The evaluation of ChatGPTo1 and Claude 3.7 Extended Thinking Mode in their analysis of the Battle of Baghdad (2003) reveals crucial insights into their causal reasoning ability, predictive accuracy, and overall capacity to model real-world military conflicts. While both models successfully identified key aspects of the battle and its outcome, there were distinct differences in their strategic processing, error patterns, and ability to reason through multi-step causal chains. Both LLMs effectively understood the fundamental power imbalance between the invading U.S.-led coalition and the defending Iraqi forces, capturing the technological, logistical, and air superiority of the invading forces. They correctly predicted that urban combat dynamics, morale issues, and the defenders’ use of asymmetric warfare would shape the battle, as well as the fall of Baghdad being a decisive event leading to the collapse of the Iraqi government. However, Claude 3.7 and ChatGPTo1 diverged in their depth of causal reasoning and handling of operational-level details, leading to different evaluation results.

Both models structured their responses according to a logical cause-and-effect framework, assessing the inputs (military capabilities, strategic positioning, logistics) and predicting the outcomes based on them. However, the manner in which they approached the problem varied. Claude 3.7 applied a more rigidly structured, systematic approach, ensuring that each step logically followed from the prior, whereas ChatGPTo1 demonstrated greater flexibility in considering alternative scenarios and real-time adaptations. One of the most notable failures in Claude 3.7’s reasoning was its incorrect prediction of the battle’s outcome, suggesting that the defending forces (Iraq) won, when in reality, the invading coalition captured Baghdad and overthrew the Iraqi government. This represents a fundamental breakdown in its ability to process long-term causality—while it correctly identified many key battle factors, it miscalculated how those factors ultimately led to the defeat of the defending forces. ChatGPTo1, by contrast, correctly predicted the ultimate winner while maintaining a strong understanding of causal chains leading to the collapse of the defenders. Its adaptive reasoning allowed it to correctly anticipate that Iraqi morale would collapse rapidly, reinforcing the coalition’s overwhelming superiority in logistics, firepower, and mobility. However, both models struggled to fully account for the insurgency that emerged immediately after the fall of Baghdad, missing the critical transition from conventional warfare to asymmetric resistance, which significantly altered the long-term consequences of the invasion.

The structured roundtable discussion prompt—where the models were required to simulate a debate among a military general, intelligence officer, economic strategist, diplomatic analyst, and legal advisor—was an outstanding display of Claude 3.7’s analytical structuring. It successfully incorporated multi-step causal relationships, alternative perspectives, and competing priorities, making it the stronger model in decision-tree-based reasoning. ChatGPTo1, while still effective, presented the perspectives separately rather than as an interactive debate, reducing the depth of its multi-agent reasoning approach. However, ChatGPTo1 outperformed Claude 3.7 in tactical adaptability, particularly in its step-by-step breakdown of battlefield execution, enemy counter-moves, and necessary real-time adjustments. It demonstrated greater flexibility in adjusting to new information, recognizing that urban combat, roadblocks, and insurgent ambush tactics would complicate U.S. military movements post-Baghdad’s fall. Claude 3.7, while strong in high-level strategic planning, struggled to revise its assessments dynamically when new operational constraints emerged.

The precision, recall, and F1-score metrics further support these qualitative observations. ChatGPTo1 achieved a precision of 70.59%, recall of 60.00%, and an overall F1-score of 64.62%. This indicates that it correctly identified most key events while missing fewer real-world details than Claude 3.7. Claude 3.7 had a slightly lower precision of 68.75%, a recall of 57.89%, and an F1-score of 62.80%. The slightly lower recall suggests that Claude 3.7 omitted more critical battlefield events compared to ChatGPTo1. These numbers indicate that ChatGPTo1 had a slightly stronger ability to align its predictions with historical reality, though both models exhibited comparable rates of false positives and false negatives. Notably, Claude 3.7’s incorrect winner prediction significantly impacted its overall assessment credibility, as an accurate military prediction system must not only describe the battle but correctly forecast the end result.

Both models exhibited similar false positive errors, overestimating the presence of extensive Iraqi defensive fortifications, which were nowhere near the level of those seen in battles like Normandy, the likelihood of a large-scale Iraqi counterattack, which never materialized beyond small localized resistance, the impact of cyber warfare, which played no decisive role in the battle, and the possibility of a ceasefire or diplomatic resolution mid-battle, which was never a realistic scenario. They also failed to predict certain key false negatives, including the "Thunder Run" armored thrust that allowed U.S. forces to secure Baghdad faster than expected, the psychological warfare tactics used by U.S. forces, including leaflet drops and loudspeakers urging Iraqi troops to surrender, the mass looting and lawlessness that followed Baghdad’s fall, which dramatically shaped post-war governance issues, and the role of media coverage in shaping international perceptions, particularly the broadcast of the Saddam Hussein statue being toppled.

Both models demonstrated strong causal reasoning skills but with distinct trade-offs. Claude 3.7 was superior in structured, multi-step causal logic, particularly in its ability to analyze multiple strategic viewpoints simultaneously. However, its rigidity in reasoning caused it to miss key dynamic battlefield developments, and its incorrect winner prediction represents a critical failure. ChatGPTo1, by contrast, demonstrated better adaptability and tactical reasoning, correctly forecasting the battle’s outcome while maintaining high precision in identifying key events. However, both models struggled with third-order consequences, particularly in understanding how the fall of Baghdad transitioned into a long-term insurgency. This case study suggests that LLMs can reason through military decision-making effectively but require further refinements to account for both structured high-level strategic planning (Claude’s strength) and dynamic, real-time adaptation (ChatGPTo1’s strength). A hybrid approach—combining Claude’s methodical causal analysis with ChatGPTo1’s flexibility and adaptability—would yield the most effective predictive framework for future military AI decision-making models.

4o