The evaluation of ChatGPTo1 and Claude 3.7 Extended Thinking in their analysis of the Battle of Stalingrad provides insight into their ability to perform causal reasoning, assess military decision-making, and predict battle outcomes. Stalingrad was one of the most pivotal battles of World War II, marked by brutal urban combat, a prolonged siege, and a decisive Soviet counteroffensive that resulted in a major turning point in the war. The battle's complexity, involving supply chain disruptions, environmental challenges, and strategic encirclement, presented a formidable test for both models in assessing military causality and predicting outcomes.

Both ChatGPTo1 and Claude 3.7 captured key aspects of the battle, such as the role of urban warfare, the challenges posed by winter conditions, and the ultimate destruction of the German Sixth Army following the Soviet counteroffensive. They successfully identified that the defending forces (the Soviet Union) leveraged urban combat tactics to neutralize the advantages of the attacking forces (Germany), that the Luftwaffe failed to supply the encircled German troops effectively, and that Soviet reinforcements helped sustain their counteroffensive. However, a major failure in both models’ reasoning was their incorrect prediction that Germany won the battle, a fundamental misinterpretation of historical reality.

ChatGPTo1 demonstrated strong analytical capabilities by correctly outlining the strategic difficulties Germany faced, including extended supply lines, harsh weather, and Hitler’s refusal to allow retreat. It also recognized that the Soviet Union used superior adaptability and manpower reserves to sustain the battle. However, the model introduced false positive errors, including the incorrect claim that the Soviet Union launched a preemptive attack (the battle actually began with a German offensive) and the prediction of large-scale tank engagements, when in reality, Stalingrad was primarily an infantry and artillery-driven urban conflict. ChatGPTo1 also missed significant real-world details (false negatives), such as the Luftwaffe’s intense initial bombardment that reduced Stalingrad to rubble, the role of Soviet women in combat and logistics, and the severe starvation and disease that plagued the trapped German forces.

Claude 3.7 similarly identified major elements of the battle but suffered from comparable errors. It correctly noted that Stalingrad was a turning point in the war, that the German supply lines were critically overextended, and that the Soviet counteroffensive was decisive. However, it introduced several historical inaccuracies, including an overstatement of German supply efforts, the incorrect assumption that a ceasefire was negotiated, and the misinterpretation of Soviet air superiority in the battle. Additionally, it failed to recognize the importance of Soviet deception tactics (maskirovka) and the long-term consequences of Axis prisoners dying in captivity.

The evaluation metrics highlight the limitations of both models. ChatGPTo1 achieved a precision of 55.56 percent, recall of 55.56 percent, and an F1-score of 55.56 percent, while Claude 3.7 performed slightly worse with a precision of 52.63 percent, recall of 50.00 percent, and an F1-score of 51.24 percent. These results indicate moderate accuracy in recognizing key battle events but also highlight significant errors in both factual correctness and causal interpretation. The most critical failure was the incorrect outcome prediction, which severely undermined both models' assessments.

From a causal reasoning perspective, ChatGPTo1 performed well in linking logistics and supply chain failures to battlefield outcomes but did not fully grasp the effects of Soviet deception strategies or the decisive moment when the German forces became irreversibly trapped. Claude 3.7 structured its analysis in a more segmented manner, effectively detailing phases of the battle, but it also failed to integrate the key tactical missteps that led to Germany’s downfall. Neither model successfully contextualized how the strategic blunders of the German high command, particularly Hitler’s insistence on holding Stalingrad at all costs, directly led to the catastrophic defeat.

This case study illustrates the challenges LLMs face in predicting military decision-making and assessing historical battles with full accuracy. While both models correctly identified many key events, they also introduced fictional elements, missed crucial historical details, and ultimately failed in their final assessment of the battle’s outcome. These errors underscore the need for improved integration of historical causality and tactical realism in LLM-based military analysis. Future refinements should focus on enhancing LLMs’ ability to process intelligence failures, leadership rigidity, and attritional warfare in complex battle scenarios. Combining ChatGPTo1’s adaptability with Claude 3.7’s structured approach may lead to a more reliable framework for evaluating historical conflicts and predicting military decision-making outcomes.