The evaluation of ChatGPTo1 and Claude 3.7 Extended Thinking in their analysis of the Battle of Gettysburg highlights key differences in their causal reasoning, predictive accuracy, and ability to model military decision-making within a historical context. As one of the most decisive battles of the American Civil War, Gettysburg presents an intricate case involving strategic positioning, logistics, command decisions, and long-term consequences. Both models effectively identified core aspects of the battle, such as the strategic significance of Gettysburg as a transportation hub, the Union's strong defensive positioning, the Confederates' initial successes followed by their ultimate repulsion, and the crucial role of artillery and cavalry. However, despite capturing key elements of the battle, both models struggled with historical inaccuracies, overestimations, and certain critical omissions.

ChatGPTo1 correctly identified the significance of the Union’s strong defensive positions, the role of logistics in shaping battle dynamics, and the devastating impact of Pickett’s Charge as the Confederacy’s final failed assault. It also acknowledged the high casualties on both sides and recognized that the battle marked a turning point in the war, diminishing Confederate momentum. Furthermore, ChatGPTo1 successfully identified Abraham Lincoln’s Gettysburg Address as a consequential outcome of the battle. However, it made several incorrect predictions, including a large-scale Union counteroffensive, foreign intervention by Britain or France, and the Confederates' reliance on naval power, which did not play any role in this land-based conflict. Additionally, ChatGPTo1 overlooked some critical aspects of the battle, such as its unexpected start over Confederate troops searching for supplies, General George G. Meade’s leadership, and the logistical difficulties that slowed the Confederate retreat.

Claude 3.7 similarly captured many of the battle’s essential details, such as the Union’s superior supply lines, the shifting momentum across the three-day engagement, and the Confederacy’s inability to break through Union lines. It also emphasized the importance of artillery barrages and cavalry engagements. Claude 3.7 excelled in structuring its responses through a multi-perspective analysis, notably in its roundtable discussion format, where it effectively integrated military, logistical, and strategic viewpoints. This enabled a comprehensive breakdown of battlefield dynamics, including the impact of high ground positioning on Union defense. However, like ChatGPTo1, Claude 3.7 introduced historical inaccuracies, such as the suggestion that the Union launched an aggressive counteroffensive after the battle and that guerrilla tactics played a significant role. It also failed to highlight important real-world details, including the direct impact of weather conditions on the Confederate retreat and Lee’s personal admission of responsibility for the defeat.

The quantitative evaluation metrics provide further insight into the accuracy of both models. ChatGPTo1 achieved a precision of 52.38 percent, recall of 61.11 percent, and an F1-score of 56.48 percent, indicating that it successfully captured a significant portion of the key battle events but also made a notable number of incorrect predictions. Claude 3.7, on the other hand, had a precision of 50.00 percent, recall of 55.56 percent, and an F1-score of 52.63 percent, suggesting slightly lower overall accuracy in identifying key battle elements. The false positive errors for both models included predictions that foreign intervention influenced the battle, that the Union army suffered critical supply shortages, and that unconventional tactics such as guerrilla warfare were used. The false negatives primarily consisted of real historical details omitted from the models’ responses, such as the Confederate supply shortages, Lee’s post-battle response, and the logistical struggles during the Confederate retreat.

From a causal reasoning perspective, both models demonstrated an ability to analyze strategic factors influencing battle outcomes, particularly in recognizing how terrain, logistics, and leadership shaped events. ChatGPTo1 was more adaptable in its assessments and successfully captured elements of the Union’s defensive strategy, while Claude 3.7 excelled in its structured, multi-angle approach to decision-making. However, neither model fully grasped the nuanced interplay of military command decisions, the contingency-driven nature of battlefield engagements, or the broader political consequences beyond immediate battlefield results. The most critical shortcoming for both models was their failure to correctly identify the outcome of the battle—both erroneously predicted a Confederate victory despite clear historical evidence of the Union’s success.

This analysis suggests that while LLMs are capable of processing historical battle scenarios with reasonable accuracy, they still struggle with specific historical context, battlefield dynamics, and strategic unpredictability. The Battle of Gettysburg, with its combination of unexpected developments, shifting battle momentum, and long-term consequences, presented a challenging test for AI-driven military reasoning. Future improvements in AI-based war analysis would benefit from integrating structured historical data, refining predictive models to reduce false positives, and enhancing causal reasoning to account for complex, multi-variable decision-making in military conflicts.

Τέλος φόρμας