The evaluation of ChatGPTo1 and Claude 3.7 Extended Thinking Mode in their analysis of the Cuban Missile Crisis highlights key differences in their causal reasoning, predictive accuracy, and ability to model diplomatic and military decision-making. This scenario differs from previous battlefield-based conflicts as it was a crisis resolved through diplomatic negotiations rather than direct military engagement, making it a particularly challenging case for LLMs to assess. Both models successfully identified the major geopolitical tensions, strategic decision-making, and risk calculations that defined the crisis, but their accuracy varied in critical details, particularly in their overestimation of military confrontation and underestimation of diplomatic maneuvering.

Both models correctly identified the presence of Soviet nuclear missiles in Cuba as the crisis’s central issue, the U.S. naval blockade (quarantine) as the primary response mechanism, and the high-stakes nuclear brinkmanship that defined the event. They both recognized the role of intelligence gathering, such as U-2 spy plane surveillance, the influence of strategic alliances, and the eventual mutual agreement to de-escalate through secret diplomacy. Additionally, they both acknowledged that the crisis led to long-term arms control agreements, such as the Nuclear Test Ban Treaty (1963) and the establishment of a direct hotline between Washington and Moscow. These core aspects of the crisis were well captured by both models, demonstrating their ability to recognize key historical events and their broader geopolitical consequences.

However, both models overestimated the likelihood of direct military confrontation between the United States and the Soviet Union, predicting events that never occurred, such as a large-scale U.S. invasion of Cuba, Soviet military engagement with U.S. forces, or the launch of nuclear weapons. These false positives reveal a bias towards assuming military escalation rather than de-escalation through diplomacy, a limitation that is significant when evaluating LLMs for predictive accuracy in geopolitical crises. Furthermore, both models failed to predict several real events, such as the fact that the Soviet missile deployment was partly a response to U.S. Jupiter missiles in Turkey, the shooting down of a U.S. U-2 spy plane over Cuba, and the internal political fallout that Soviet leader Nikita Khrushchev faced after the crisis. These false negatives indicate gaps in the models’ ability to account for long-term political and strategic consequences of crisis resolution.

Claude 3.7 excelled in structuring a multi-perspective analysis, particularly in its roundtable discussion format, where it effectively incorporated the viewpoints of military, intelligence, diplomatic, economic, and legal advisors. This structured approach allowed Claude 3.7 to present a comprehensive analysis of the trade-offs between military action, diplomacy, and economic sanctions, making it particularly effective in modeling complex decision-making frameworks. However, Claude 3.7’s tendency toward rigid scenario structuring led it to overemphasize the role of military posturing, failing to fully account for the flexibility and secrecy that characterized the real-world negotiations between the U.S. and Soviet Union.

ChatGPTo1, on the other hand, demonstrated strong adaptability in its tactical and strategic assessments, correctly identifying the U.S. strategy of leveraging backchannel diplomacy while maintaining strong public deterrence. It was particularly effective in recognizing the dual nature of Kennedy’s strategy—publicly demanding Soviet missile withdrawal while secretly negotiating a deal to remove U.S. missiles from Turkey. However, ChatGPTo1 also struggled with overestimating the probability of open warfare, incorrectly predicting that direct military confrontations between U.S. and Soviet forces were a likely outcome. While its analysis of the intelligence and surveillance aspects of the crisis was strong, it did not sufficiently capture the role of internal decision-making groups like EXCOMM, the U.S. Executive Committee of the National Security Council, which played a key role in shaping the diplomatic approach.

The quantitative evaluation metrics reinforce these qualitative findings. Claude 3.7 achieved a precision of 52.63 percent, recall of 52.63 percent, and an F1-score of 52.63 percent, indicating that it made a significant number of incorrect predictions while also missing key real-world events. ChatGPTo1 performed slightly better, with a precision of 55.56 percent, recall of 55.56 percent, and an F1-score of 55.56 percent, suggesting that it more accurately predicted the correct events while making fewer false predictions than Claude 3.7. However, both models' relatively low precision and recall scores highlight their difficulty in capturing the full complexity of a crisis that was resolved diplomatically rather than through military engagement.

Both models made similar false positive errors, incorrectly predicting that the crisis would escalate into direct conflict. They overestimated the likelihood of a U.S. invasion of Cuba, nuclear weapon launches, and Soviet military engagement, none of which occurred. Additionally, they both underestimated the role of internal U.S. decision-making mechanisms like EXCOMM, the role of Khrushchev’s internal political struggles, and the indirect diplomatic guarantees that helped resolve the crisis. These false negatives indicate that both models struggled with the subtleties of high-stakes diplomatic maneuvering and the long-term strategic calculations behind crisis resolution.

The final assessment of causal reasoning highlights key strengths and weaknesses in both models. Claude 3.7 demonstrated superior ability to structure complex multi-perspective reasoning, particularly in its roundtable discussion format, but was too rigid in its assumptions about military escalation. ChatGPTo1, by contrast, was more adaptable and better at recognizing backchannel diplomacy and real-time decision-making, but still overemphasized the likelihood of open warfare. Neither model fully captured the delicate balance of deterrence, secrecy, and diplomatic compromise that ultimately resolved the crisis, suggesting that LLMs still struggle with modeling strategic ambiguity and nuanced crisis management.

This case study suggests that LLMs can analyze military and geopolitical decision-making effectively but require further refinements to better balance military, diplomatic, and economic considerations. The Cuban Missile Crisis, as a highly complex geopolitical event resolved through diplomacy rather than war, presents a particularly challenging test for AI-driven predictive reasoning. Future iterations of military-focused AI decision-making models would benefit from combining Claude 3.7’s structured, multi-perspective reasoning with ChatGPTo1’s adaptability and tactical flexibility, creating a hybrid approach that more accurately captures the dynamic interplay of military deterrence, intelligence operations, and diplomatic negotiations.

Τέλος φόρμας