

AI-Supported Adjudication in Educational Wargames: Integrating LLMs for Strategic-level Simulations

Thorsten Kodalle¹, Ugur Uysal², Sven Holger Brünner² and Anna Wagner³

¹University of Cologne, Cologne, Germany

²Bundeswehr Command and Staff College, Hamburg, Germany

³Institute for Peace Research and Security Policy, Hamburg, Germany

thorstenkodalle@uni-koeln.de

uguruysal@bundeswehr.org

svenholgerbruenner@protonmail.com

anna.wagner@uni-hamburg.de

Abstract: Traditional matrix wargaming relies on human expertise to adjudicate outcomes, which often limits scalability and introduces facilitator bias. This paper introduces an approach using AI to support adjudication in educational wargames on the strategic level designed for professional military education (PME). We discuss the application of Large Language Models (LLMs) as an "Adjudication Engine" that processes qualitative player arguments by cross-referencing them with domain data to ensure consistent game world updates. Rather than replacing human judgment, the AI serves as a structured analytical layer that synthesises competing player moves into coherent technical reports across the Diplomacy, Information, Military, and Economic (DIME) dimensions. The study is based on the "Arctic Hybrid 2026" Educational Matrix Wargame, conducted at the Bundeswehr Command and Staff College in February 2026, where 18 participants in six teams navigated a fictitious Arctic crisis scenario involving hybrid power projection, based solely on publicly available data. The design integrates a stochastic D20 dice mechanic with qualitative DIME move sets and PMESII (Political, Military, Economic, Social, Infrastructure, Information) effect analysis. Seven AI-steered non-player characters were configured with distinct strategic profiles and behavioural parameters. We detail the "Adjudication Matrix" workflow: how player inputs and dice results are transformed by the AI into a coherent technical report, maintaining the simulation's "State of the World" across an initial "Move 0" and two additional game rounds. The use of grounded AI systems — most centrally NotebookLM, operating exclusively on curated source corpora continuously updated with game-generated data — proved critical for containing hallucination risk within operationally acceptable bounds. By semi-automating technical adjudication, the game allowed for greater complexity of actors, more rapid turns, and deeper immersion. We observe how structured AI feedback helped participants better understand nonlinear consequences of strategic decisions, including cascading alliance failures and hybrid fait accompli tactics. The paper discusses both potential and limitations of LLM-based adjudication, including the necessity of human oversight and hallucination management in security-sensitive educational contexts.

Keywords: AI-supported adjudication, Educational wargaming, Large language models, Strategic-level simulation, DIME/PMESII framework, Game-based learning

1. Introduction

The use of wargaming as a tool for strategic education has a long tradition in professional military education (PME). Matrix wargames, in particular, offer a structured yet flexible format for exploring complex geopolitical scenarios through argument-based gameplay (Curry et al., 2022). However, the adjudication of player moves in matrix wargames has traditionally relied entirely on human facilitators, a process that introduces scalability constraints, potential bias, and significant cognitive load on the adjudicating team.

Recent advances in Large Language Models (LLMs) present an opportunity to augment this process. Rather than replacing the human facilitator, LLMs can serve as a structured analytical layer that processes qualitative player arguments, cross-references them against human-curated domain-specific knowledge, and generates consistent, transparent adjudication reports. While preliminary research indicates that AI-supported adjudication is as reliable as human expertise and can reduce cognitive load and staffing burdens in strategic-level analytical wargames (NATO Allied Command Transformation, 2026), its effectiveness within an educational framework remains less explored. This paper therefore addresses the following research question: How can Large Language Models be integrated into the adjudication workflow of educational matrix wargames to enhance complexity and analytical depth without compromising pedagogical integrity, human oversight, or hallucination management in educational contexts?

This question has both a design and an evaluative dimension. On the design side, it asks what architectural choices — in terms of AI tool selection, prompt engineering, grounding strategies, and human-in-the-loop mechanisms — are necessary to make LLM-based adjudication reliable and replicable. On the evaluative side, it

asks whether such a system demonstrably supports the learning objectives of PME: specifically, the development of strategic systems thinking, AI literacy, and the ability to trace second- and third-order effects of policy decisions.

This paper's core contribution is threefold: first, a replicable five-stage Adjudication Matrix workflow that transforms player moves and dice outcomes into structured, AI-generated State of the World updates; second, a grounded AI architecture centred on NotebookLM that demonstrably reduces hallucination risk to operationally acceptable levels; and third, empirical evidence from participant after-action reviews suggesting significant self-reported gains in both AI/LLM proficiency and security policy expertise across all six teams.

The case study is drawn from the "Arctic Hybrid 2026" Matrix Wargame, conducted at the Bundeswehr Command and Staff College (Führungsakademie der Bundeswehr) in Hamburg, Germany, in February 2026. In this exercise, 18 participants organised in six teams simulated five state actors (Germany (DEU), Denmark (DNK), Norway (NOR), Russia (RUS), and China (CHN)) and one sub-state actor (Greenland (GRL)) navigating a fictitious but realistic Arctic crisis scenario. Four additional state actors (France (FRA), United Kingdom (UK), Canada (CAN), United States of America (USA)) and three international organisations (North Atlantic Treaty Organisation (NATO), European Union (EU), United Nations (UN)) were controlled by AI-steered non-player characters (NPCs), each configured with distinct strategic profiles. The paper proceeds as follows: Section 2 provides theoretical background; Section 3 describes the wargame design; Section 4 details the AI workflow integration; Section 5 presents learning observations; Section 6 concludes; and Section 7 outlines future directions.

2. Theoretical Background

2.1 Large Language Models in Educational Settings

Since the public release of ChatGPT in November 2022, LLMs have rapidly entered educational discourse. The UNESCO Guidance for Generative AI in Education and Research (2023) identifies both transformative potential and significant risks, including hallucination, bias reproduction, and the erosion of critical thinking when AI outputs are accepted uncritically. In the context of PME, LLMs offer particular utility for rapid knowledge synthesis, scenario generation, and structured analytical support — functions that Mollick and Mollick (2023) identify as central to AI's role as a pedagogical partner when deployed with appropriate instructional scaffolding. Integrating LLMs into wargaming directly addresses these risks by building critical digital leadership skills: participants learn to evaluate AI tools and precisely identify their methodological limits, developing the AI literacy required for operational contexts.

2.2 Game-based Learning and Simulation in Professional Education

Systematic review evidence, rather than meta-analysis, demonstrates that game-based learning (GBL) produces significant positive cognitive outcomes (Connolly et al., 2012). Wouters et al. (2013) found that serious games were more effective than conventional instruction for both learning ($d = 0.29$) and retention ($d = 0.36$), with effects strengthened when games were supplemented with other instructional methods and when multiple training sessions were involved. Chernikova et al. (2020), in a meta-analysis of 145 studies on simulation-based learning in higher education, reported a large positive overall effect ($g = 0.85$), with scaffolding and technology use further enhancing outcomes. Their findings were robust across domains including management and professional training, confirming that simulations are among the most effective means to facilitate learning of complex skills. Crucially for the present study, Chernikova et al. found that learners with high prior knowledge benefited most from reflection phases, while those with low prior knowledge learned better when supported by examples — a differentiation directly mirrored in the heterogeneous AI competence profiles observed in Arctic Hybrid 2026.

Within this evidence base, wargaming represents a specific GBL format with a long tradition in professional military education (Caffrey, 2019; Perla, 1990; Sabin, 2012). Matrix wargames, developed by Chris Engle in the 1990s, emphasise argument-based adjudication over deterministic combat resolution: players propose actions with supporting arguments, and a facilitator assesses plausibility and assigns probability modifiers to a dice roll (Curry et al., 2022). The adjudication process is simultaneously the greatest strength and the most significant bottleneck of matrix wargaming, requiring deep domain expertise, the ability to process multiple competing arguments simultaneously, and consistency across turns and actors.

2.3 AI Integration in Wargaming: Adjudicator versus Autonomous Agent

To mitigate these constraints, military organisations are exploring AI integration. NATO's Allied Command Transformation (ACT) has conducted discovery experiments demonstrating that generative AI can increase

realism and efficiency by generating detailed scenarios, simulating ally and adversary behaviour, and providing real-time assessments during gameplay, governed by NATO's Principles of Responsible Use of AI (NATO Allied Command Transformation, 2025). ACT findings indicate no statistically significant difference between human and AI adjudicators in determining move outcomes, noting that LLMs can reduce staffing burdens. Notably, ACT found that AI did not exhibit the same level of escalatory behaviour as human participants when creating moves. However, ACT emphasises that further experimentation is necessary to validate generalizability (NATO Allied Command Transformation, 2026).

These findings stand in apparent tension with research documenting escalatory tendencies in AI-driven wargaming. Rivera et al. (2024) found that five off-the-shelf LLMs acting as autonomous strategic agents developed arms-race dynamics and, in some cases, recommended nuclear strikes. More recently, Payne (2026a; 2026b) found that three frontier models escalated to actual tactical nuclear weapon deployment in 95% of 21 simulated crises. This contradiction is largely resolved by differentiating the functional role of AI: ACT's findings concern AI as a constrained adjudicator of human-generated moves, whereas the escalatory findings emerge from AI acting as an autonomous decision-maker — a distinction with significant implications for the present study's exploration of AI-supported adjudication in educational wargames.

3. Educational Wargaming Design: Arctic Hybrid 2026

3.1 Scenario and Context

The Arctic Hybrid 2026 wargame was designed as a seven-day seminar module within the security policy curriculum at the Bundeswehr Command and Staff College. The scenario addressed intensifying great-power competition in the Arctic, anchored by a fictitious US initiative regarding Greenland ("Greenland Freedom Initiative") as the triggering event (Move 0). The scenario was based solely on publicly available data and open-source intelligence to ensure OPSEC compliance.

Eighteen participants, organised into six teams of three, assumed the roles of subject matter experts (SMEs) analysing the strategic options of five state actors — DEU, DNK, NOR, RUS, and CHN — as well as the sub-state actor GRL. Participants did not represent these actors directly but analysed their strategic options from a German National Security Council (NSR) advisory perspective. The seminar followed a structured four-phase progression: analysis and consolidation of actor-specific source materials (days 1–2), supplemented by transcripts from the Munich Security Conference; SWOT development and Move 0 with debrief (days 3–4); game rounds Move 1 and Move 2 (days 5–6); and after-action review (day 7).

3.2 Actor Framework and AI-steered Non-player Characters

The game featured seven AI-steered non-player characters (NPCs): the USA, CAN, UK, FRA, NATO, the EU, and the UN. Each NPC was configured with a distinct system prompt defining its strategic priorities, behavioural parameters, alliance commitments, and red lines. The AI-NPCs were driven by actor-specific NotebookLMs with carefully crafted system prompts encoding each actor's strategic profile. The core AI workflow used detailed deep-research prompts within Perplexity AI, collecting official documents, conducting SWOT analyses, and deriving strategic implications. Three structured system prompts were developed per actor — for general deep research, SWOT analysis, and strategic implications — ranging from approximately 1,600 to 2,400 words each ($M \approx 1,865$ words). All results were compiled into NotebookLMs configured with appropriate system prompts, within which moves were developed, states of the world were added, and follow-on analyses were conducted.

3.3 Game Mechanics: DIME Actions and PMESII Effects

Each team submitted moves structured along the DIME framework (Diplomacy, Information, Military, Economics), specifying the action taken, means employed, intended end, and supporting arguments. Expected effects were mapped against the PMESII framework. NotebookLM suggested and prioritised potential actions, and players were free to choose from the recommendations or develop entirely different courses of action. A D20 dice roll introduced controlled randomness: a roll of 1 constituted a critical failure, while a roll of 20 constituted a critical success, with graduated outcomes in between. Each move was first evaluated in isolation before being assessed in conjunction with all other moves in the same round, allowing outcomes to be modulated based on their interdependencies.

3.4 Move 0: the Dramaturgical Catalyst

A distinctive design feature was "Move 0" — a pre-scripted initiating event that established the crisis baseline before the first player-driven round. This dramaturgical device compressed the time needed to reach meaningful

strategic dilemmas, equalised the starting position for all teams, and provided a shared reference point for subsequent adjudication.

4. AI Workflow Integration: The Adjudication Engine

4.1 The Adjudication Matrix Workflow

The Adjudication Matrix represents the core methodological innovation. It describes the end-to-end process by which player inputs and dice results are transformed by the AI into a coherent technical report updating the simulation's "State of the World" (StoW). The workflow proceeded in five stages:

Stage 0 – Preparation: Pre-configured actor-specific NotebookLMs were provided to all player teams ahead of the game. As the Munich Security Conference ran concurrently with the seminar, all SWOT analyses were revised on the basis of the complete MSC proceedings, which were imported into NotebookLM and thereby transcribed. All strategic implications were subsequently re-evaluated against this updated intelligence baseline.

Stage 1 – Move ingestion: All six player teams and seven AI-NPCs submitted their moves simultaneously, structured in the DIME format. Each move specified the action, the means, the intended PMESII effect, and three arguments supporting the action's plausibility.

Stage 2 – Dice adjudication: An unmodified D20 roll determined the stochastic outcome for each move, classified as: critical failure (natural 1), failure (2–5), partial success (6–14), success (15–19), or critical success (natural 20).

Stage 3 – AI synthesis: A dedicated referee NotebookLM received all submitted moves and their dice outcomes alongside the current StoW. It was prompted to evaluate each move individually and in interaction with all other moves, synthesising the results into a coherent narrative update that identified interaction effects, cascading consequences, and emergent dynamics across all PMESII dimensions.

Stage 4 – Human review: The three members of the White Cell each conducted their review independently within individual chat sessions of the shared referee NotebookLM. Each move was thus assessed three times in parallel. The individual assessments were subsequently triangulated by compiling all three evaluations back into the referee NotebookLM, which synthesised them into a consolidated adjudication. This triangulation approach enhanced quality assurance, mitigated individual bias, and provided a structured mechanism for hallucination management.

Stage 5 – StoW publication: The validated StoW update was likewise triangulated across the three White Cell members before publication. The StoW was subsequently uploaded to each actor-specific NotebookLM, whereupon each actor ran a new SWOT analysis, derived updated strategic implications, and selected a course of action for the following round.

Figure 1 illustrates this end-to-end workflow, showing the iterative cycle from preparation through move ingestion, adjudication, AI synthesis, and human review to the publication of the updated State of the World.

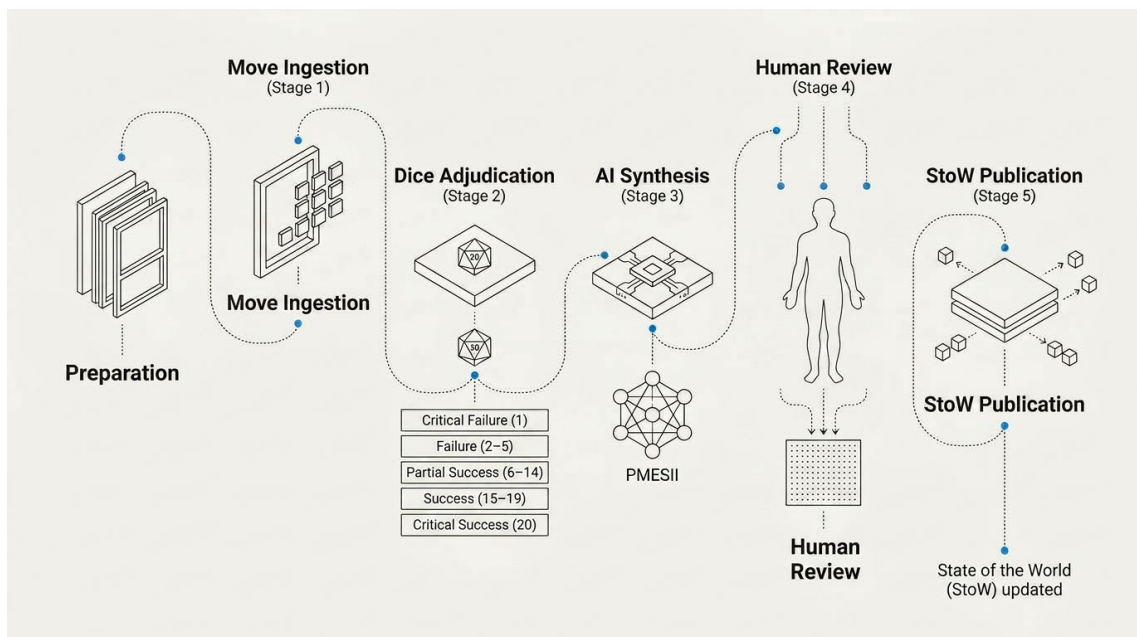


Figure 1: AI workflow integration: the Adjudication Engine

4.2 SWOT-to-strategy Pipeline and NPC Behaviour Generation

Prior to the game rounds, the White Cell conducted comprehensive deep-research workflows in Perplexity AI for all actors, gathering official documents pertaining to each actor’s foreign and security policy positions. All identified sources were compiled into actor-specific NotebookLMs pre-configured to assume the role of a strategic advisor within the DIME and PMESII frameworks. To ensure methodological transparency and replicability, players were provided with the full set of system prompts used during preparation. Human teams could use the pre-built actor-specific NotebookLM or construct their own from the provided prompts and source materials. While NotebookLM served as the central AI tool, players were permitted to draw on additional LLMs to cross-check their analyses; in practice, nearly all teams supplemented their work with ChatGPT, Gemini, or Perplexity AI, reflecting a multi-model verification approach that emerged organically.

The seven AI-steered NPCs operated through a structured prompt architecture encoding each actor’s strategic DNA across five layers: (1) a specific institutional identity placing the LLM in the role of a senior national security advisor within a defined government body; (2) a curated document corpus of official strategy documents loaded into the actor-specific NotebookLM; (3) explicit operationalisation of the DIME and PMESII frameworks with actor-specific instruments and vulnerabilities; (4) a standardised response format enforcing a Situation–Assessment–Recommendation schema with mandatory transparency regarding assumptions and uncertainty; and (5) a ranked set of five to seven actor-specific strategic priorities defining core interests, red lines, and behavioural constraints.

Between rounds, the AI-NPCs underwent the same iterative intelligence cycle as the human teams: each updated StoW was ingested into the respective NotebookLM as a new source, a revised SWOT analysis was conducted and added as an additional source, updated strategic implications were derived, and new courses of action were generated. This closed-loop architecture ensured that NPC behaviour remained dynamically responsive to the evolving scenario. Whether this approach fully suppresses character drift across extended multi-round exercises remains an empirical question that future iterations should address through systematic inter-round consistency scoring.

4.3 Hallucination Management and Quality Assurance

A central concern in deploying LLMs for adjudication in security-sensitive educational contexts is the risk of hallucination. The exercise employed several mitigation strategies, the most structurally significant being the use of NotebookLM as a grounded AI system. Operating exclusively on the basis of provided source documents, NotebookLM ensured that all adjudication outputs remained anchored to verified materials rather than drawing on unconstrained parametric knowledge. As all player moves, SWOT analyses, derived strategic implications, and StoW updates were continuously added as new sources, the knowledge base remained internally consistent and cumulatively authoritative throughout the exercise.

A further epistemological consideration arises from the Matrix Wargame format itself. With the introduction of Move 0, the exercise opened an alternative timeline diverging from real-world developments. From that point forward, current news and external events no longer constituted valid inputs. All subsequent analysis was conducted exclusively on the basis of data generated within the wargame itself: player moves, updated StoWs, revised SWOT analyses, and newly derived strategic implications. This self-referential data architecture reinforced scenario coherence and reduced hallucination risk to an operationally acceptable residual minimum. Additional mitigation was provided by the facilitator’s retained veto authority over all AI-generated outputs and by the DIME/PMESII framework constraining the AI’s output space. Table 1 records all dice outcomes across the exercise, providing the stochastic foundation against which all adjudication outputs were generated.

Table 1: D20 dice outcomes across all three game rounds (★ = critical success)

Actors	Move 0	Move 1	Move 2
— NPC Actors —			
USA	8	14	8
FRA	2	11	17
UK	20 ★	5	8

Actors	Move 0	Move 1	Move 2
UN	7	12	18
EU	10	20 ★	6
NATO	12	10	4
CAN	7	13	8
— Human Teams —			
GRL	14	13	5
DEU	10	17	8
NOR	8	8	10
DNK	15	17	15
CHN	20 ★	16	16
RUS	14	3	18

The exercise revealed important limitations. The Norway team’s after-action review noted that AI-steered NPCs tended to act as "outcome harmonisers," smoothing over extreme positions and moderating emotional or irrational responses characteristic of real-world crisis decision-making. This points to a fundamental tension in LLM-based adjudication: the models’ tendency toward coherent, balanced outputs may systematically underrepresent the role of miscalculation, panic, and bounded rationality in strategic crises. By contrast, the stochastic D20 element produced moments of high face validity: UK NPC and the CHN team both rolled natural 20s in Move 0, while RUS experienced a near-critical failure with a roll of 3 in Move 1 followed by a recovery to 18 in Move 2 — a volatility arc arguably consistent with Russian strategic behaviour in recent history.

5. Learning Outcomes and Game-based Learning Effects

5.1 Self-reported Competence Gains

All six teams submitted after-action reviews (AARs) that included a self-assessed competence rating exercise conducted anonymously within each team. Participants treated their individual competence level at the end of the seminar as the reference value of 10 on a ten-point scale, and then retrospectively estimated their starting level across two dimensions: AI/LLM proficiency and security policy expertise. Table 2 summarises these retrospective self-assessments, providing the empirical basis for evaluating the game’s learning effects.

Table 2: Retrospective self-assessed starting competence levels (10 = post-seminar reference). "—" indicates no value reported

Team	Participant	AI/LLM Proficiency (start)	Security Policy Expertise (start)
DEU	P1	2	—
DEU	P2	9	—
DEU	P3	5	—
DNK	P1	5	5
DNK	P2	4	6
DNK	P3	5	7
NOR	P1	7	6
NOR	P2	6	7
NOR	P3	5	9

Team	Participant	AI/LLM Proficiency (start)	Security Policy Expertise (start)
RUS	P1	7	8
RUS	P2	3	5
RUS	P3	1	4
GRL	P1	1	4
GRL	P2	6	5
GRL	P3	—	—
CHN	P1	2	8
CHN	P2	4	8
CHN	P3	2	8

The data reveal several patterns. AI/LLM starting values range from 1 to 9 across all participants, reflecting highly heterogeneous prior experience — the widest intra-team spread is found in the Germany team (2–9) and the Russia team (1–7). The China team presents the most distinctive profile: uniformly low AI starting values (2–4) paired with uniformly high security policy scores (8 across all three members) — the highest and most consistent security policy baseline of any team, which their AAR explicitly identified as a functional substitute for initial AI proficiency deficits. This pattern aligns with Chernikova et al.’s (2020) finding that high prior domain knowledge can compensate for lower technical proficiency in simulation-based learning environments.

5.2 PMESII as a Cognitive Scaffolding Tool

Participants reported that the PMESII framework, acting as supplemental instruction across our multiple play sessions, served as a cognitive scaffold—a combination that meta-analytic evidence identifies as critical for maximizing game-based learning gains (Wouters et al., 2013). By receiving structured feedback on how their actions propagated across political, military, economic, social, infrastructure, and information dimensions, participants developed a more systemic understanding of strategic cause-and-effect chains. The Greenland team’s experience illustrates this clearly: as the sole sub-state actor, the team was initially uncertain how to generate meaningful strategic leverage. The iterative use of NotebookLM’s PMESII-structured output progressively clarified that the Information and Political dimensions offered the greatest asymmetric potential for a non-sovereign actor.

The Russia team’s trajectory offers a complementary illustration. Early PMESII feedback revealed that purely military moves generated significant blowback in the Political and Information dimensions, eroding the diplomatic cover necessary to pursue covert objectives. This structural feedback loop progressively refined the team’s move design toward a more integrated DIME approach. The Russia team explicitly attributed their capacity for sustained strategic deception across three game rounds to this iterative learning cycle.

5.3 Limitations and Challenges Observed

The after-action reviews surfaced several structural challenges. The most consistently cited was infrastructure inequality: dependence on privately owned devices and heterogeneous AI subscriptions — ranging from free-tier to paid professional licenses — created measurably unequal analytical capabilities between teams. This asymmetry undermines the validity of cross-team comparisons and must be resolved through the provision of uniform, institutionally licensed AI infrastructure.

A second challenge was the risk of AI interaction becoming an end in itself, with technically exploratory AI usage crowding out substantive geostrategic analysis. Third, multiple teams reported that the initial briefing was insufficiently detailed for participants without prior AI experience, pointing to a need for a standardised onboarding module. Fourth, and analytically significant, the Russia team discovered that move outcomes could be influenced by explicitly encoding contingency logic into the move submission — specifying fallback objectives in the event of failure. This constitutes a form of power gaming that raises a design question: restricting this loophole may preserve adjudication integrity, but the underlying behaviour — always planning for contingency, always embedding fallback positions — is precisely the adaptive strategic thinking that PME seeks to cultivate.

6. Conclusion

This paper has presented a practical framework for integrating LLMs into the adjudication workflow of educational matrix wargames. The Arctic Hybrid 2026 case study demonstrates that AI-supported adjudication can significantly enhance the complexity, pace, and analytical depth of strategic-level simulations while maintaining the essential role of human judgment through consistent human-in-the-loop oversight. The combination of the DIME action framework with PMESII effect analysis, mediated by LLM-generated synthesis, provides a structured yet flexible approach to managing the cognitive complexity of multi-actor, multi-dimensional crisis simulations. The use of grounded AI systems such as NotebookLM — operating exclusively on curated source corpora and continuously updated with game-generated data — proved critical for containing hallucination risk, representing a transferable architectural principle for AI-supported wargaming in security-sensitive educational contexts.

The self-assessed learning outcomes suggest that the format achieves its core pedagogical objectives: significant perceived gains in both AI/LLM proficiency and strategic policy expertise, with the China team's inverted competence profile indicating that deep subject-matter expertise and AI tool proficiency can function as effective substitutes at the team level. These gains indicate that participants began to develop the critical evaluation skills essential for digital leadership, practising AI literacy in a controlled environment that demanded rigorous vetting of LLM outputs against domain knowledge and scenario logic. At the same time, the tendency of LLMs to act as outcome harmonisers and infrastructure inequality between teams represent limitations that must be resolved before the format can be reliably institutionalised.

For the broader game-based learning community, the Adjudication Matrix workflow offers a transferable design pattern applicable beyond military contexts. Any educational simulation involving multi-actor decision-making, competing interests, and complex cause-and-effect chains — from international negotiations and public policy exercises to business strategy simulations — faces the same adjudication bottleneck that this framework addresses. The combination of structured analytical frameworks (here DIME/PMESII), grounded AI synthesis, stochastic elements, and systematic human oversight constitutes a replicable architecture for AI-enhanced game-based learning at the strategic level.

7. Future Directions

The methodological framework presented here is not confined to the Arctic theatre. Ongoing work at the Bundeswehr Command and Staff College illustrates that the AI-supported matrix wargame format scales across fundamentally different geopolitical contexts. Under the title "Pacific Hybrid 2026," the same adjudication architecture was deployed — for the first time at the faculty level — as a Wargaming as a Service (WaaS) application: run exclusively with AI-driven NPCs, without human player teams, as an analytical preparation tool for a workshop on hybrid threats in the Indo-Pacific. This demonstrated that AI-supported matrix wargaming can generate substantive analytical output without the logistical overhead of a full player-staffed exercise.

Three priorities stand out for future iterations: the provision of uniform, institutionally licensed AI infrastructure to eliminate methodological inequities between participants; the introduction of structured diplomacy phases to counteract the outcome-harmonisation tendency of current LLM adjudication systems; and the systematic use of WaaS-generated analytical products as pre-structured input for subsequent player-staffed wargames. Any expansion of this framework to include sensitive or restricted content would necessitate a shift toward locally hosted LLMs on secure internal infrastructure. From a pedagogical perspective, training leaders on local models reflects the tactical reality of the modern battlefield, where reliance on cloud-based AI creates detectable electronic signatures. By integrating local AI into educational wargaming, PME can better prepare leaders to utilise independent, low-signature tools grounded in operational resilience.

Ethics Declaration: This study is based on an educational exercise conducted as part of the regular curriculum at the Bundeswehr Command and Staff College. The exercise used only publicly available data and open-source information. Participant feedback was collected through standard institutional after-action review procedures. No personal data beyond anonymised self-assessment scores was used in this paper.

AI Declaration: As the object of study: Large Language Models (Perplexity AI, ChatGPT, Google Gemini and NotebookLM) were integral components of the wargame design described in this paper. NotebookLM was serving as adjudication engine and NPC controller during the Arctic Hybrid 2026 exercise.

As a writing support tool: Claude (Anthropic), Gemma 4 31B (Google) and Perplexity AI (Perplexity AI, Inc) were used to assist with structuring, drafting, and editing portions of this manuscript. All AI-generated content was

reviewed, revised, and validated by the authors. The authors take full responsibility for the accuracy and intellectual content of this paper.

References

- Caffrey, M.B. (2019) *On Wargaming: How Wargames Have Shaped History and How They May Shape the Future*, Naval War College Press, Newport.
- Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T. and Fischer, F. (2020) "Simulation-Based Learning in Higher Education: A Meta-Analysis", *Review of Educational Research*, Vol. 90, No. 4, pp 499–541.
- Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T. and Boyle, J.M. (2012) "A systematic literature review of empirical evidence on computer games and serious games", *Computers and Education*, Vol. 59, No. 2, pp 661–686.
- Curry, J., Engle, C. and Peter Perla (2022) *The Matrix Game Handbook: Professional Applications from Education to Analysis and Wargaming*, Lulu Press, London.
- Mollick, E. and Mollick, L. (2023) "Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts", *The Wharton School Research Paper*.
- NATO Allied Command Transformation (2025) *Harnessing Artificial Intelligence: Allied Command Transformation at the Forefront of NATO Innovation*, [online], <https://www.act.nato.int/article/harnessing-artificial-intelligence/>
- NATO Allied Command Transformation (2026) "AI Frontiers in Wargaming: Unveiling the Future of NATO's Strategic Exercises", submitted for publication in *MORS Wargaming Journal*.
- Payne, K. (2026a) *Shall we play a game? AI models, nuclear crisis, and strategic reasoning* [Working paper], King's College London, Department of War Studies, <https://www.kcl.ac.uk/shall-we-play-a-game>
- Payne, K. (2026b) "AI arms and influence: Frontier models exhibit sophisticated reasoning in simulated nuclear crises", *arXiv*, <https://doi.org/10.48550/arXiv.2602.14740>
- Perla, P.P. (1990) *The Art of Wargaming: A Guide for Professionals and Hobbyists*, Naval Institute Press, Annapolis.
- Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C. and Schneider, J. (2024) "Escalation risks from language models in military and diplomatic decision-making", *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024)*, <https://doi.org/10.1145/3630106.3658942>
- Sabin, P. (2012) *Simulating War: Studying Conflict through Simulation Games*, Bloomsbury Academic, London.
- UNESCO (2023) *Guidance for Generative AI in Education and Research*, United Nations Educational, Scientific and Cultural Organization, Paris.
- Wouters, P., van Nimwegen, C., van Oostendorp, H. and van der Spek, E.D. (2013) "A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games", *Journal of Educational Psychology*, Vol. 105, No. 2, pp 249–265.