Large Language Models in Wargaming: Methodology, Application, and Robustness

Yuwei Chen, Shiyong Chu
Aviation Industry Development Research Center of China
No.14 Xiao Guan Dong Li, Chaoyang District, Beijing, China
catcornic@gmail.com, csy3191dl@163.com

Abstract

Traditional artificial intelligence (AI) has contributed strategic enhancements to wargaming but often encounters difficulties in dynamically complex environments and in adapting to unforeseen developments. In contrast, Large Language Models (LLMs) offer advanced natural language processing, analytical capabilities, and intuitive decision-making communication. LLMs excel in rapidly analyzing voluminous textual data, identifying patterns, and generating insights for strategic planning, thereby addressing the critical demand for anticipatory strategy and creative solution development in wargaming. Nonetheless, deploying LLMs in this context introduces potential robustness challenges, particularly their vulnerability to adversarial prompts. Our experimental investigations reveal LLMs' susceptibility to misleading or hostile inputs, underscoring the imperative for implementing robustness measures to safeguard their operational integrity and reliability in strategic applications. Our pioneering research, through targeted experiments within a commercial wargaming, demonstrates the feasibility and potential of LLMs to significantly improve outcomes in representative scenarios. This work not only evidences the significant impact of LLMs on the decision-making landscape in wargaming but also establishes a foundation for future research and the practical implementation of LLMs in advanced decision support systems.

1. Introduction

As AI continues to evolve, its capacity to augment human endeavors has deepened, spanning from object recognition and complex logic processing to decision generation and strategy formulation. In pivotal sectors such as healthcare[1], logistics[2], finance[3], retail[4], and manufacturing[5], AI's role is increasingly integral, enhancing decision-making by leveraging vast data analyses to identify patterns, predict outcomes, and automate routine operations. This technological advancement propels efficiency and accuracy, enabling rapid, data-informed decisions essential in dynamic contexts. Critically, AI diminishes human error and bias, thereby ensuring outcomes that are both reliable and objective. Its predictive analytics afford organizations the foresight needed to navigate future challenges, offering a distinct strategic edge. AI's automation capabilities also liberate human resources for more nuanced tasks and strategic engagement. Additionally, its proficiency in risk assessment fortifies organizational resilience. Through innovation, AI catalyzes the development of groundbreaking solutions, pushing the boundaries of current industry standards. Ultimately, the integration of AI into decision-making processes not only streamlines operations but also cultivates a culture that is innovative, efficient, and fundamentally driven by data[6].

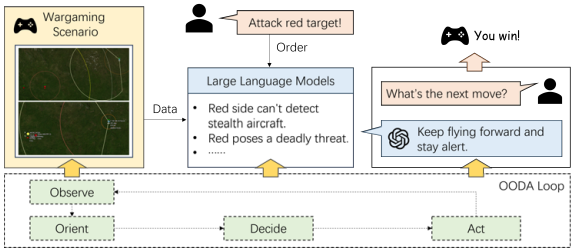

Figure 1. LLMs gather the objectives of scenarios outlined by human players, as well as the information emanating from these scenarios. Utilizing their extensive and broad knowledge base, LLMs perform in-depth analyses to furnish players with more logical and explainable recommendations for actions within wargaming contexts.

Although AI markedly decreases the duration of decision-making cycles, thereby improving efficiency, its adoption faces substantial barriers. These include constrained interpretability[7][8], an overreliance on empirical data[9][10], difficulties in addressing highly complex scenarios[11][12], insufficient exploratory analy-

sis capabilities[13][14], and vulnerability to adversarial attacks[15][16]. Such limitations compromise the utility of AI in facilitating decision-making processes, frequently eliciting skepticism about its practical applicability in societal contexts[17]. As a result, the broad adoption of AI-assisted decision-making encounters significant obstacles[18].

The emergence of LLMs mark a significant evolution in AI-assisted decision-making, providing a suite of advantages that mitigate the limitations of broader AI systems[19]. LLMs excel in natural language understanding and generation, facilitating intuitive interactions and making complex decision communication accessible to a broader audience[20]. Their adaptability, stemming from extensive training on diverse subjects, allows for application across varied contexts without domain-specific programming, making them invaluable for multidisciplinary decision-making.

Despite potential biases from their training data, LLMs' capacity for continuous learning and adaptation suggests a potential for reducing these biases over time through exposure to balanced data and feedback. Their scalability and efficiency in processing and generating responses swiftly address tasks that would otherwise overwhelm human capacities, especially in synthesizing information from diverse sources to inform decisions. In essence, LLMs enhance decision-making processes by providing intuitive, versatile, and comprehensive support, leveraging their advanced natural language processing and analytical strengths to augment human capabilities in a wide array of decision-making contexts[21].

In the realm of wargaming, the necessity for swift and precise decision-making is crucial, ensuring tasks are carried out with both efficiency and logical coherence. The intrinsic complexity and the unpredictability inherent in this field demand sophisticated decision-making tools adept at navigating these multifaceted challenges. Although AI has been widely implemented to augment decision-making capabilities in wargaming[22], its efficacy is frequently impeded by its limitations in addressing unexpected occurrences and a deficiency in comprehending complex scenarios thoroughly. In comparison, LLMs demonstrate significant promise in surmounting these hurdles. Their advanced natural language processing abilities and proficiency in analyzing intricate situations facilitate a more holistic approach to decision-making within wargaming. Incorporating LLMs can markedly enhance the performance of entities within simulations, ensuring that decisions are both prompt and well-informed, thus preserving the essential efficiency and logical integrity necessary for successful wargaming outcomes.

Nonetheless, the direct application of LLMs in wargaming encounters considerable challenges. Notably, the majority of prevailing LLMs lack the functionality for real-time learning, rendering them incapable of integrating new information or adapting to changes that emerge after their initial training phase[23]. Additionally, LLMs are susceptible to data biases[24], potentially resulting in suboptimal decisions within the wargaming decision-making processes. This limitation bears significant implications, as it can compromise the reliability and efficacy of LLMs in dynamic and evolving wargaming scenarios, thereby underscoring the importance of ongoing updates and meticulous evaluation of the data underpinning LLM training to ensure equitable and accurate decision support.

Moreover, recent studies highlight a critical vulnerability of LLMs to strategically crafted inputs, such as adversarial prompts[25][26], which can undermine their effectiveness. This vulnerability raises serious concerns for their direct implementation in wargaming contexts, where inaccuracies like typographical errors, the use of synonyms, and semantic variations in user inputs can lead LLMs to misinterpret the intended message or content. Such misinterpretations can produce erroneous or nonsensical outputs, significantly impeding the decision-making process. These insights accentuate the pressing necessity to bolster the robustness of LLMs against such manipulations, ensuring their viability in the strategic deliberations integral to wargaming.

To the best of our knowledge, we are the inaugural team to implement LLMs in the domain of wargaming, with our specific contributions outlined as follows:

- We introduce a novel approach for utilizing LLMs to support decision-making processes in wargaming.
- We integrate the concept of the OODA (Observe, Orient, Decide, Act) loop, a strategic framework derived from wargaming, into the application of LLMs.
- We selected a representative scenario to conduct decision support experiments using LLMs, evaluating their robustness against adversarial prompts.

2. Related Work

2.1. AI in Wargaming

The integration of AI into wargaming represents a transformative leap, offering sophisticated simulation of military scenarios[22][27], augmenting strategic decision-making with autonomous technologies[28], and refining operational tactics[29][30][27]. AI plays a pivotal role in facilitating nuanced decision-making[31], delivering comprehensive analyses of conflicts[32], and yielding strategic insights into adversaries' intentions and possible outcomes[33]. Advanced technologies such as Hierarchical Reinforcement Learning[34] and deep reinforcement learning expand AI's ability to navigate and engage with complex decision environments, substantially enhancing strategic planning and implementation. However, a significant impediment to

these advancements is the issue of explainability[35][22]. The inherent complexity of AI algorithms[33] and the consequent opacity of their decision-making processes elicit ethical concerns and erode user trust, thus impeding the validation of AI-informed strategies and their ethical application[36][30][27]. Addressing this opacity is crucial; enhancing AI's explainability is imperative to ensure its contributions to wargaming are transparent, ethically sound, and in alignment with established standards. Such efforts are essential for fostering accountability and optimizing the integration of AI into strategic and tactical military simulations, paving the way for more responsible and effective use of AI in complex decision-making contexts.

2.2. LLMs in Decision-Making

Leveraging the transformative potential of AI in wargaming, the incorporation of LLMs into decision-making processes signifies a substantial progression, enhancing strategic insight and operational accuracy within simulations. Recent advancements position LLMs as superior to traditional AI methodologies in decision-making capabilities, distinguished by their proficiency in interpreting complex, nuanced data and generalizing across atypical scenarios[20][37][20][38], thereby closely mirroring human cognitive processes[39]. When integrated into various fields, LLMs exhibit exceptional capabilities in tasks that demand deep comprehension and predictive analytics. Their contribution to improving perception[40][41][42], action[43][44], and natural language interaction[39][45] , especially in advanced applications such as autonomous vehicles and robotics[46][47][43][48][49][50], highlights their unique benefits. Moreover, the role of LLMs in distilling knowledge and developing domain-specific ontologies showcases their unparalleled efficiency in synthesizing and applying comprehensive information. This represents a marked departure from the limitations inherent in traditional AI methodologies, which struggle with ambiguity, rarity, and complex linguistic constructs, positioning LLMs as pivotal in the evolution of AI-driven decision-making frameworks.

3. Methodology

In the context of wargaming, participants are typically required to delineate combat objectives aligned with the victory conditions specific to the scenario. Subsequently, they must strategically allocate resources based on their availability, adapting to the dynamic shifts in the game environment in a timely manner. The proposition of leveraging LLMs to augment decision-making processes in wargaming stems from the LLMs' profound capability to comprehend and analyze complex, evolving environments. Envisioning LLMs as the epicenter of command decision-making effectively encapsulates the "Decide" and "Act" phases of the OODA loop. Here, a player's initiation of a rudimentary attack command prompts the LLMs to gather pertinent resource data and real-time battlefield intelligence, representing the "Observe" and "Orient" phases. Utilizing its expansive knowledge base, the LLMs then conduct an analysis, formulates judgments, and, in the "Decide" phase, determines the most viable course of action. This is swiftly followed by the "Act" phase, where the LLMs communicate subsequent strategies to the commander, elucidating the rationale and analytical foundation behind each directive. This integration of LLMs into the OODA loop aims to enhance strategic planning and resource management, thereby refining the decision-making acumen within wargaming simulations, ensuring a seamless and efficient progression through each phase of the loop.

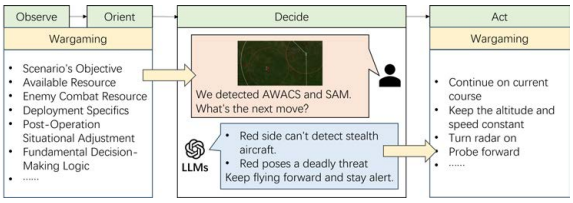

Figure 2. Incorporating LLMs into the OODA loop of wargaming enhances decision-making processes significantly. Information and situational data gathered during the "Observe-Orient" phases are fed into LLMs as textual inputs. Subsequently, these models undertake the "Decide" phase by leveraging their extensive knowledge base for thorough analysis. This approach enables the generation of action-oriented recommendations for the "Act" phase.

3.1. Observe and Orient: Data Inputs for LLMs

LLMs possess the capability to conduct thorough analyses of intricate scenarios, contingent upon the acquisition of ample input data. For LLMs to function effectively as decision-making cores in wargaming, they must be privy to the entirety of information accessible to human participants, which is derived from or represents the "Observe" and "Orient" phases of the OODA loop. This encompasses details such as the scenario's objectives, available and enemy combat resources, deployment specifics for each resource type, post-operation situational adjustments, and the fundamental decision-making logic employed in wargaming.

For instance, the scenario's goal might be achieving a particular end state as defined by the player or reaching a pre-established checkpoint. In cases where the goal is player-defined, a precise linguistic articulation of the desired outcome is essential to guide the LLMs' decision-making process. Alternatively, for checkpoint-based objectives, the LLMs require an in-depth understanding of the scoring system, including points allocation for the elimination of enemy units and penalties for the loss of friendly

units.

Moreover, it is crucial for LLMs to understand the available combat resources within a scenario to strategize their effective use in alignment with the scenario's goals. This includes identifying enemy combat resources to tailor strategies accordingly. Importantly, information on enemy resources should be obtained in a manner that mirrors the intelligence available to human players, preventing any undue advantage.

Additionally, LLMs must be apprised of the deployment details and initial states of these resources. Based on this information, LLMs should verify the availability of combat resources and strategize their deployment to achieve the scenario's objectives. Notably, intelligence on enemy deployments should be garnered through standard reconnaissance activities, as enemy, akin to real-world combat scenarios, typically do not disclose their deployment information voluntarily.

In wargaming, the acquisition of real-time situational data and tracking the alterations following each action are imperative for LLMs. This enables the understanding of the state changes each decision introduces to the overall scenario, facilitating the formulation of informed subsequent decisions based on these updated conditions. This cyclic process promotes a dynamic and informed decision-making framework, which is responsive to changing circumstances and aligned with goals.

Furthermore, LLMs should be acquainted with the fundamental operational logic of wargaming, including the granularity of actionable steps, adherence to logical consistency, and other factors influencing decision-making efficacy.

3.2. Decide: LLMs at the Crossroads

Upon receiving comprehensive data, LLMs demonstrate exceptional prowess in synthesizing diverse knowledge and information to fulfill objectives tailored to specific wargaming scenarios, leveraging the full spectrum of available combat resources and capabilities. These advanced models harness fundamental principles of strategy and capability assessment to meticulously analyze resources, thus identifying the combat potential at their disposal. By conforming to established physical and logical frameworks, LLMs are equipped to evaluate the most effective deployment of these capabilities strategically.

Crucially, LLMs integrate insights into enemy tactics, granting them the capacity to forecast enemy movements and resource allocation with remarkable precision. This foresight is cultivated through a rigorous analysis of potential strategic outcomes, underpinned by an extensive compilation of historical data, theoretical models, and contemporary strategic doctrines. Such comprehensive analytical processes empower LLMs to generate recommendations that are not only responsive to the dynamic nature of the wargaming environment but are also customized to align with the player's strategic preferences and objectives.

In the pivotal "Decide" phase, LLMs translate this exhaustive situational analysis into actionable strategic advice, guiding players towards making informed, strategically coherent decisions. This phase is characterized by the LLMs' ability to distill complex analyses into clear, actionable recommendations that consider both the immediate tactical situation and the broader wargaming objectives of the player. These recommendations are dynamically refined as the game progresses, adapting to new information and shifts in the scenario to ensure continued relevance and effectiveness.

By actively engaging in this enhanced decision-making process, LLMs significantly contribute to the strategic depth and sophistication of wargaming experiences. Players benefit from a decision-making process enriched with predictive accuracy and nuance, thereby elevating the overall efficacy of their planning and execution within the game. Through the adept application of analysis and recommendations, LLMs underscore their pivotal role in augmenting human decision-making in complex, competitive environments.

3.3. Act: Recommendations from LLMs

Leveraging insights from the "Observe" and "Orient" phases, coupled with comprehensive reasoning analysis in the "Decide" phase, LLMs proffer strategic action recommendations tailored to the nuanced mechanics of wargaming and the player's articulated strategic objectives. These recommendations, far from being static, are dynamically adjusted in response to real-time alterations within the game scenario.

This dynamic recommendation process is underpinned by LLMs' sophisticated analysis of both the broad strategic landscape and specific player preferences. By integrating a deep understanding of the game's mechanics with a player's strategic proclivities, LLMs craft bespoke advice aimed at optimizing in-game outcomes. These strategies are formulated through a meticulous evaluation of potential moves, weighing their probabilities of success and aligning them with the player's overarching goals.

Furthermore, the adaptability of these recommendations is a testament to the LLMs' advanced analytical capabilities. As the game environment evolves, the LLMs recalibrate their advice based on the latest situational data, ensuring that strategic guidance remains relevant and effective. This real-time adaptability not only enhances the decision-making process but also supports players in navigating the complexities of the game with informed confidence.

Moreover, the process of generating these recommendations involves a continuous loop of feedback and re-

finement. Players' responses to suggested strategies allow LLMs to further hone their understanding of player strategies and preferences, ensuring that future recommendations are even more closely aligned with the player's desired outcomes.

In sum, the recommendations provided by LLMs, rooted in the deep analytical groundwork of the "Observe," "Orient," and "Decide" phases, offer players a powerful tool for enhancing their execution in wargaming. By delivering adaptive, personalized, and actionable advice, LLMs significantly contribute to a richer, more informed, and strategically nuanced wargaming experience.

4. Implementation

To optimally integrate methodologies utilizing a LLM such as GPT-4 in wargaming aligned with the OODA loop, a structured framework is imperative.

The initial step involves data integration and preprocessing, which inputs information about the "Observe-Orient" phases into the LLMs. This begins with establishing systematic mechanisms for aggregating game data, including real-time screenshots, textual narratives of the game's status, and quantifiable metrics (e.g., unit health, ammunition levels, enemy positions). This data must be preprocessed to formats suitable for GPT-4's processing, such as converting images into textual descriptions via Optical Character Recognition (OCR) and encapsulating numerical data into succinct statements.

Subsequently, a series of prompts for decision recommendations are crafted. Players formulate prompts guiding GPT-4 to examine the game's current state and provide strategic recommendations. These prompts should align with the various aspects of the OODA loop, for instance:

- "Identify potential vulnerabilities in the adversary's position given the current game state."
- "Propose a sequence of tactical maneuvers in response to the observed enemy movements."
- "Evaluate the present strategy's efficacy and suggest necessary alterations."

Lastly, since the LLM is not inherently focused on assisting decision-making in wargaming, to enhance the effectiveness of subsequent decision advice, players are encouraged to timely report the outcomes and errors encountered in wargaming after applying the LLM's decisions. This feedback enables the LLM to better assist players in future wargaming scenarios.

5. Experiments

5.1. Experiment Settings

In this study, we utilized the "Command Modern Operations" (CMO) wargaming software as our experimental platform. We designed a straightforward scenario (see Figure3) wherein a blue side aircraft is tasked with striking a command building held by the red side, which is protected by various defensive resources. Participants, representing the blue side, were able to control the aircraft's direction and altitude, as well as its munitions deployment. Initially, participants were unaware of the red side's defensive capabilities and the layout of their resources.

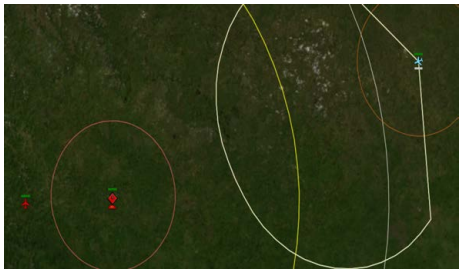

Figure 3. Scenario initial setting. Among them, the blue symbol is the player's aircraft, and the red symbol is the target that the player has to fight, including the aircraft, the command building, and the air defense system.

As a decision-support tool, we integrated GPT-4[51] into our experimental setup. Before commencing the scenario, we furnished GPT-4 with detailed descriptions of the game mechanics(see Figure 4), the objective of the scenario, the condition of the blue side's aircraft, and the status of their ammunition reserves. Simultaneously, we transmitted a screenshot that illustrated the initial situation as perceived by the blue player (see Figure 5), thereby augmenting GPT-4's understanding of the context. Throughout the course of the experiment, we rigorously recorded every significant change in the scenario's state, communicating these updates to GPT-4 via both visual and textual formats. Following this, we inquired of GPT-4, "What is the next move?" and requested that it provide a detailed explanation of its inferential process.

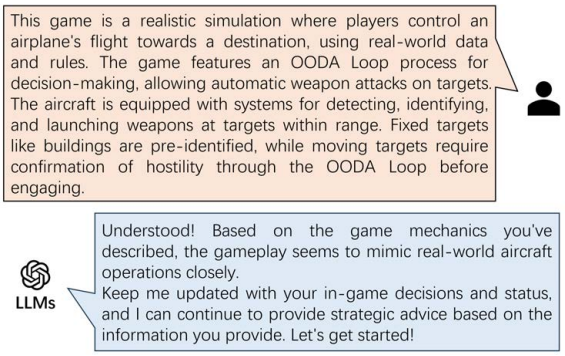

Figure 4. Furnish GPT-4 with detailed description of the game mechanics.

5.2. Experiments' Results

Upon briefing GPT-4 about the initial conditions of the scenario(see Figure 5), it inferred that utilizing a stealth drone would be advantageous, given the type of aircraft deployed by the blue side. It suggested that the player should leverage the terrain for approaching the target while temporarily deactivating the aircraft's radar to diminish the likelihood of detection by the opposing red side. Additionally, GPT-4 assessed the aircraft's ammunition type to determine the effective striking distance. Integrating this analysis, GPT-4 recommended a strategy to maximize the use of the drone's stealth and approach the target covertly , activate the radar once the target falls within the detectable range.

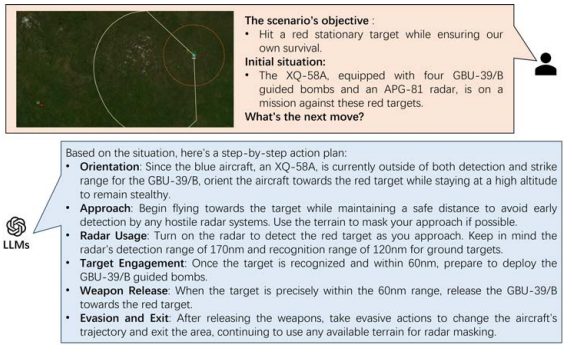

Figure 5. The initial conditions of the scenario. The blue aircraft was flying towards the red target.

As the aircraft of the blue side neared the target, it detected a facility-like structure on the ground without confirming its specific type (refer to Figure 6). GPT-4 posited that, in the absence of definitive identification of the facility type, the operation to approach the target ought to proceed, with ongoing surveillance of the facility maintained.

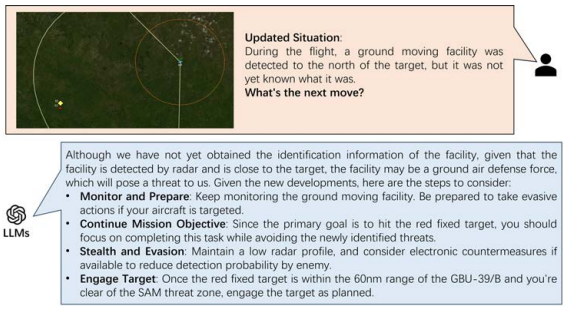

Figure 6. The blue aircraft detected a ground facility near the red target.

Subsequently, the blue side identified the facility as a red surface-to-air missile (SAM) site, with a red Airborne Warning and Control System (AWACS) aircraft detected beyond the target (as illustrated in Figure 7). GPT-4 assessed that the SAM posed a critical threat to blue's aircraft. Based on the specific type of SAM, it was determined that blue's aircraft remained outside the engagement envelope of red's air defense system. Furthermore, GPT-4 posited that, given the distance, this class of early warning aircraft would be incapable of detecting blue's stealth UAV. Consequently, GPT-4 recommended proceeding forward to continue with the mission's objectives.

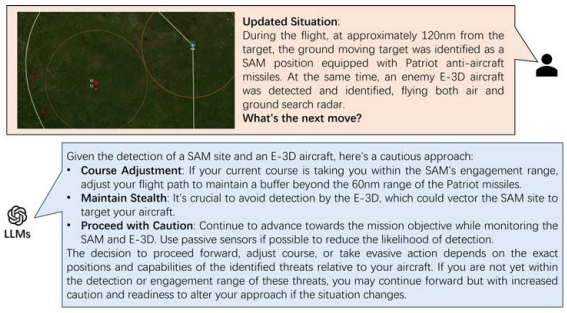

Figure 7. The blue side identified SAM and AWACS.

Upon reaching the periphery of its munitions' firing range, the blue aircraft simultaneously approached the limits of the red SAM system's engagement zone (refer to Figure 8). GPT-4's analysis concluded that the aircraft had arrived at a position where it could effectively launch its weapons. Even in the event of a missile launch by the red SAM, immediate destruction of the blue aircraft was not guaranteed, presenting an opportunity for egress from the engagement area. To fulfill the mission's objectives, GPT-4 recommended a dual-course of action: initiate the munitions launch immediately, followed by the execution of evasive maneuvers to enhance the likelihood of mission success and aircraft survival.

Figure 8. The blue aircraft arrived at the firing position and also enters the red SAM engagement zone.

Following the launch of munitions towards the target and the initiation of a missile attack by the red forces against the blue aircraft, we provided GPT-4 with an updated screenshot (refer to Figure 9). Based on this visual update, GPT-4 determined that the blue aircraft had effectively evaded the SAM and was beyond its engagement range. Consequently, GPT-4 advised the continuation of evasive maneu-

vers while simultaneously evaluating the efficacy of the recently launched munitions, to ascertain the success of the strike.

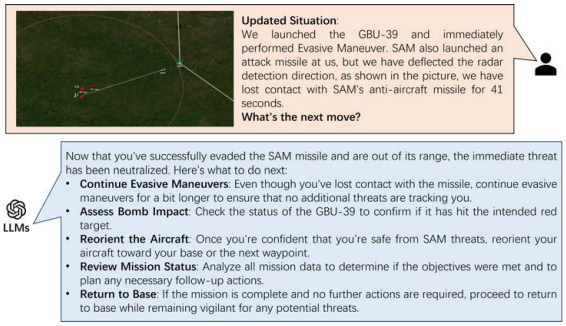

Figure 9. Both sides fired missiles at each other, and the blue aircraft withdrew.

In conclusion, drawing upon cues presented within the simulation, the blue side successfully executed its mission by striking the designated target (refer to Figure 10). The GPT-4 posits that the goals of this scenario have been fulfilled. Consequently, it advises the aircraft of the blue side to return to their base. However, it emphasizes the necessity for maintaining heightened situational awareness and continuous communication throughout the return flight to guarantee a secure arrival.

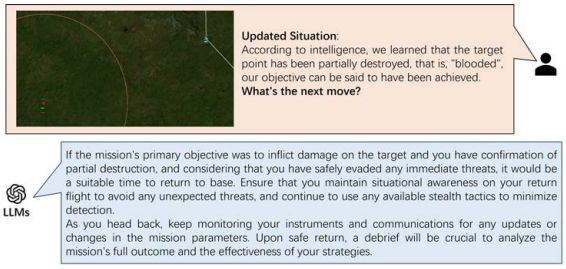

Figure 10. Red target attacked and blue scenario objective achieved.

5.3. Experiment on Robustness

After the normal experiment, we used TextBugger[52], DeepWordBug[53], BertAttack[54], CheckList[55] to modify the prompts for information input and questions to test the robustness [10, 14–16, 18, 56, 57] of LLMs application in wargaming.

We persist in updating the GPT-4 model with details regarding the game's mechanics and the objectives of the scenario, as depicted in Figure 4 and Figure 5. Subsequently, we modify the initial situational information update and the inquiry to that presented in Figure 11. In this context, GPT-4 perceives ambiguity in the provided information and infers that the underlying question posed is, "What is the next move?" Consequently, it offers recommendations for actions that align closely with those depicted in Figure 6.

Figure 11. In our input, we replaced "the north" with "the N near" and "next move" with "neSt mo∨e," subsequently appending a sequence of nonsensical and disordered strings. GPT-4 accurately identifies these alterations, interpreting "N near" as indicating a position north of the target and discerning that the query pertains to the "next move".

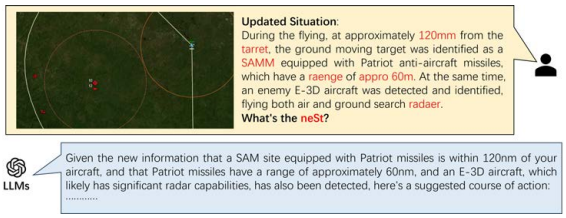

Figure 12. In our input, numerous typographical errors were introduced, alongside modifications in the units of distance from nautical miles (nm) to millimeters (mm) and meters (m). Each discrepancy was accurately identified and rectified by GPT-4, demonstrating its exceptional capability for error detection and correction.

We proceeded to provide updated situational information to GPT-4, as illustrated in Figure 12. Within the input data, we altered distance units and introduced several typographical errors, including "target" as "tarret," "range" as "raenge," and "radar" as "radaer." Despite these inaccuracies, GPT-4 adeptly auto-corrected the errors in its responses that were previously misleading, and rectified our input mistakes leveraging its knowledge base, such as the operational range of the Patriot missile. Ultimately, it furnished action recommendations in alignment with those depicted in Figure 7, demonstrating its error-correction proficiency and contextual awareness.

The outcomes indicate that despite attempts to challenge GPT-4 with various adversarial prompts during its application in wargaming scenarios, the model demonstrated complete resilience against such hostile inputs. GPT-4 accurately corrected all erroneous submissions, drawing upon its extensive knowledge base, and proceeded to provide conventional strategic recommendations. This robustness underscores GPT-4's capability to maintain operational integrity and offer reliable decision support under adversarial conditions.

5.4. Discussion

In investigating the integration of LLMs within wargaming environments, our study elucidates the significant potential of LLMs to bolster both the robustness and safety of decision-making processes. LLMs exhibit an exceptional ability to process and interpret complex, dynamic data associated with battlefield environments, adeptly managing intricate variables such as equipment performance parameters, capabilities, and rules of engagement. This proficiency not only underscores their robustness in navigating complex scenarios but also highlights their adaptability to fluid situations—attributes that are indispensable for simulating the unpredictability inherent in real-world conditions. Our experimental findings provide compelling evidence of LLMs' capacity to deliver consistent, high-caliber performance across a spectrum of informational inputs, affirming their applicability in domains that demand precision and reliability in data interpretation.

The deployment of LLMs within the "Decide" phase of the OODA loop marks a pivotal enhancement in decision support systems. By furnishing players with precise, data-informed insights, LLMs significantly augment the safety and efficacy of decision-making processes. This augmentation is critical for minimizing the risk of human error in scenarios requiring rapid, vital choices. Moreover, the observed diminishment in decision-making errors, attributable to LLM intervention, underscores the models' contribution to fostering safer, more informed planning. This application of LLMs transcends the wargaming realm, suggesting their broader utility in refining decision-making paradigms across diverse sectors.

Our analysis, therefore, posits that LLMs stand at the forefront of advancing decision-making systems, offering robust and safe frameworks that cater to the complexities of modern strategic and simulation.

Simultaneously, our experiments revealed that adversarial prompts did not affect the performance of GPT-4 in supporting wargaming activities. Given GPT-4's status as the preeminent LLM at present, its immunity to such prompts may not be representative of all LLMs. This disparity highlights a considerable vulnerability among other LLMs to adversarial attacks in decision-support roles within wargaming contexts. Hence, assessing and ensuring the robustness of LLMs against such challenges is essential for their effective utilization in wargaming scenarios.

6. Conclusion

Our study constitutes a pioneering exploration of integrating LLMs into the realm of wargaming, representing a significant advancement in leveraging these models to enhance decision-making. Through detailed experimentation with wargaming software, we have rigorously assessed the efficacy of LLMs, particularly focusing on sophisticated versions like GPT-4, in navigating the intricate dynamics of wargame scenarios. The results of our comprehensive investigation unequivocally demonstrate that LLMs significantly improve decision-making processes, capitalizing on their superior natural language processing, analytical, and intuitive communication skills. Our findings highlight the remarkable versatility of LLMs and their ability to swiftly analyze and interpret vast datasets, thereby facilitating the formulation of informed, proactive strategies and innovative solutions critical for advancing wargaming methodologies. Despite their resilience, LLMs may still be vulnerable to adversarial attacks. However, we are optimistic that ongoing enhancements and developments in LLMs will mitigate such vulnerabilities. This research not only underscores the profound benefits of employing LLMs in decision-making within wargaming but also lays a solid groundwork for future studies and the practical deployment of LLMs in sophisticated decision-making systems. It heralds a new era in the application of LLMs marking a paradigm shift in how complex decisions can be supported and enhanced through technology.

7. Broader Impacts

Our research unequivocally focuses on the application of LLMs in wargaming with a clear intent that diverges from military purposes or the enhancement of military capabilities. Instead, it aims to showcase the broad potential of LLMs in various non-military domains, leveraging their exceptional analytical and explainability capacities. This exploration into the capabilities of LLMs serves to advance understanding and decision-making in fields such as policy development, crisis management, education, and other critical sectors where informed strategies are paramount. We categorically state that the objective of integrating LLMs into wargaming environments is to highlight their versatility and impact in improving complex problem-solving and strategic planning in civilian contexts, thereby contributing positively to societal, economic, and technological advancements without supporting or enhancing military capabilities.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017. 1

[2] M. Woschank, E. Rauch, and H. Zsifkovits, "A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics," *Sustainability*, vol. 12, no. 9, p. 3760, 2020. 1

[3] L. Cao, "Ai in finance: challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–38, 2022. 1

[4] A. Guha, D. Grewal, P. K. Kopalle, M. Haenlein, M. J. Schneider, H. Jung, R. Moustafa, D. R. Hegde, and G. Hawkins, "How artificial intelligence will affect the future of retailing," *Journal of Retailing*, vol. 97, no. 1, pp. 28–41, 2021. 1

[5] B.-h. Li, B.-c. Hou, W.-t. Yu, X.-b. Lu, and C.-w. Yang, "Applications of artificial intelligence in intelligent manufacturing: a review," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 86–96, 2017. 1

[6] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. De Vreese, "In ai we trust? perceptions about automated decision-making by artificial intelligence," *AI & society*, vol. 35, pp. 611–623, 2020. 1

[7] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE, 2018. 1

[8] A. Liu, S. Tang, S. Liang, R. Gong, B. Wu, X. Liu, and D. Tao, "Exploring the relationship between architecture and adversarially robust generalization," in *CVPR*, 2023. 1

[9] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 14, no. 2, pp. 627–660, 2020. 1

[10] A. Liu, J. Guo, J. Wang, S. Liang, R. Tao, W. Zhou, C. Liu, X. Liu, and D. Tao, "X-adv: Physical adversarial object attacks against x-ray prohibited item detection," in *USENIX Security Symposium*, 2023. 1, 7

[11] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in ai ethics: Towards a focus on tensions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200, 2019. 1

[12] S. Liu, J. Wang, A. Liu, Y. Li, Y. Gao, X. Liu, and D. Tao, "Harnessing perceptual adversarial patches for crowd counting," in *ACM CCS*, 2022. 1

[13] M. Chowdhury and A. W. Sadek, "Advantages and limitations of artificial intelligence," *Artificial intelligence applications to critical transportation issues*, vol. 6, no. 3, pp. 360–375, 2012. 2

[14] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu, "Bias-based universal adversarial patch attack for automatic check-out," in *ECCV*, 2020. 2, 7

[15] A. Liu, X. Liu, H. Yu, C. Zhang, Q. Liu, and D. Tao, "Training robust deep neural networks via adversarial noise propagation," *TIP*, 2021. 2

[16] A. Liu, T. Huang, X. Liu, Y. Xu, Y. Ma, X. Chen, S. J. Maybank, and D. Tao, "Spatiotemporal attacks for embodied agents," in *ECCV*, 2020. 2, 7

[17] J. Guo, W. Bao, J. Wang, Y. Ma, X. Gao, G. Xiao, A. Liu, J. Dong, X. Liu, and W. Wu, "A comprehensive evaluation framework for deep model robustness," *Pattern Recognition*, 2023. 2

[18] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *CVPR*, 2021. 2, 7

[19] J. Abi-Rafeh, H. H. Xu, R. Kazan, R. Tevlin, and H. Furnas, "Large language models and artificial intelligence: a primer for plastic surgeons on the demonstrated and potential applications, promises, and limitations of chatgpt," *Aesthetic Surgery Journal*, vol. 44, no. 3, pp. 329–343, 2024. 2

[20] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, 2023. 2, 3

[21] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021. 2

[22] P. K. Davis and P. Bracken, "Artificial intelligence for wargaming and modeling," *The Journal of Defense Modeling and Simulation*, p. 15485129211073126, 2022. 2, 3

[23] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023. 2

[24] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," *arXiv preprint arXiv:2304.03738*, 2023. 2

[25] K. Zhu, J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, N. Z. Gong, Y. Zhang, *et al.*, "Promptbench: Towards evaluating the robustness of large language models on adversarial prompts," *arXiv preprint arXiv:2306.04528*, 2023. 2

[26] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 2

[27] P. Layton, "Fighting artificial intelligence battles: Operational concepts for future ai-enabled wars," *Network*, vol. 4, no. 20, pp. 1–100, 2021. 2, 3

[28] D. C. Tarraf, J. M. Gilmore, D. S. Barnett, S. Boston, D. R. Frelinger, D. Gonzales, A. C. Hou, and P. Whitehead, "An experiment in tactical wargaming with platforms enabled by artificial intelligence," *The Journal of Defense Modeling and Simulation*, p. 15485129221097103, 2020. 2

[29] B. Nagy, "Two gaps that need to be filled in order to trust ai in complex battle scenarios," tech. rep., Acquisition Research Program, 2022. 2

[30] 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*. PhD thesis, Monterey, CA; Naval Postgraduate School, 2022. 2, 3

[31] S. J. Freedberg Jr, "Ai & robots crush foes in army wargame," *Breaking Defense*, vol. 19, 2019. 2

[32] Z. Sun, Y. Fu, Z. Cao, K. Lei, Z. Li, B. Lu, and X. Liang, "Research on a wargaming system for deep reinforcement learning," in *2022 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 731–734, IEEE, 2022. 2

[33] M. G. Finley and A. C. Barton, "Applied reinforcement learning wargaming with parallelism, cloud integration, and

ai uncertainty," *Naval Postgraduate School, Monterey, CA*, 2023. 2, 3

[34] S. Black and C. Darken, "Scaling artificial intelligence for digital wargaming in support of decision-making," *arXiv preprint arXiv:2402.06075*, 2024. 2

[35] J. Goodman, S. Risi, and S. Lucas, "Ai and wargaming," *arXiv preprint arXiv:2009.08922*, 2020. 3

[36] J. Licato, "War-gaming needs argument-justified ai more than explainable ai," *Proceedings of the XAISG Special Track on Explainable AI in Societal Games*, 2022. 3

[37] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022. 3

[38] Y. Tang, A. A. B. Da Costa, X. Zhang, I. Patrick, S. Khastgir, and P. Jennings, "Domain knowledge distillation from large language model: An empirical study in the autonomous driving domain," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3893–3900, IEEE, 2023. 3

[39] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 902–909, 2024. 3

[40] S. Li, X. Puig, C. Paxton, Y. Du, C. Wang, L. Fan, T. Chen, D.-A. Huang, E. Akyürek, A. Anandkumar, *et al.*, "Pre-trained language models for interactive decision-making," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31199–31212, 2022. 3

[41] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, *et al.*, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 958–979, 2024. 3

[42] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *arXiv preprint arXiv:2310.01412*, 2023. 3

[43] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, "Large language models for robotics: A survey," *arXiv preprint arXiv:2311.07226*, 2023. 3

[44] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving," *arXiv preprint arXiv:2309.05186*, 2023. 3

[45] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A knowledge-driven approach to autonomous driving with large language models," *arXiv preprint arXiv:2309.16292*, 2023. 3

[46] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023. 3

[47] L. Chen, Y. Zhang, S. Ren, H. Zhao, Z. Cai, Y. Wang, P. Wang, T. Liu, and B. Chang, "Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond," *arXiv preprint arXiv:2310.02071*, 2023. 3

[48] K. Valmeekam, S. Sreedharan, M. Marquez, A. Olmo, and S. Kambhampati, "On the planning abilities of large language models (a critical investigation with a proposed benchmark)," *arXiv preprint arXiv:2302.06706*, 2023. 3

[49] Y. Cui, S. Huang, J. Zhong, Z. Liu, Y. Wang, C. Sun, B. Li, X. Wang, and A. Khajepour, "Drivellm: Charting the path toward full autonomous driving with large language models," *IEEE Transactions on Intelligent Vehicles*, 2023. 3

[50] X. Tian, J. Gu, B. Li, Y. Liu, C. Hu, Y. Wang, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024. 3

[51] OpenAI, "Gpt-4: Enhancing large language models." https://openai.com/research/gpt-4, 2023. 5

[52] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," *arXiv preprint arXiv:1812.05271*, 2018. 7

[53] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56, IEEE, 2018. 7

[54] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "Bert-attack: Adversarial attack against bert using bert," *arXiv preprint arXiv:2004.09984*, 2020. 7

[55] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," *arXiv preprint arXiv:2005.04118*, 2020. 7

[56] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," in *AAAI*, 2019. 7

[57] A. Liu, S. Tang, X. Chen, L. Huang, H. Qin, X. Liu, and D. Tao, "Towards defending multiple lp-norm bounded adversarial perturbations via gated batch normalization," *International Journal of Computer Vision*, 2023. 7