BattleAgent: Multi-modal Dynamic Emulation on Historical Battles to Complement Historical Analysis

Shuhang Lin[1*] **Wenyue Hua**[1*] **Lingyao Li**[2] **Che-Jui Chang**[1] **Lizhou Fan**[2] **Jianchao Ji**[1]
Hang Hua[3] **Mingyu Jin**[1] **Jiebo Luo**[3] **Yongfeng Zhang**[1]
[1]Department of Computer Science, Rutgers University, New Brunswick
[2]School of Information, University of Michigan, Ann Arbor
[2]School of Computer Science, University of Rochester
[*]Shuhang Lin and Wenyue Hua contribute equally.

Abstract

This paper presents **BattleAgent**, a detailed emulation demonstration system that combines the Large Vision-Language Model (VLM) and Multi-Agent System (MAS). This novel system aims to emulate complex dynamic interactions among multiple agents, as well as between agents and their environments, over a period of time. The emulation showcases the current capabilities of agents, featuring fine-grained multi-modal interactions between agents and landscapes. It develops customizable agent structures to meet specific situational requirements, for example, a variety of battle-related activities like scouting and trench digging. These components collaborate to recreate historical events in a lively and comprehensive manner. This methodology holds the potential to substantially improve visualization of historical events and deepen our understanding of historical events especially from the perspective of decision making. The data and code for this project are accessible at https://github.com/agiresearch/battleagent. The demo is accessible at https://drive.google.com/file/d/1I5B3KWiYCSSP1uMiPGNmXlTmild-MzRJ/view?usp=sharing.

1 Introduction

An agent is defined as a system that has the ability to perceive its surroundings and make informed decisions based on these perceptions to achieve specific objectives (Xi et al., 2023). Recent progress in large language models (LLMs) (Zhao et al., 2023; Fan et al., 2023) has demonstrated impressive reasoning capabilities (Huang and Chang, 2022; Jin et al., 2024), indicating their potential to serve as the foundation for agents. Additionally, the development of large Vision Language Models (VLM) (Zhang et al., 2024) has facilitated the creation of various agent applications that support multi-modal information interaction (Durante et al., 2024; Xie et al., 2024b). When combined with external tools, either physical or virtual, these agents employ LLM or VLM as their reasoning backbone to determine how tasks should be addressed, how tools should be utilized, and what information should be retained in memory. This enhancement equips agents to manage an array of natural language processing tasks and engage with their environment using language.

Numerous agent applications have been created using LLM and VLM, with a focus on improving reasoning (Du et al., 2023; Chan et al., 2023; Sun et al., 2023; Liang et al., 2023), production capabilities (Hong et al., 2023; Liu et al., 2023a; Ge et al., 2023a; Yang et al., 2023; Mei et al., 2024; Ge et al., 2023b), gaming (Gong et al., 2023; Xu et al., 2023; Lan et al., 2023; Hu et al., 2024), and social simulation (Pang et al.; Zhou et al., 2024; Sreedhar and Chilton, 2024; Xie et al., 2024a; Hua et al., 2023), among others. WarAgent (Hua et al., 2023) is the pioneering LLM-based MAS simulation of historical events, examining the behaviors of systems at the macro level, such as nations and governments, rather than the micro-level simulation of detailed and dynamic events occurring during battles or individual experiences in such dynamic time periods. Therefore, BattleAgent, building on the foundation laid by WarAgent in historical event simulation, investigates the potential of LLM and VLM for detailed historical situation recovery and the exploration of individual experiences within the simulation.

To emulate such a complex scenario, our emulation incorporates the following three key features:
Enhanced 2-D Realism Features: BattleAgent emulates detailed interactions within environments, including terrain engagement, temporal progression, and interactions between agents.
Immersive Multi-agent Interactions: It integrates MAS to facilitate dynamic interactions among agents in battle emulations, accurately reflecting the historical milieu and the intricacies of military engagements, from strategic maneuvers to logisti-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 172–181
November 12-16, 2024 ©2024 Association for Computational Linguistics

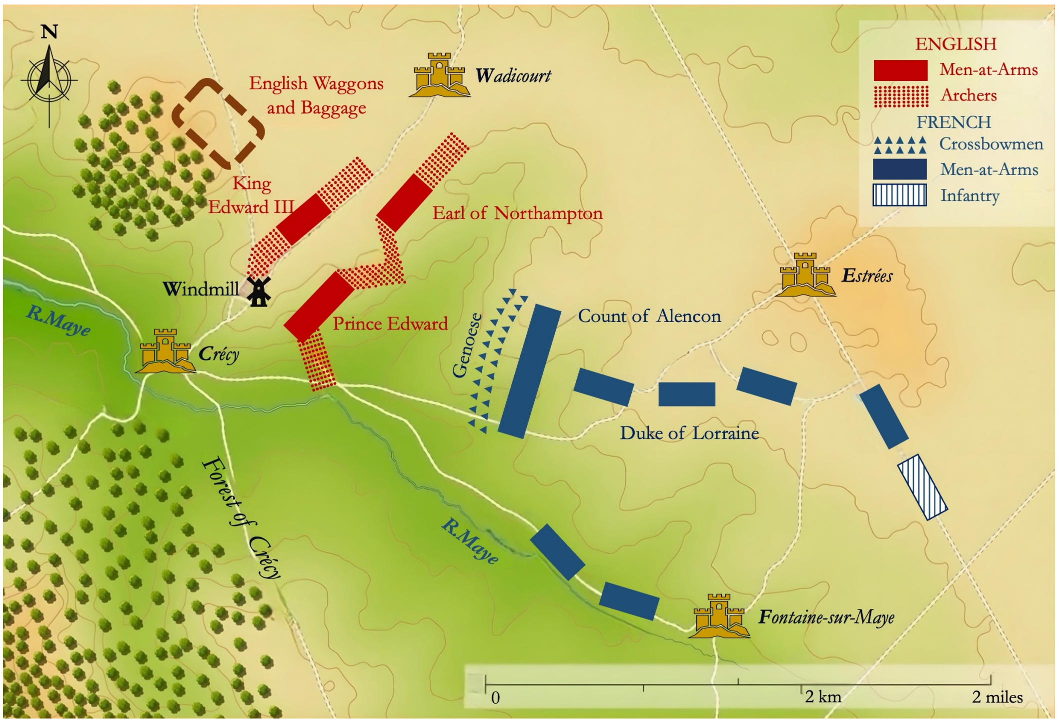

Figure 1: Demonstration of the emulated Battle of Crécy, 1346: Troop formations and movements depicting the positions of the English and French forces during the historical engagement, with key locations and leaders marked. Image adjusted from https://the-past.com/feature/the-battle-of-crecy-26-august-1346/

cal considerations and communication dynamics.
Dynamic Agent Structure: The framework introduces adaptable agent configurations and multi-modal interactions. The system can "self improvise" its structure to fork, merge, and prune agents to continuously maintain the emulation effectiveness. It boasts the capability to autonomously adjust its architecture to optimize emulation fidelity.

The contributions of our study to historical analysis and society can be summarized as follows:
Connection and resonance with the past: Helping to prevent future conflicts by learning from the detailed analysis of past mistakes and human costs. This platform fosters empathy and a deeper connection to the past by humanizing the experiences of those involved in historical battles.
Educational tool for understanding history: Providing an educational tool to help people understand the intricacies of history and the harsh realities of historical events. Its immersive and interactive platform can foster empathy and a more nuanced perspective on the past, making it a valuable resource for students and history enthusiasts.
Potential as a next-generation game engine: Providing a fully automated process to create immersive and dynamic historical emulations, making it a potential next-generation game engine. By using LLM-based agents and VLM-based agents, it can generate detailed and realistic environments, characters, and events, offering a unique and engaging gaming experience.

2 Emulation Setting

This section outlines the emulation framework and setting for our research demonstration. We commence with an exposition of the historical context of the four significant European battles that our emulation seeks to emulate: the Battle of Crécy, the Battle of Agincourt, the Battle of Poitiers, and the Battle of Falkirk. Each battle has been selected for its notable use of cold weapons and the strategic bipartite confrontations that characterized warfare during their respective periods. Building upon the historical context, we elaborate on the configuration of agents and their designated roles within our emulation framework.

2.1 Agent Definition

Each agent represents an army. Decisions and strategies of the agent will be made based on the general information in the army profile, which includes the following aspects: **(1) ID**: The ID of a agent is represented by a hash code that is generated to uniquely identify each agent within the emulation sandbox. This is necessary due to the dynamic agent structure employed in our emulation, which allows for the creation of additional agents beyond the initial (two) agents as the emulation progresses. The use of a hash code ensures that each agent can be accurately identified and tracked throughout the course of the emulation. **(2) Military Command Structure**: This involves the hierarchical organization and leadership dynamics within each military faction. **(3) Morale and Discipline**: An assessment of the troops' psychological readiness, their discipline levels, and overall morale. **(4) Military Strategy**: The overarching tactical approaches and plans employed by each side in the conflict. **(5) Military Capability**: An inventory of the weapons and defense tools at each side's disposal. **(6) Force size and composition**: This aspect includes the total number of soldiers and their composition including information about the types of troops, their roles, and their proportions in the overall force. **(7) Location**: The current location of the agent is represented by its coordinates. These coordinates provide a precise indication of the agent's position within the sandbox environment, allowing for accurate tracking and analysis of its movements and interactions with other agents and the environment.

2.2 Action Space

Our emulation framework contains an action space with 51 distinct actions. Agents within the emulation have the flexibility to select any combination of these actions at each decision point. The actions available in the action space are organized into six categorically distinct groups: **(1) Reposition**. This category includes actions that involve the movement of an army or a subsection thereof to a different location: *Reposition Forces, Create Decoy Units* **(2) Preparation**. Actions in this group are geared towards readying forces for an impending attack: *Deploy Longbows, Rally Troops, Employ Artillery, Use of Gunpowder Weapons, Resupply Archers, Destroy Enemy Morale, Deploy Archers in Flanking Positions, Organize Night Raids, Organize Raiding Parties, Digging trenches* **(3) Attack**. This group encapsulates a variety of common attack strategies, such as skirmishing, ambushing, besieging, cavalry charges, and direct firing, among others: *Initiate Skirmish, Charge Cavalry, Ambush Enemy, Launch Full Assault, Archery Duel, Siege Tactics, Hand-to-Hand Combat, Counterattack, Conduct Reconnaissance, Direct Artillery Fire, Engage in Siege Warfare, Execute Flanking Maneuvers, Use Cavalry for Shock Tactics, Employ Archers Strategically* **(4) Defense**. Encompasses actions such as shielding, fortification, and the creation of obstacles: *Construct Defenses, Prepare Defenses, Develop Counter-Siege Measures, Form Defensive Shields, Establish Defensive Fortifications, Fortify Rear Guards, Fortify Position, Create Obstacles for Enemy Cavalry, Form Defensive Pike Formations, Set Traps* **(5) Observation**. Focused on gathering information about the surrounding area and the current situation of the enemy: *Scout Enemy Position, Gather Intelligence, Intercept Enemy Supplies, Establish Communication Lines* **(6) Retreat**. Actions related to strategic withdrawal in the face of adverse conditions: *Retreat and Regroup, Tactical Retreat, Plan Feigned Retreat*

3 Emulation Sandbox

In our emulation framework, we concentrate on a relatively straightforward scenario: a bipartite battle. The process begins with (1) setting up the geographical context for the entire scenario, both textual description as well as a visual map, and (2) define the two initial opposing agents, each represents the army of one country. This section will introduce the emulation process: we first present an overview of the sandbox emulation process from a high-level perspective and then delve into the details of the process. This includes how time and location are represented and processed, how agent actions are determined, and how the results of these actions are computed.

3.1 General Sandbox Emulation Process

Here we provide a very simple and crude overview of the emulation sandbox. We initiate the emulation based on historical map which contain information about geography as well as the position of the armies. The following represents a high-level overview of the steps involved in the emulation process: **Step 1:** Each agent starts by observing its surroundings and gathering information. This observation process involves text-based description of overall environment which are inputted to the agent by prompt as well as direct visual information taking the map as input. **Step 2:** Based on the gathered information, each agent decides on its actions, such as preparing for battle (e.g., digging trenches, reinforcing troops), collecting further information, or making organizational changes to dynamically split armies into smaller units or merge armies with other allied armies. **Step 3:** For every

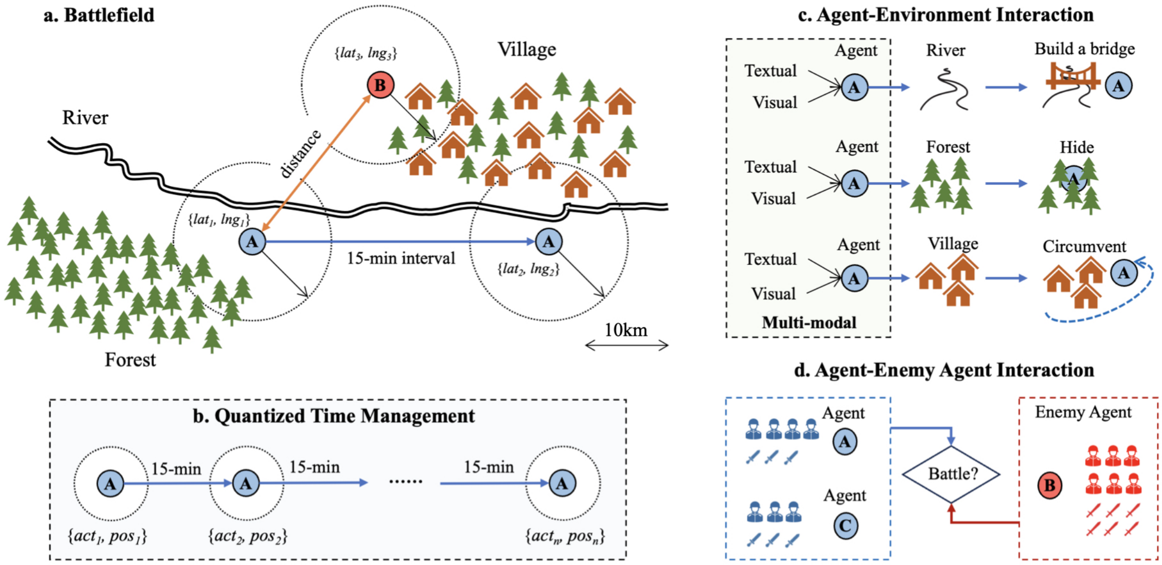

Figure 2: Battlefield interaction (a) Battlefield environment, (2) Quantized time management, (c) Agent-environment interaction, and (d) Agent and enemy agent interaction.

15-minute interval in the emulation sandbox, agent information such as their locations and properties and corresponding visual change in map is updated according to the actions taken by all agents. **Step 4:** An objective LLM-based observer computes the impact of agent actions especially casualty loss in agent. **Step 5:** The process then loops back to Step 1, with agents continuing to observe, make decisions, and act based on the updated information and evolving battlefield situation.

3.2 Time and Space in Sandbox

In order to accurately emulate the dynamics of historical battles, it is crucial to effectively manage the time and space within the sandbox environment. In this section, we introduce our approach to time and space management in the sandbox.

Quantized Time Management The battlefield environment is characterized by continuous dynamic changes. Therefore, to emulate these dynamics while preserving the discrete decision-making process in our agent-based emulation, we employ a time quantization approach. Specifically, we discretize the continuous flow of time (Matsuoka et al., 2001; Al Rowaei et al., 2011) into 15-minute intervals in sandbox. For each quantized time block, agents have the flexibility to either maintain their current action or adapt their actions.

Coordinate Generation based on Map We obtain the initial map of the battlefield from historical documents (Kiffer, 2019; Curry, 2000). These agents take both textual description of the map as well as the visual map as input (for agents with multi-modal LLM as backbones). Thus we need to generate the coordinates from the original image for description. We use one army position as the reference point, designated as the (0,0) position. We then use a scale of 10 yards as one unit of the coordinate system. The coordinates of key landscapes on the map such as villages and castles and their distances with each other and with agents are estimated and provided.

3.3 Action Planning

At each discrete time point, an agent has the ability to choose from a multitude of potential actions. In this part, we will outline four common types of actions that agents typically engage in: location movement, dynamic agent structure, interaction with the landscape, and interaction with other agents. These actions require a range of strategic considerations that agents must take into account when making decisions in the context of the battlefield.

Location Movement In the context of location movement, an agent possesses the capability to traverse to a different location for strategic purposes. This may involve moving closer to enemy agents to initiate an attack, or distancing itself from potential threats. In terms of the mechanics of location movement, the agent will generate the coordinates of its intended final destination, which it aims to reach within the subsequent 15-minute timeframe.

Dynamic Agent Structure The battlefield environment is highly dynamic and fluid, with a multitude of situations arising unpredictably. To address this complexity, we propose a dynamic agent structure (Liu et al., 2023b; Han et al., 2024) that enables agents to adapt their organizational configurations according to the current situation. Our proposed dynamic agent structure supports several adaptive mechanisms, as shown in Figure 3:

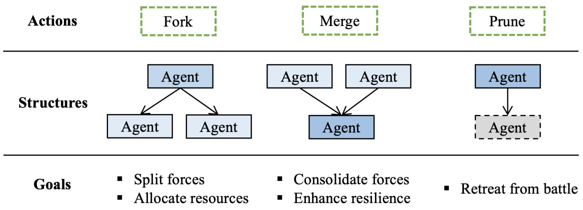

Figure 3: Dynamic agent structure.

Fork: An agent may decide to fork another autonomous agent for a specific task, splitting its forces and allocating resources to address multiple objectives simultaneously. **Merge:** In scenarios where an agent is under significant pressure but chooses to continue fighting, it may merge with the closest allied agent to consolidate forces and enhance its resilience. **Prune:** In cases where an agent is overwhelmed or retreats from the battlefield, the dynamic agent structure accommodates this change by pruning it from the active force.

Each newly created agent will inherit profile information of the country army that it belongs to, but also includes more granular and unique information: (1) Initial mission assigned when being created (2) Current location represented by coordinates (3) The number of soldiers at its disposal (4) The type of soldiers under its command. These properties are subject to evolution over time. For instance, the number of soldiers associated with an agent may fluctuate as soldiers joining the agent, thereby increasing its forces, or from soldiers being killed or wounded in battle, leading to a decrease in its forces. The current location of the agent may also change as it navigates the battlefield, and its initial mission may adapt in response to shifting circumstances and strategic considerations.

Interaction with Landscape Environment To accurately emulate battle dynamics, it is crucial for agents to be able to interact with the physical surroundings as shown in Figure 2 (c), such as rivers, forests, villages, and other features. For example, when encountering a river, agents may build a bridge to cross it; when encountering a forest, agents might choose to hide within it to ambush enemies; and when encountering a village, agents could decide to circumvent it. To facilitate these interactions, it is essential to maintain a relative distance between agents and specific locations on the map, as well as between agents themselves.

Interaction with Other Agents Given the observation agents make about their surrounding situations, agents will make decisions regarding whether and when to engage in interactions with other agents, particularly those identified as enemies, as depicted in Figure 2 (d). The specific nature and timing of these interactions are not predetermined; rather, they are initiated by the agents themselves. For instance, when an enemy agent is within close proximity, an agent may opt to engage in combat or launch an attack. The outcome of these interactions between agents is contingent upon various factors, such as the number of soldiers at their disposal and the types of weapons they possess.

3.4 Casualty Evaluation by Observer

In the event that one agent initiates an aggressive action towards another, hereafter referred to as the target agent, both parties may sustain casualty losses. The loss is evaluated by an objective evaluator supported by GPT-4, which can be seen as an observer. The observer determines the casualties based on several factors: (1) Current information of the agents, including their force size, force composition, and command architecture. (2) The actions undertaken by the agents, including the action name and a more detailed description of the action generated alongside the action name by the agent. For example, "Deploy Longbows: Deploying longbows in coordination with nearby friendly forces to initiate a skirmish against the nearest enemy cavalry unit and disrupt their advance." (3) The location and relative distance between the agents, as well as relevant landscape information surrounding them. (4) Objective information about the specific weapon utilized, including weapon parameters, such as range and damage.

Evaluation aspect	Description
Final battle casualty	Comparison with historical data, focusing on the final casualty figures for both armies
Human analysis on location movement	Assessment of the dynamic structure of agents and their movement on the battlefield as a whole
Human analysis of agent action	Evaluation of the reasonableness of the actions conducted by the agents.

Table 1: Three aspects of evaluation and demonstration.

Battle	**Model**	**France/Scotland**		**England**	
		Casualties	**Historical Casualties**	**Casualties**	**Historical Casualties**
Crécy	Claude-3	19.2k ± 8.3k	10k - 30k	7.7k± 2.5k	100 - 300
	GPT-4	10.1k±2.5k		3.8k± 2.0k	
	GPT-4-vision	14.0k ± 2.5k		4.5k ± 2.0k	
Agincourt	Claude-3	27.5k ± 5.0k	4k - 10k	5.7k ± 0.1k	0.1k - 1.5k
	GPT-4	5.3k ± 0.4k		2.8k ± 0.1k	
	GPT-4-vision	8.3k ± 0.1k		2.9k ± 0.1k	
Poitiers	Claude-3	10.1k ± 2.3k	5k - 7k	3.6k ± 1.3k	40
	GPT-4	6.8 k ± 1.0k		1.9k ± 0.7k	
	GPT-4-vision	4.8k ± 1.8k		2.3k ± 0.5k	
Falkirk	Claude-3	5.4k ± 0.4k	2k	8.1k ± 1.6k	2k
	GPT-4	2.2k ± 1.0k		1.9k± 0.7k	
	GPT-4-vision	2.0k ± 1.3k		1.9k ± 0.9k	

Table 2: Casualties in historical battles predicted by different models with mean and standard deviation

4 Experiment

The primary objective of these experiments is to investigate the extent to which agents based on LLMs and VLMs can reasonably emulate historical battles, which are characterized by a high degree of complexity and dynamism. We conduct experiments on 4 distinct historical scenarios, namely the Battle of Crécy, the Battle of Agincourt, the Battle of Falkirk, and the Battle of Poitiers. The experiments are performed using 3 strong language models and vision-language models: Claude-3-opus (Anthropic, 2024), GPT-4-1106-preview (Achiam et al., 2023), and GPT-4-vision (OpenAI, 2023). For each scenario and each language model, we execute the emulation 5 times using the same setting to account to randomness, continuing until the casualty figures for both armies converge, or in other words, reach a state of stability.

We employ three evaluation metrics as described in Table 1. The final battle casualty metric quantitatively assesses whether the simulation's final prediction of losses aligns with historical records. Given the challenge of directly evaluating the validity or authenticity of the simulation process due to the typical scarcity of detailed historical documentation, we rely on evaluating the final casualty results. Table 2 presents a comparison of the emulated casualties and historical casualties for all experiments, with more detailed results provided in Appendix A.2. The evaluation of location movement and agent actions is based on human analysis and visualization, with example visualizations available in Appendix A.1 and Appendix A.3 respectively. In general, we observed that current LLMs exhibit a limited understanding of distance, which affects location movement decisions.

5 Conclusions and Future Work

In this study, we have demonstrated the potential of LLM and VLM to support highly complex and dynamic simulations of historical battles. Our emulation sandbox provides a comprehensive evaluation of the emulated battles, including a comparison of casualty figures with historical data and a human analysis of the strategies and tactical maneuvers employed by both armies. We believe that our work can also provide new pedagogical methods for students and researchers interested in historical analysis. By simulating historical battles and presenting the results in an interactive and intuitive way, students can gain a deeper understanding of the complexities and dynamics of warfare.

Limitations

The present study has illustrated the potential of Large Language Models (LLMs) and Visual Language Models (VLMs) in facilitating intricate and dynamic simulations of historical battles. However, as a pioneering work in complex situational event simulation, there are several areas that warrant improvement and further development.

Firstly, the current evaluation methods are constrained. Quantitative evaluation is predominantly limited to casualty counts, particularly at the conclusion of battles. For other aspects, such as the decisions made by agents and their movements, the analysis is heavily reliant on manual methods. Therefore, there is a need for **additional evaluation metrics** to comprehensively establish the effectiveness of these dynamic simulations. Such metrics would enable a more thorough assessment of the accuracy and reliability of the simulation results and help identify areas for enhancement.

Secondly, the current scope of our simulation is restricted to **different types of battles beyond barpitite medieval battles**. Future work should aim to extend these simulations to a more diverse range of scenarios. This expansion will allow for a more robust evaluation of the versatility of our approach and its applicability to a broader spectrum of historical battles.

Thirdly, the current system does not **integrate expert systems** for various components of the simulation, such as information gathering for observation and casualty estimation. Incorporating such systems would enhance the accuracy and realism of the simulation results, while LLMs would continue to be responsible for decision-making processes.

In summary, our future work aims to extend and refine our approach to provide even more realistic and comprehensive simulations of historical battles. This will involve capturing the complexities and dynamics of warfare and offering valuable insights into the strategies and tactics employed by both armies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmed A Al Rowaei, Arnold H Buss, and Stephen Lieberman. 2011. The effects of time advance mechanism on simple agent behaviors in combat simulations. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pages 2426–2437. IEEE.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

Anne Curry. 2000. *The battle of Agincourt: sources and interpretations*. Boydell Press.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2023. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*.

Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023a. OpenAGI: When LLM meets domain experts. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yingqiang Ge, Yujie Ren, Wenyue Hua, Shuyuan Xu, Juntao Tan, and Yongfeng Zhang. 2023b. Llm as os, agents as apps: Envisioning aios, agents and the aios-agent ecosystem. *arXiv e-prints*, pages arXiv–2312.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, et al. 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. 2024. A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

André Geraque Kiffer. 2019. *Battle Of Falkirk, July 22, 1298*. Clube de Autores.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2023. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. *arXiv preprint arXiv:2310.14985*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, et al. 2023a. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents. *arXiv preprint arXiv:2308.05960*.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023b. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*.

Toshifumi Matsuoka, Takahiro Hasegawa, Yasuhiro Yamada, Tetsuya Tamagawa, and Yuzuru Ashida. 2001. Computer simulation for sandbox experiments. In *SEG International Exposition and Annual Meeting*, pages SEG–2001. SEG.

Kai Mei, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. 2024. Llm agent operating system. *arXiv preprint arXiv:2403.16971*.

OpenAI. 2023. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. Self-alignment of large language models via multi-agent social simulation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Karthik Sreedhar and Lydia Chilton. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms. *arXiv preprint arXiv:2402.08189*.

Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2023. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024a. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024b. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*.

A More Experiment Result

A.1 Human analysis on location movement

Figure. 4 illustrates the general agent location dynamics of a single emulation of the Battle of Crécy using GPT-4. The English army is represented by red symbols, while the French army is represented by blue symbols. The sizes of the symbols are normalized to correspond to the number of soldiers contained in each agent. Different line types represent different types of agents.

At a glance, we can observe that as the emulation progresses, both armies are gradually split into smaller teams, especially the English army. In particular, some longbowmen tend to maintain a safe distance from the enemy for extended periods, using their longbows to inflict casualties from afar. As time progresses, the advantage of the French army's larger number of soldiers is diminishing over time, particularly in the case of the heavy cavalry and heavy knights. This is likely due to the effectiveness of the English longbowmen in inflicting casualties from a safe distance, as well as the challenging terrain of the battlefield, which made it difficult for the heavily armored French knights to maneuver effectively.

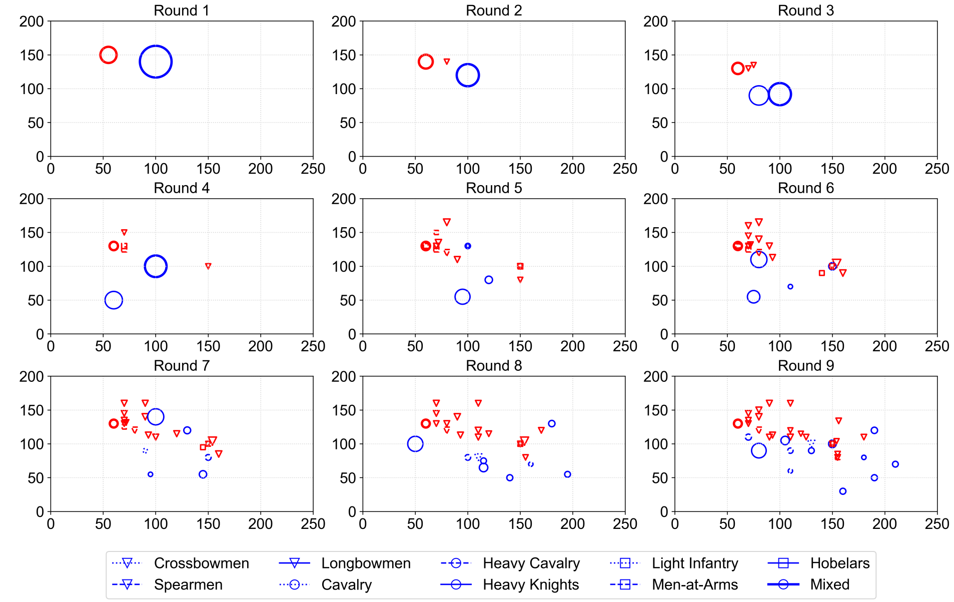

Figure 4: All agent movement and dynamic agent structure on battlefield.

To further evaluate the performance of the LLMs and VLMs in simulating historical battles, we can examine the paths taken by individual agents over time. This can provide insights into whether these models have a good sense of distance and can make reasonable decisions based on the overall environment.

A.2 Final Battle Casualty

Each of the four series of figures illustrates the time-series casualty data at each quantized time interval for the models Claude-3, GPT-4, and GPT-4-vision, presented from left to right. Within each image, the mean and standard deviation of casualties for both parties are displayed. Generally, it is evident that the Claude-3 model generates simulations resulting in significantly higher casualty figures compared to the other two models.

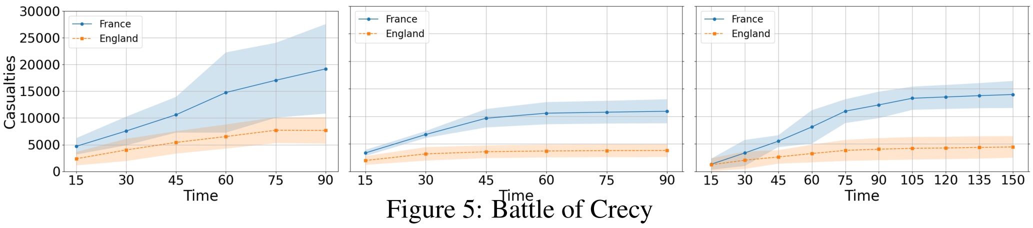

Figure 5: Battle of Crecy

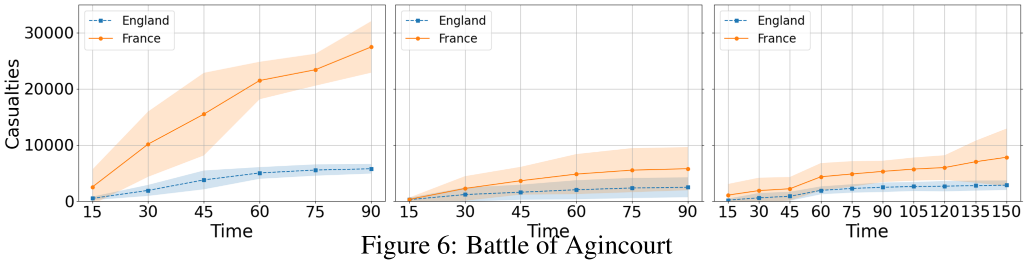

Figure 6: Battle of Agincourt

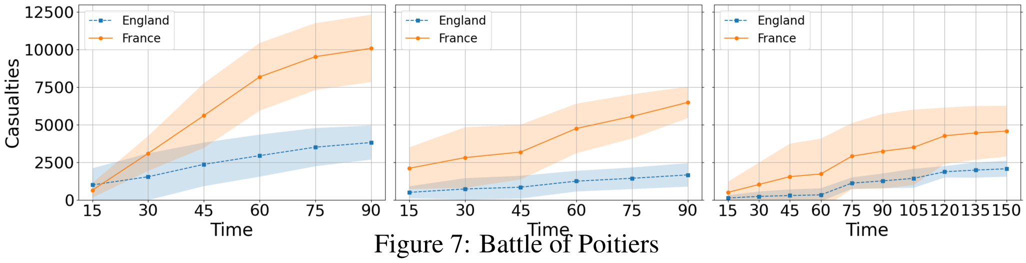

Figure 7: Battle of Poitiers

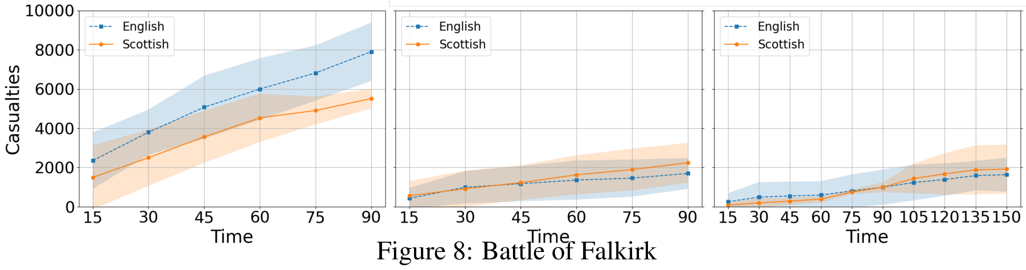

Figure 8: Battle of Falkirk

A.3 Human analysis on agent action

Figure 9 provides an illustrative example of the actions undertaken by two agents, one representing a part of the army belonging to England and the other representing a part of the army belonging to France, throughout the entire emulation time. The English agent's cautious approach is reflected in its movements and actions, while the French agent's aggressive strategy is evident in its frequent attacks and resulting losses. This example provides a reasonable representation of how historical battles may have unfolded.

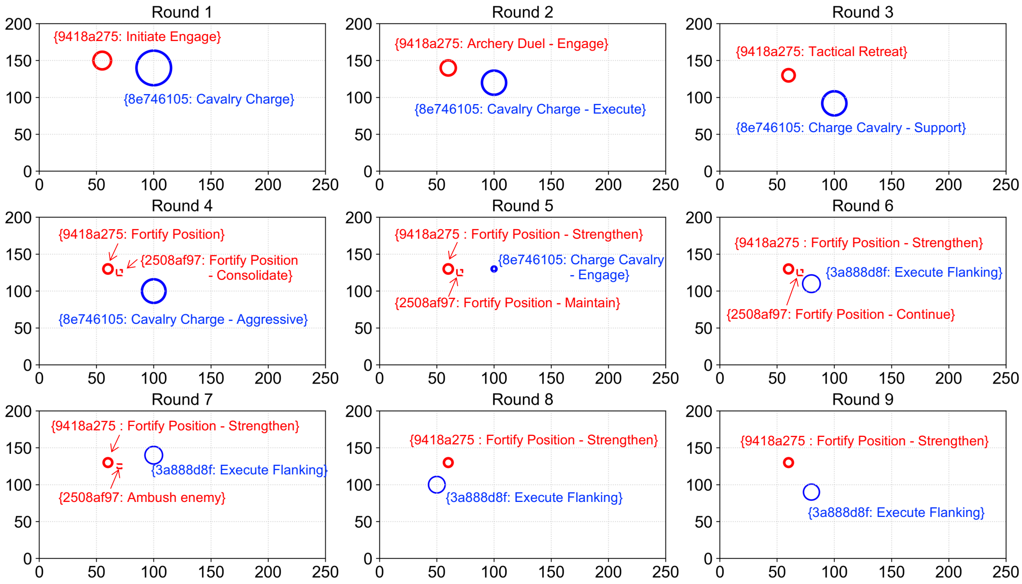

Figure 9: Agent action tracker over time.