulas, "How Large Language Models are Transforming Modern Warfare," Joint Air Power Competence Centre, May 2024, [Online]. Available: https://www.japcc.org/articles/how-large-language-models-are-transforming-modern-warfare/

[34] M. Lamparth, A. Corso, J. Ganz, O. S. Mastro, J. Schneider, and H. Trinkunas, "Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations," arXiv [cs.CY]. 2024. [Online]. Available: http://arxiv.org/abs/2403.03407

[35] H. Naveed et al., "A Comprehensive Overview of Large Language Models," arXiv [cs.CL]. 2024. [Online]. Available: http://arxiv.org/abs/2307.06435

[36] H. Jung, "A Glimpse of the Future Battlefield: AI-Embedded Wargames," Proc. AMIA Annu. Fall Symp., vol. 150, no. 6, Jun. 2024, [Online]. Available: https://www.usni.org/magazines/proceedings/2024/june/glimpse-future-battlefield-ai-embedded-wargames

[37] D. P. Hogan and A. Brennen, "Open-Ended Wargames with Large Language Models," arXiv [cs.CL], Apr. 17, 2024. [Online]. Available: http://arxiv.org/abs/2404.11446

[38] M. Liu et al., "Introduction of a new dataset and method for location predicting based on deep learning in wargame," J. Intell. Fuzzy Syst., vol. 40, no. 5, pp. 9259–9275, 2021, doi: 10.3233/JIFS-201726.

[39] E. Kania and I. B. Mccaslin, "Learning Warfare from the Laboratory - China's Progression in Wargaming and Opposing Force Training," Military Learning and the Future of War, Sep. 2021, [Online]. Available: https://www.understandingwar.org/report/learning-warfare-laboratory-china%E2%80%99s-progression-wargaming-and-opposing-force-training

[40] R. S. Badalyan, A. D. Graham, M. W. Nixt, and J.-E. Sanchez, "APPLICATION OF AN ARTIFICIAL INTELLIGENCE-ENABLED REAL-TIME WARGAMING SYSTEM FOR NAVAL TACTICAL OPERATIONS." [Online]. Available: https://apps.dtic.mil/sti/trecms/pdf/AD1184745.pdf

[41] S. Black and C. Darken, "Scaling artificial intelligence for digital wargaming in support of decision-making," arXiv [cs.LG], Feb. 08, 2024. [Online]. Available: https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-MSG-207/MP-MSG-207-23.pdf

[42] M. Farmer, "Four Dimensional Planning at the Speed of Relevance - Artificial Intelligence Enabled Military Decision Making Process," MILITARY REVIEW - THE PROFESSIONAL JOURNAL OF THE U.S. ARMY, November-Dicember 2022, Vol. 102, No. 6, Nov. 2022. [Online]. Available: https://www.armyupress.army.mil/Portals/7/military-review/Archives/English/ND-22/Farmer/Farmer-Clausewitz%E2%80%99s-Ghost-UA.pdf

[43] D. C. Tarraf et al., An Experiment in Tactical Wargaming with Platforms Enabled by Artificial Intelligence. Santa Monica, CA: RAND Corporation, 2020. doi: 10.7249/RRA423-1.

[44] P. Sabin, Simulating War: Studying Conflict through Simulation Games. Bloomsbury Academic, 2014. [Online]. Available: https://play.google.com/store/books/details?id=x8jYngEACAAJ

[45] Bundesamt für Sicherheit in der Informationstechnik (BSI), "Die Lage der IT-Sicherheit in Deutschland 2015." Nov. 2015. [Online]. Available: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2015.pdf%3F__blob%3DpublicationFile

[46] M. Caffrey (Jr.), On wargaming: How wargames have shaped history and how they may shape the future. 2019.

[47] C.-P. Dai and F. Ke, "Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review," Computers and Education: Artificial Intelligence, vol. 3, p. 100087, Jan. 2022, doi: 10.1016/j.caeai.2022.100087.

[48] J. Appleget, R. Burks, and F. Cameron, The craft of wargaming the craft of wargaming: A detailed planning guide for defense planners and analysts. US Naval Institute Press, 2020. [Online]. Available: https://books.google.com/books/about/The_Craft_of_Wargaming.html?id=DTr1DwAAQBAJ

[49] A. M. Law, Simulation Modeling and Analysis. McGraw-Hill Education, 2014.

[50] J. Hodicky and J. Melichar, "Role and Place of Modelling and Simulation in Wargaming." NATO S&T Organisation, 2017. [Online]. Available: https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-MSG-149/MP-MSG-149-12.pdf

[51] I. Umarov and M. Mozgovoy, "Believable and effective AI agents in virtual worlds," Int. J. Gaming Comput. Mediat. Simul., vol. 4, no. 2, pp. 37–59, Apr. 2012, doi: 10.4018/jgcms.2012040103.

[52] M. Colledanchise and P. Ögren, "Behavior Trees in Robotics and AI: An Introduction," arXiv [cs.RO], Aug. 31, 2017. [Online]. Available: http://arxiv.org/abs/1709.00084

[53] G. W. Ng and W. C. Leung, "Strong Artificial Intelligence and Consciousness," J. AI. Consci., vol. 07, no. 01, pp. 63–72, Mar. 2020, doi: 10.1142/S2705078520300042.

[54] M. V. Butz, "Towards Strong AI," KI - Künstliche Intelligenz, vol. 35, no. 1, pp. 91–101, Mar. 2021, doi: 10.1007/s13218-021-00705-x.

[55] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," arXiv [cs.HC], Apr. 07, 2023. [Online]. Available: http://arxiv.org/abs/2304.03442

[56] G. Jäger and D. Reisinger, "Can We Replicate Real Human Behaviour Using Artificial Neural Networks?," arXiv [cs.MA], Jul. 09, 2021. [Online]. Available: http://arxiv.org/abs/2107.04267

[57] A. F. Mendi, F. N. Büyükoflaz, T. Erol, and E. Kalfaoglu, "Transition from Rule-based Behaviour Models to Learning-based Behaviour Models in Tactical Simulation," in NATO STO IST-190 Research Symposium (RSY) on Artificial Intelligence, Machine Learning and Big Data for Hybrid Military Operations (AI4HMO), unknown, Oct. 2021. doi: 10.14339/STO-MP-IST-190-10-PDF.

[58] H. C. Siu et al., "Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi," arXiv [cs.AI], Jul. 15, 2021. [Online]. Available: http://arxiv.org/abs/2107.07630

[59] R. S. Sutton and A. G. Barto, Reinforcement Learning, second edition: An Introduction. London, England: MIT Press, 2018.

[60] S. Miyashita, X. Lian, X. Zeng, T. Matsubara, and K. Uehara, "Developing game AI agent behaving like human by mixing reinforcement learning and supervised learning," in 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), IEEE, Jun. 2017, pp. 489–494. doi: 10.1109/SNPD.2017.8022767.

[61] D. Silver et al., "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," arXiv [cs.AI], Dec. 05, 2017. [Online]. Available: http://arxiv.org/abs/1712.01815

[62] O. Vinyals et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," Nature, vol. 575, no. 7782, pp. 350–354, Nov. 2019, doi: 10.1038/s41586-019-1724-z.

[63] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.