The evaluation of ChatGPTo1 and Claude 3.7 Extended Thinking in their analysis of the Battle of Midway provides insight into their ability to perform causal reasoning, predict military decision-making, and assess historical conflicts. As one of the most decisive battles of the Pacific War during World War II, Midway serves as a crucial test case due to its reliance on intelligence, strategic deception, carrier-based warfare, and the decisive impact of a single engagement on the course of the war. Both models correctly identified the overall victor of the battle, recognizing that the defending forces, the United States, successfully repelled the Japanese offensive. However, while they captured many key elements of the battle, they also introduced historical inaccuracies, misrepresented certain aspects of the combat, and failed to fully integrate the intelligence and tactical missteps that led to Japan’s defeat.

ChatGPTo1 successfully identified the role of aircraft carriers, the impact of U.S. codebreaking in predicting the Japanese attack, and the strategic importance of Midway Atoll. It recognized that Japan began with numerical superiority, that the U.S. set a successful ambush, and that multiple Japanese carriers were destroyed, shifting the balance of power in the Pacific. However, it introduced several inaccuracies, including the incorrect prediction that Japan successfully landed troops on Midway, that a large-scale submarine engagement took place, and that Japan used radar defenses, which were not a significant factor in the battle. ChatGPTo1 also failed to highlight certain critical real-world details, such as the unsuccessful U.S. torpedo bomber attacks that nevertheless paved the way for the decisive dive bomber assault, the delays in Japanese scout plane reconnaissance, and the vulnerability of Japanese carriers while refueling and rearming aircraft.

Claude 3.7 similarly identified the carrier-based nature of the battle, the role of intelligence and deception, and the strategic importance of Midway in shifting the war’s trajectory. It was particularly effective in structuring its responses, breaking the battle into phases and analyzing different strategic perspectives. However, like ChatGPTo1, it made several incorrect assumptions, including an overestimation of Japanese resilience, an exaggerated role of ground combat, and the prediction of a temporary ceasefire, which did not occur. Additionally, it failed to recognize the critical moment when U.S. dive bombers struck the Japanese fleet at its most vulnerable point.

The quantitative evaluation metrics highlight the strengths and weaknesses of both models. ChatGPTo1 achieved a precision of 55.56 percent, recall of 52.63 percent, and an F1-score of 54.02 percent, while Claude 3.7 scored slightly lower, with a precision of 52.63 percent, recall of 50.00 percent, and an F1-score of 51.24 percent. These scores indicate that while both models successfully captured major aspects of the battle, they also made numerous errors and failed to fully integrate critical historical details. False positive errors included assumptions that Japan successfully landed troops, that submarines played a major role, and that there was a ceasefire. False negatives included missing key intelligence failures, the precise timing of U.S. airstrikes, and the Japanese inability to detect U.S. carriers before it was too late.

From a causal reasoning perspective, ChatGPTo1 was more adaptable in assessing intelligence, deception, and real-time battlefield conditions, while Claude 3.7 structured its analysis with a strong focus on breaking down strategic phases. However, neither model fully integrated all the key decision points that led to the U.S. victory. Their inability to recognize the decisive moment in the battle, when dive bombers struck Japanese carriers at a critical moment, suggests that LLMs struggle with understanding how intelligence failures, logistical constraints, and tactical shifts interact in real time.

This analysis underscores the challenges LLMs face in military decision-making, particularly when predicting historical battle outcomes and assessing battlefield dynamics. While both models identified many core elements of the Battle of Midway, their introduction of false events, their omission of key tactical details, and their overestimation of Japan’s ability to sustain the battle reveal significant gaps in their reasoning. Future improvements in AI-driven military analysis should focus on refining models’ abilities to assess intelligence failures, logistical constraints, and the unpredictability of combat environments. A hybrid approach that combines ChatGPTo1’s adaptability with Claude 3.7’s structured reasoning could lead to a more accurate framework for analyzing historical battles and strategic decision-making.