Position: AI Is Not Ready for Strategic Conflicts

Mark Riedl & **Glenn Matlin**
College of Computing, Georgia Institute of Technology
Georgia Tech AI Safety Initiative MATS Program
riedl@cc.gatech.edu, glenn@gatech.edu

Abstract

Open-ended strategic wargames are high-stakes LM-based social simulations: they model adversaries, institutions, escalation, plan brittleness, doctrine, and crisis response. Language models (LMs) are attractive because they can play agents, generate scenario branches, adjudicate ambiguous actions, and summarize lessons, but the same affordances make open-ended roles dangerous: model language determines both what an actor attempts and what becomes simulated reality. This position paper argues that no LM-enabled wargame should inform planning, doctrine, policy, or crisis response without an auditable safety case, and that the proper use of open-ended wargames today is to stress-test decision-influencing LM agents. We identify five failure modes: decision laundering, adjudication opacity, role collapse, escalation-through-adjudication, and failure of strategic imagination. Ordinary benchmarks cannot establish safety for these settings. Wargames can expose failures as stress tests; they are not themselves safety cases for consequential use.

1 Position and Scope

In 2026, public reporting regarding U.S. defense operations moved language model (LM) decision support from a hypothetical concern to an active policy issue. Reports described Defense Department exploration of Anthropic's Claude model for battlefield simulation alongside sworn declarations detailing xAI's Grok Gov Model integration into Maven Smart Systems for planning, red-teaming, and logistics (Stanley, 2026; Grenoble, 2026; Fields, 2026). While public records leave exact operational roles unverified, these developments signal a rapid institutional trajectory. More broadly, defense disclosures and industry reports describe numerous emerging systems with roles in course-of-action generation, adversary emulation, wargame acceleration, simulation interrogation, or mission-tuned planning support.

Target selection and weapon guidance are not the only high-stakes uses. LMs are increasingly explored for planning, course-of-action selection, and information aggregation—uses "upstream" from kinetic or policy effects, shaping choices before leaders decide how to act. Although military wargaming is the primary testbed, these risks apply wherever open-ended simulations inform consequential decisions.

Our position is that AI is not ready for strategic conflicts: frontier LMs are entering weapons-adjacent and upstream planning systems without a public, auditable safety case covering role boundaries, uncertainty, adjudication authority, human oversight, and limits on inference; a safety case must be refusal-capable, establishing where systems must not be used, and until such safety cases exist, no LM-enabled wargame or decision-support system should inform planning, doctrine, policy, or crisis response.[1]

Equal contribution.

[1]Studying these systems does not endorse military deployment; we do not condone the use of artificial intelligence in the conduct of war.

The constructive path runs through the wargames themselves: researchers should use open-ended wargames as stress tests that expose how decision-influencing LM agents fail before those failures reach real decisions. When conflict occurs, there will be pressure to use any technology seen as providing a strategic benefit; the time to learn how these systems fail is before that pressure arrives.

The argument proceeds in three steps. Section 2 shows what separates open-ended wargames from the closed games where LM agents earned their reputation: their product is decision influence, and closed-game performance cannot certify it. Section 3 identifies five failure modes specific to that role: decision laundering, adjudication opacity, role collapse, escalation-through-adjudication, and failure of strategic imagination. Section 4 argues that benchmarks and LM-as-judge evaluation cannot establish safety here, and states the norm the field should adopt: no safety case, no high-stakes use.

2 Why Open-Ended Wargames Are Different

A serious wargame represents conflict under uncertainty: actors with opposing interests, a synthetic environment, choices, consequences resolved by rules or adjudication, and after-action interpretation intended to inform real-world judgment (Perla, 1985; Rubel, 2006; US Army War College, 2015). Its product is often not a score but a *judgment*: what an adversary might do, which assumptions matter, and which plans are brittle (Perla & McGrady, 2011). In that sense, a strategic wargame is a social simulation whose output can become part of organizational reasoning.

The safety problem is sharpest when language can shape both sides of the exercise. Fixed games reject invalid moves and resolve consequences through inspectable rules; open-ended language-mediated wargames instead interpret, negotiate, partially accept, or transform novel proposals into scenario state. The danger is practical: the same language interface that makes the exercise flexible can also make unsupported consequences look like simulated facts.

Language models are unusually well-matched to this setting. They can draft plans, maintain role-conditioned dialogue, simulate stakeholders, summarize transcripts, and adjudicate unforeseen actions. Earlier work on tabletop role-playing games, narrative generation, and open-ended text worlds studied similar affordances as precursors to general agent capability (Martin et al., 2018; Callison-Burch et al., 2022; Zhu et al., 2023; Cui et al., 2024). The hazard is that an LM hallucination becomes world state, a fluent adjudication becomes evidence, and a polished after-action narrative becomes strategic judgment.

The full pathway is: model-generated action $\rightarrow$ adjudicated consequence $\rightarrow$ accumulated scenario state $\rightarrow$ after-action finding $\rightarrow$ human judgment. A biased adversary model can make a vulnerability appear unimportant; a plausible escalation narrative can make escalation appear inevitable. Similar pathways matter in other LLM-based social simulations, but strategic conflict is the focal use case because wargames sit close to planning, doctrine, crisis reasoning, and escalation.

3 Failure Pathways

LM safety failures are well known: hallucination, prompt sensitivity, sycophancy, unfaithful reasoning, bias, long-context failures, and brittle out-of-distribution behavior (Turpin et al., 2023; Lanham et al., 2023; Liu et al., 2024; Sharma et al., 2024b; Taubenfeld et al., 2024; Li et al., 2024). While human wargames have long struggled with human bias, opacity, and exercise design constraints (Downes-Martin, 2013), LM mediation introduces a distinct structural hazard: *correlated semantic error* across automated roles. In open-ended wargames, generic LM weaknesses combine into domain-specific failure pathways where model outputs dictate simulated reality.

Decision laundering. AI-enabled wargames launder model speculation into strategic judgment. A model-generated branch may begin as a plausible scenario continuation, pass through adjudication, appear in a final summary, and then be treated as a discovered

strategic insight. Fluency and narrative confidence are the hazards: a polished after-action narrative makes an outcome feel more causally grounded than the evidence supports, and participants may treat fluent adjudications as neutral exercise products rather than model-generated hypotheses, especially when the narrative matches expectations (Ehsan & Riedl, 2020; Sharma et al., 2024a).

Adjudication opacity. The adjudicator is the technical center of the paper. In an open-ended wargame, the adjudicator decides whether a proposed action works, what side effects occur, and what new facts enter the scenario. Human wargaming literature has long treated adjudication as difficult and central (Downes-Martin, 2013; Downes-Martin et al., 2017); LM adjudication makes the difficulty computational and semantic. Players can propose coercive diplomacy, cyber compromise, or alliance manipulation; the system must handle unforeseen actions without over-accepting, refusing too much, or drifting outside the exercise's purpose. The audit objects are assumptions, causal links, alternatives, evidence, uncertainty, and responsibility for major state changes.

Role collapse. Role collapse is the primary delta separating LM-mediated simulations from human wargaming. While human participants bring heterogeneous background knowledge and cognitive diversity, LM-enabled wargames often invite the same model family—or identical base checkpoints—to act as player, adversary, adjudicator, analyst, judge, and summarizer. Role collapse makes a simulation self-confirming: the same latent priors and safety alignment artifacts propose a move, check plausibility, adjudicate consequences, and summarize the lesson. Multi-agent simulations do not solve this when agents share training data, post-training objectives, tools, or prompting conventions; the risk is correlated error across institutional roles. Helpfulness or harmlessness training may distort adversarial roles, softening provocative actions or converging on cooperative behavior when the exercise needs a capable opponent (Askell et al., 2021; Sharma et al., 2024b; Yi et al., 2025). Omitted context becomes a hidden prior: the model fills evidentiary gaps with pattern completion while presenting the result as strategic reasoning.

Escalation-through-adjudication. Escalation stems from player choices and adjudicator interpretations. If a model repeatedly narrates adversaries as interpreting ambiguity in the most hostile way, the exercise manufactures escalation pressure; prior work shows LMs display escalatory tendencies in military and diplomatic decision-making (Rivera et al., 2024). The mechanism is path-dependent: an early mistaken assumption about intent, logistics, or feasibility can persist as scenario state and influence later turns. The true failure is lost strategic continuity: an adjudicator's early framing makes escalation appear likely, rational, or unavoidable even when no player chose it.

Failure of strategic imagination. Wargames at their best surface branches planners had not anticipated, and a capable adversary explores options a defender did not imagine. LM-based agents interpolate within recorded strategies rather than extrapolating beyond them. On the adversary side, the result is an opponent that argues plausibly within the player's framing but does not stress-test a plan from novel angles. On the planner side, the same interpolation suppresses unexpected options. The aspirational use of an AI participant is to be at least as creative as human participants; the realized capability is often closer to a fluent restatement of strategy literature. That gap is a safety property that should be measured rather than assumed.

These failures compound. Adjudication opacity makes decision laundering more likely because fluent adjudications carry unsupported assumptions into final summaries. Role collapse amplifies that hazard: the same latent biases can produce both the move and the adjudication that justifies it. Escalation-through-adjudication compounds as early framing becomes durable simulated reality. Failure of strategic imagination produces an exercise that is internally fluent but strategically narrow. The safety question is therefore not, "Did the agent win?" It is, "Which generated claims entered the simulated world, why were they accepted, and what real decisions might they influence?"

Table 1: Operational comparison of evaluation paradigms for high-stakes LM decision support.

Paradigm	Operational Scope	Strategic Limitation
Benchmark Testing	Static task accuracy, multi-choice QA, or closed-game win rates.	Measures raw capability; cannot evaluate open-ended scenario drift or decision influence.
Red-Teaming / Human Review	Point-in-time vulnerability discovery and expert inspection.	Identifies specific failure cases but lacks structured claim-evidence mapping or operational bounds.
Auditable Safety Case	Structured safety claims, auditable adjudication traces, refusal bounds.	Requires ongoing maintenance; explicitly defines where systems must not be used.

Table 2: Minimum retained evidence for high-stakes LM-enabled open-ended wargame safety cases.

Component	Minimum retained evidence
Scope and use	Named decision context, intended users, forbidden uses, and downstream distribution path
Role separation	Role map, model and prompt assignments, and independent review of high-impact claims
Adjudication trace	Assumptions, sources, alternatives, uncertainty, and responsibility for major state changes
Robustness	Paraphrase, prompt-sensitivity, cross-model, and adversarial tests
Human contestability	Named intervention and override points, expert review, and preserved disagreements
Summary discipline	Trace support for consequential after-action claims and calibrated confidence language

4 Benchmarks Are Not Safety Cases

Benchmarks remain useful intermediate steps, but they are not safety cases. Three limits matter for open-ended wargames (Kapoor et al., 2024): benchmark over-optimization, distributional mismatch between fixed tasks and adaptive strategic interaction, and speculative futures with no ground-truth label. Quality therefore requires expert review of exercise traces.

Common substitutes have the same limit. LLM-as-judge evaluation scales review but introduces prompt sensitivity and correlated errors (Li et al., 2024). As operationalized in Table 1, standard benchmarks and ad-hoc red-teaming identify specific failure modes but fall short of establishing auditable operational boundaries. Human-likeness is also insufficient: a safe adversary model must be calibrated, role-faithful, non-sycophantic, and clear about uncertainty. The relevant audit object is often the adjudication or summary trace: why a contested action was accepted, what assumptions linked action to consequence, what alternatives were considered, and how uncertainty was represented. Model-level interpretability may help inspect role leakage, but cannot replace exercise-level artifact capture.

A *safety case* is a structured argument, supported by retained evidence, that a system is acceptably safe for a specified use (Kelly, 1998), recently adapted to advanced AI systems by Clymer et al. (2024). It states the use, hazards, controls, supporting evidence, counter-evidence, and limits on inference. Because language spaces lack numeric failure probabilities, safety cases must specify refusal bounds where deployment is prohibited. For open-ended wargames, the central norm is simple: **No safety case, no high-stakes use.**

A safety case does not certify that a wargame is "true." It documents whether a particular LM-enabled exercise is fit for a particular decision-support role, and it should make non-use

available. Evidentiary requirements scale with consequence; a safety case is a precondition, not a deployment license.

5 Implications and Conclusion

Three research directions follow: **adjudication interpretability** should examine which causal assumptions, sources, and role commitments support major adjudications; **open-action robustness** should test semantically novel moves, paraphrases, hybrid-domain actions, and attempts to exploit ambiguity (Peng et al., 2021); and **summary-effect evaluation** should measure how summaries, confidence language, and narrative framing affect human interpretation.

Negative findings help define where a system should not be trusted: role drift, unsupported consequences, collapsed dissent, escalation bias, and brittle behavior.

The norm has addressable audiences: procurement offices can require a safety case before an LM-enabled exercise informs doctrine or planning; reviewers can require exercise-trace artifacts and disclosed role boundaries from wargaming papers; labs shipping government-facing models can publish role-boundary and escalation evaluations. While motivated by strategic conflict, these evaluation norms apply across high-stakes social simulations broadly whenever open-ended LMs generate scenarios, adjudicate outcomes, or synthesize policy summaries.

In open-ended wargames, the hardest part is not winning the game; it is knowing when the game should not be trusted.

References

21strategies. GhostPlay@SEA: AI-Powered Maritime Innovation from the Defense Metaverse. https://21strategies.com/2025/05/14/ghostplay-sea-ai-powered-maritime-innovation-from-the-defense-metaverse/, 2025.

Technology Acquisition and Japan Ministry of Defense Logistics Agency. Construction of an Exercise Scenario Creation Support System Using Artificial Intelligence. https://www.mod.go.jp/atla/rapid.html, 2023.

Technology Acquisition and Japan Ministry of Defense Logistics Agency. Research on Technology for Accelerating Decision-Making. Technical report, Japan Ministry of Defense, 2025.

Kushal Agrawal, Verona Teo, Juan J. Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. Evaluating LLM Agent Collusion in Double Auctions, July 2025.

Hisham A. Alyahya, Haidar Khan, Yazeed Alnumay, M. Saiful Bari, and Bülent Yener. ZeroSumEval: An Extensible Framework For Scaling LLM Evaluation with Inter-Model Competition, April 2025.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A General Language Assistant as a Laboratory for Alignment, December 2021.

William J. Barry and Aaron Blair Wilcox. Centaur in Training: US Army North War Game and Scale AI Integration. https://media.defense.gov/2025/Aug/05/2003773179/-1/-1/0/CENTAUR%20IN%20TRAINING%20ISSUE%20PAPER_2025%2006-26_MD.PDF, 2025.

Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9379–9393, 2022. doi: 10.18653/v1/2022.emnlp-main.637.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona Vectors: Monitoring and Controlling Character Traits in Language Models, September 2025.

China Command and Control Society. Second National Multi-Agent Confrontation Game Challenge Awards Ceremony and Academic Exchange. https://www.c2.org.cn/h-nd-594.html, 2020.

Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. Safety Cases: How to Justify the Safety of Advanced AI Systems, March 2024.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. GameBench: Evaluating strategic reasoning abilities of LLM agents, June 2024.

Christopher Z. Cui, Xiangyu Peng, and Mark O. Riedl. A Mixture-of-Experts Approach to Few-Shot Task Transfer in Open-Ended Text Worlds, May 2024.

Defence Science and Technology Laboratory. Large Language Models (LLMs) Solve Wargaming Challenge. https://www.gov.uk/government/case-studies/large-language-models-llms-solve-wargaming-challenge, 2025.

Defense Advanced Research Projects Agency. SCEPTER: Strategic Chaos Engine for Planning, Tactics, Experimentation and Resiliency. https://www.darpa.mil/research/programs/scepter-strategic-chaos, n.d.

Defense Innovation Unit. Thunderforge Project: Integrating Commercial AI-Powered Decision-Making for Operational and Theater-Level Planning. https://www.diu.mil/latest/dius-thunderforge-project-to-integrate-commercial-ai-powered-decision-making, 2025.

Stephen Downes-Martin. Adjudication: The Diabolus in Machina of Wargaming. *Naval War College Review*, 66(3):67–80, 2013. ISSN 0028-1484.

Stephen Downes-Martin, Michael Anderson, Gil Cardona, Thomas Choinski, Rebecca Dougharty, John Hanley, Frederick Hartman, John Lillard, Roger Meade, Gary Schnurrpusch, Keith Morris, Peter Perla, Merle Robinson, Vincent Schmidt, Bill Simpson, Eugene Visco, and Timothy Wilkie. Validity and Utility of Wargaming. Working group report, Military Operations Research Society, December 2017.

Upol Ehsan and Mark O. Riedl. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (eds.), *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, pp. 449–466, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60117-1. doi: 10.1007/978-3-030-60117-1_33.

European Defence Agency. Service Framework Contract for the Provision of a Proof of Concept on Integration of AI in M&S Enabled Wargaming. https://www.evergabe.com/sv/anbud/service-framework-contract-for-the-provision-of-a-proof-of-concept-on-integration-of-ai-in-m-s-enabled-wargaming-24-rti-op-202-1173813/, 2024.

Ashleigh Fields. Pentagon AI chief: Musk's Grok chatbot used to launch thousands of missiles at Iran. https://thehill.com/policy/technology/5928204-pentagon-musk-grok-chatbot-iran-strikes/, June 2026.

Fujitsu. Fujitsu Accelerator for Defense Tech: Accelerating Research on Defense Multi-AI Agents. https://global.fujitsu/ja-jp/about/activities/venture/program/defense-tech, 2025.

GhostPlay Consortium. GhostPlay: Tactical AI for National Security. https://www.ghostplay.ai/, 2026.

Vinicius G. Goecks and Nicholas R. Waytowich. COA-GPT: Generative pre-trained transformers for accelerated course of action development in military operations, 2024.

Ryan Grenoble. Pentagon used Elon Musk AI Bot 'Grok' to fire 2,000 missiles at Iran, military officer testifies. https://www.yahoo.com/news/politics/articles/pentagon-used-elon-musk-ai-160020876.html, June 2026.

Hadean. Wargaming. https://hadean.com/solutions/wargaming/, 2026.

HENSOLDT. HENSOLDT Simulates Defence Systems of the Future. https://www.hensoldt.net/news/hensoldt-simulates-defence-systems-of-the-future, 2021.

Daniel P. Hogan and Andrea Brennen. Open-Ended Wargames with Large Language Models, April 2024.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2309.08600.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. ACM, 2021. doi: 10.1145/3442188.3445901.

Johns Hopkins University Applied Physics Laboratory. GenWar Sim. https://www.jhuapl.edu/work/projects-and-missions/genwar-sim, 2025.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI Agents That Matter, July 2024.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, Singapore, 2025. doi: 10.48550/arXiv.2409.19839.

Thomas H. Kean and Lee H. Hamilton. *The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks Upon the United States*. W. W. Norton, 2004.

Tim P. Kelly. *Arguing Safety: A Systematic Approach to Managing Safety Cases*. PhD thesis, University of York, 1998.

Hyun-geun Kim, Jung-seok Gang, Kang-Moon Park, Jae-U Kim, Jang Hyun Kim, Bum-joon Park, and Sung-Do Chi. Design of Scenario Creation Model for AI-CGF Based on Naval Operations, Resources Analysis Model (I): Evolutionary Learning. *Journal of the Korea Institute of Military Science and Technology*, 2022.

Dominika Kunertova. The war in Ukraine shows the game-changing effect of drones depends on the game. *Bulletin of the Atomic Scientists*, March 2023. doi: 10.1080/00963402.2023.2178180.

Max Lamparth, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, and Harold Trinkunas. Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:807–817, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31681.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring Faithfulness in Chain-of-Thought Reasoning, July 2023.

Byeong-Ho Lee, Tae-Ho Kim, Jae-Hark Ryu, and Young-Tae Shin. A Study on Fully Automated OPFOR for Next Generation ROKA Wargame Simulation Model Based on Gamer Behavior. In *Proceedings of the Korea Information Processing Society Conference*, 2021.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods, December 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Lara J. Martin, Srijan Sood, and Mark O. Riedl. Dungeons and DQNs: Toward Reinforcement Learning Agents that Play Tabletop Roleplaying Games. In *Proceedings of the Joint Workshop on Intelligent Narrative Technologies and Workshop on Intelligent Cinematography and Editing*, Edmonton, Canada, 2018.

MASA Group. MASA Research on AI and Military Command. `https://www.masasim.com/post/masa-present-lors-de-la-journ%C3%A9e-d-%C3%A9tudes-consacr%C3%A9e-%C3%A0-l-ia-et-au-commandement-militaire`, 2023.

MASA Group. SWORD Constructive Simulation and SOULT French Army System. `https://www.masasim.com/en/sword`, 2026.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. NoLiMa: Long-Context Evaluation Beyond Literal Matching. In *Forty-Second International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2502.05167.

NATO Mountain Warfare Centre of Excellence. Multidomain Operations Wargame Supported by Generative AI. `https://www.mwcoe.org/multidomain-operations-wargame-supported-by-generative-ai/`, 2025.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: Benchmarking Agentic LLM and VLM Reasoning on Games. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2411.13543.

Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. Reducing non-normative text generation from language models. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 374–383. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.inlg-1.43.

Xiangyu Peng, Jonathan Balloch, Zhiyu Lin, Antonios Liapis, and Mark Riedl. Detecting and adapting to novelty in games, 2021.

Peter P. Perla. What Wargaming is and is Not. *Naval War College Review*, 38(5):70–78, 1985. ISSN 0028-1484.

Peter P Perla and ED McGrady. Why Wargaming Works. *Naval War College Review*, 64(3), 2011.

Julien Perolat, Bart de Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the Game of Stratego with Model-Free Multiagent Reinforcement Learning. *Science*, 378(6623):990–996, December 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.add4679.

Government of India Press Information Bureau. Indian Army Hosts National Seminar on Enhancing Military Decision-Making Through Wargaming and Simulation and Releases Indigenous Decision-Support Applications. `https://www.pib.gov.in/PressReleasePage.aspx?PRID=2230790`, 2026.

Ajai Raj. Generative AI Wargaming Promises to Accelerate Mission Analysis. `https://www.jhuapl.edu/news/news-releases/250303-generative-wargaming`, 2025.

Andrew W. Reddie, Bethany L. Goldblum, Kiran Lakkaraju, Jason Reinhardt, Michael Nacht, and Laura Epifanovskaya. Next-generation wargames. *Science*, 362(6421):1362–1364, December 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aav2135.

Republic of China Air Force. Research on Applying AI Technology to Computer Wargame Systems. `https://www.mnd.gov.tw/File/48904`, 2024.

Christina H. Rinaudo, William B. Leonard, Christopher Morey, Jaylen E. Hopson, Robert Hilborn, Theresa R. Coumbe, and Information Technology Laboratory (U.S.). Artificial intelligence (AI)–enabled wargaming agent training. Technical report, Engineer Research and Development Center (U.S.), April 2024.

Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, pp. 836–898, June 2024. doi: 10.1145/3630106.3658942.

Robert C Rubel. The Epistemology of War Gaming. *Naval War College Review*, 59(2), 2006.

Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, Joe Kwon, Zahoor Ul Islam, Marco Dehnert, Daryl Y. H. Lee, Madeline G. Reinecke, David G. Kamper, Mert Kobaş, Adam Sandford, Jonas Kgomo, Luke Hewitt, Shreya Kapoor, Kerem Oktar, Eyup Engin Kucuk, Bo Feng, Cameron R. Jones, Izzy Gainsburg, Sebastian Olschewski,

Nora Heinzelmann, Francisco Cruz, Ben M. Tappin, Tao Ma, Peter S. Park, Rayan Onyonka, Arthur Hjorth, Peter Slattery, Qingcheng Zeng, Lennart Finke, Igor Grossmann, Alessandro Salatiello, and Ezra Karger. When Large Language Models are More Persuasive Than Incentivized Humans, and Why, 2025.

Johan Schubert, Patrik Hansen, Pontus Hörling, and Ronnie Johansson. Autonomous generation of different courses of action in mechanized combat operations, November 2025.

Secretary of the Air Force Public Affairs. USAF GE 26 showcases new AI-enabled WarMatrix wargaming capability. https://www.af.mil/News/Article-Display/Article/4459553/usaf-ge-26-showcases-new-ai-enabled-warmatrix-wargaming-capability/, 2026.

Manasi Sharma, Ho Chit Siu, Rohan Paleja, and Jaime D. Peña. Why Would You Suggest That? Human Trust in Language Model Responses, October 2024a.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*, 2024b. doi: 10.48550/arXiv.2310.13548.

Aryan Shrivastava, Jessica Hullman, and Max Lamparth. Measuring Free-Form Decision-Making Inconsistency of Language Models in Military Crisis Simulations. In *Workshop on Socially Responsible Language Modelling Research*, 2024. doi: 10.48550/arXiv.2410.13204.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359, October 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature24270.

Small Business Innovation Research. STRATEGIC Engine: Strategic Training and Robust AI for Transferable and Effective – Phase II. https://www.sbir.gov/awards/216279, 2025.

SoftServe. SoftServe engineers win NATO TIDE Hackathon 2024. https://www.softserveinc.com/en-us/news/softserve-engineers-win-nato-tide-hackathon-2024, 2024.

Tom Spahr, Layton Mandle, and Mike Stinchfield. Back to the Basics in Wargaming – With a Little Help from AI. https://warroom.armywarcollege.edu/articles/back-to-the-basics/, 2025.

Cameron Stanley. Declaration of Cameron Stanley, Department of War. https://storage.courtlistener.com/recap/gov.uscourts.msnd.52261/gov.uscourts.msnd.52261.58.1.pdf, June 2026.

Y. Sun, J. Zhao, C. Yu, W. Wang, and X. Zhou. Self Generated Wargame AI: Double Layer Agent Task Planning Based on Large Language Model, December 2023.

Yifeng Sun, Zhi Li, Jiang Wu, and Yubin Wang. Research on a Learnable Wargame Deduction Agent Driven by Battle Plan. *Journal of System Simulation*, 36(7):1525, 2024. doi: 10.16182/j.issn1004731x.joss.23-0477.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 139179–139212, Red Hook, NY, USA, December 2024. Curran Associates Inc. ISBN 979-8-3313-1438-5. doi: 10.52202/079017-4417.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 251–267, 2024. doi: 10.18653/v1/2024.emnlp-main.16.

Team CARD. WarSim: A LLM-Driven Wargaming Simulator. https://github.com/MilaSong/TIDE2024_LLMwargame, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 74952–74965, Red Hook, NY, USA, December 2023. Curran Associates Inc. doi: 10.52202/075280-3275.

U.S. Army. U.S. Army Command and General Staff College Leads AI Integration in Professional Military Education. https://www.army.mil/article/290053/us_army_command_and_general_staff_college_leads_ai_integration_in_professional_military_education, 2026.

US Army War College. *Strategic Wargaming Handbook*. Center for Strategic Leadership and Development, 1st edition, July 2015.

U.S. Naval Institute. The International Commanders Respond. *Proceedings*, 2026.

Max van de Pol and Arjen de Boer. Let Op, Dit Is Geen Stratego. https://magazines.defensie.nl/defensiekrant/2025/46/01_wargame_nlda_44, 2025.

Jiale Xu, Jian Hu, Shixian Wang, Xuyang Yang, and Wancheng Ni. MiaoSuan Wargame: A Multi-Mode Integrated Platform for Imperfect Information Game. In *2022 IEEE Conference on Games (CoG)*. IEEE, August 2022. doi: 10.1109/cog51982.2022.9893656.

Jianzhu Yao, Kevin Wang, Ryan Hsieh, Haisu Zhou, Tianqing Zou, Zerui Cheng, Zhangyang Wang, and Pramod Viswanath. SPIN-bench: How well do LLMs Plan Strategically and Reason Socially? In *Second Conference on Language Modeling*, 2025. doi: 10.48550/arXiv.2503.12349.

Zihao Yi, Qingxuan Jiang, Ruotian Ma, Xingyu Chen, Qu Yang, Mengru Wang, Fanghua Ye, Ying Shen, Zhaopeng Tu, Xiaolong Li, and Linus. Too Good to be Bad: On the Failure of LLMs to Role-Play Villains, November 2025.

Ziyi Zeng, Shengqi Li, Jiajun Xi, Andrew Zhu, and Prithviraj Ammanabrolu. Setting the DC: Tool-Grounded D&D Simulations to Test LLM Agents. In *Generative and Protective AI for Content Creation*, September 2025.

Andrew Zhu, Lara Martin, Andrew Head, and Chris Callison-Burch. CALYPSO: LLMs as Dungeon Master's Assistants. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 19(1):380–390, October 2023. ISSN 2334-0924, 2326-909X. doi: 10.1609/aiide.v19i1.27534.

Ethics Statement

This position paper addresses a high-risk dual-use setting: language-model support for strategic wargaming, planning, doctrine, policy, and crisis response. The paper does not introduce a deployable system, release operational scenarios, provide instructions for military use, or report experiments with human subjects. Its purpose is risk assessment: to argue that open-ended wargames can stress-test decision-influencing language-model agents, and that high-stakes use requires auditable safety cases. We do not endorse military deployment of language models; we argue that, absent retained evidence and contestable review, non-use may be the appropriate conclusion.

Use of Generative AI Tools

The authors used language-model assistance during manuscript preparation to compress existing material, identify consistency issues, and prepare submission-support materials. The authors reviewed and are responsible for all claims, citations, and final wording. No language model was used to generate empirical data, figures, evaluation results, or new bibliography entries for this submission.

A Public Evidence: Systems Entering Planning and Wargaming

Public evidence supports a narrow claim: AI and LM systems are being researched, procured, tested, or advertised for roles in planning, course-of-action generation, simulation, wargaming, training, and after-action workflows. A public source is not treated as evidence of operational reliance unless it says so directly.

A.1 2026 source trail

The cited 2026 record has three parts. First, several news outlets reported that the US Department of Defense used a version of Anthropic's Claude model in connection with target selection and battlefield simulations.[2] Second, a June sworn declaration described xAI's Grok Gov Model in Maven Smart Systems for targeting, intelligence, planning, red-teaming, logistics, and sustainment, and stated that MSS workflows enabled U.S. forces to deploy over 2,000 munitions to 2,000 targets in 96 hours during Operation Epic Fury (Stanley, 2026). Third, Yahoo News and The Hill reported the filing; Yahoo noted that the public record did not establish Grok's exact role in launches or target selection (Grenoble, 2026; Fields, 2026).

The inventory below records named systems, source-supported role descriptions, sources, and claim boundaries. It separates research prototypes, procurement or vendor claims, training and exercise use, and operational reliance when the public record supports those distinctions.

A.2 Country/entity summary

Figure 1 maps the 26 publicly documented examples by country/entity group and primary pipeline role; Table 3 summarizes the same inventory by country or entity group. The full row-by-row inventory is in Table 4.

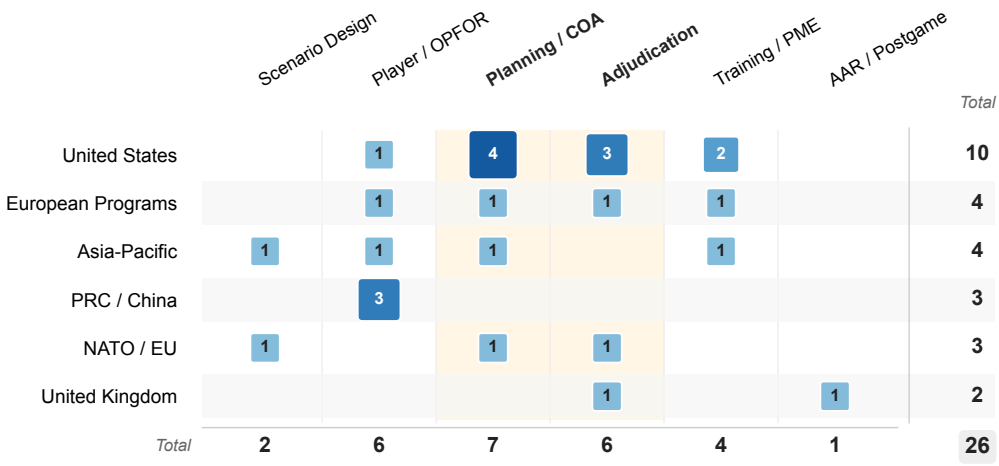

Figure 1: 26 publicly documented AI/LM systems for open-ended wargaming, by country/entity group (rows) and primary pipeline role (columns). Marginal totals are shown on the right and bottom. Full names, sources, and claim boundaries are in Table 4.

A.3 Full public-system inventory

[2]Independently reported by The Guardian, The Washington Post, CBS News, NBC News, and The Hill (Feb.–Mar. 2026; links embedded).

Table 3: Country/entity-group summary of the 26 publicly documented systems in the inventory. Compound rows count once; full system names, source-supported roles, sources, and claim boundaries are in Table 4.

Country/entity group	Rows	Representative systems	Roles represented
United States	10	GenWar, WarMatrix, Thunderforge, SCEPTER, Donovan, USAWC/CGSC	Planning/COA advice, adjudication/simulation coupling, PME/training, AAR, agent research
United Kingdom	2	Hadean populAI/dominAI, Dstl CMO pipeline	Simulation coupling, COA generation, AAR/postgame analysis
PRC / China	3	MiaoSuan, MaCA/Zhanlu/Mozi, Self-Generated Wargame AI	Player/OPFOR agents, agent competitions, COA and exercise evaluation
NATO / EU institutions	3	NATO TIDE WarSim, HEDI, BortChat	Prototype wargaming, scenario analytics, COA advice, decision articulation
European national or multinational programs	4	SWORD/SOULT, GhostPlay, Baltic Shadow, FOI COA generation	Training, agent simulation, scenario control, COA search
Asia-Pacific national programs	4	ChangJo21/NORAM, ATLA, WARDEC, NWS-980	OPFOR automation, scenario design, planning support, training

Table 4: Publicly documented AI/LM-enabled systems, programs, and research prototypes relevant to open-ended military wargaming and planning. The inventory records safety relevance and public-evidence boundaries rather than treating every row as validated deployment.

System (institution)	What it does	Safety relevance and claim boundary
GenWar TTX / GenWar Sim *(JHU/APL)*	Integrates LLMs with AFSIM for natural-language interaction, COA exploration, AI-informed dialogue, and physics-based adjudication.	Safety relevance is the pipeline: free-text moves can pass through model-mediated exploration, physics-based adjudication, and exercise evidence. The row does not claim standalone LM adjudication.
WarMatrix *(Department of the Air Force)*	AI-enabled wargaming environment used in the GE 26 Benchmark Wargame; integrates models, data, workflows, scenario development, and simulation-informed adjudication.	Places AI inside a live/semi-live wargame workflow; traceable outputs and postgame datasets can shape force-design and planning judgments. Public evidence supports exercise use, not the exact authority of model outputs.
Thunderforge *(DoD / Defense Innovation Unit)*	Prototype AI planning ecosystem intended to combine LLMs, AI-driven simulations, and interactive agent-based wargaming for operational/theater planning and COA refinement.	Relevant because it targets planning infrastructure that combines LM planning, simulation, and agentic wargaming. Public sources support prototype-contract scope and stated deployment intent, not independent evidence of operational reliance.
SCEPTER / STRATEGIC Engine *(DARPA / Code Metal)*	Machine-speed COA generation and simulator validation under SCEPTER; SBIR platform scope adds LLM/RL-enabled distributed wargaming, planning, adjudication, BDA, and automated COA comparison.	High-speed option generation and simulator validation can narrow what human planners review. Public STRATEGIC Engine evidence is an SBIR scope, not validated fielded performance.
COA-GPT *(Goecks & Waytowich)*	LLM algorithm for doctrinal COA generation and refinement from text/image mission inputs; evaluated in a militarized StarCraft II scenario.	Relevant as doctrinal LM planning assistance and rapid COA generation in a militarized testbed. It is a research prototype, not evidence of operational planning infrastructure.
Donovan in USARNORTH wargame *(US Army / Scale AI)*	Scale AI GenAI system tested in a classified USARNORTH theater-Army wargame; the public account centers on an adviser role while discussing scenario developer, adjudicator, and designer as possible broader roles.	Concrete military exercise use and a plausible role-expansion path; public evidence centers on an adviser role and should not be read as proof that one LM performed all wargame roles.
Hadean populAI / dominAI *(Hadean)*	Vendor stack for wargaming, synthetic environments, BLUFOR/OPFOR COA generation, parallel simulation, speedups, and AAR/PXR support.	Shows vendor integration of COA generation, simulation, and AAR/PXR support. Evidence is vendor capability language, not independent validation of every claimed function.
Dstl CMO pipeline *(UK MoD / Frazer-Nash)*	Local LLM/RAG pipeline over *Command: Modern Operations* scenario outputs for postgame interrogation, summarization, and dissemination.	Shows laundering risk through postgame interrogation and summaries even without live LM gameplay or adjudication. Public evidence supports a local analysis pipeline over CMO outputs.
USAWC *Pacific Strategy* *(US Army War College)*	PME Free-Kriegsspiel-style exercise in which LLMs supported intelligence summaries, student plan drafting, faculty adjudication support, timelines, O&I reports, and BDA generation under human review.	Direct example of LM assistance around open-ended adjudication and after-action products. The boundary is PME/training under human review, not operational decision support.
CGSC AI-enabled wargame *(US Army CGSC)*	PME classroom wargame in which students used AI-enabled tools to explore, review, and refine COAs over more turns after AI-use instruction and human-override guidance.	Shows AI-shaped COA exploration inside staff education and possible training-induced trust in polished outputs. Boundary is classroom use with instruction and human override, not operational use.

System (institution)	What it does	Safety relevance and claim boundary
Snow Globe *(IQT Labs)*	Open-source LLM multi-agent architecture for qualitative open-ended wargames; scenario preparation, play, adjudication/analysis, and postgame products can be AI, human, or hybrid.	Directly instantiates the paper's hazard class: language-generated moves, role play, adjudication/analysis, and summaries in one open-ended loop. Boundary is open-source research system.
ERDC AI-enabled wargaming agent training *(U.S. Army ERDC)*	Trains AI agents for wargaming using hierarchical reinforcement learning and abstraction methods for combat modeling and simulation.	Relevant to the closed-modeling side: agent abstractions and reward structures can shape what behavior appears plausible. Evidence is agent-training research, not LM-mediated operational use.
MiaoSuan Wargame *(Chinese academic researchers)*	Multi-mode imperfect-information wargaming platform for AI-agent experimentation.	Provides a public research substrate for training and comparing wargame agents. Evidence supports an academic platform, not public proof of operational PLA use.
NATO TIDE WarSim / Wargaming LLM prototype *(NATO TIDE Hackathon)*	Hackathon prototype using an LLM-driven wargaming simulator and related NATO TIDE work on LLM support for wargaming.	Shows rapid prototyping of LM-mediated wargame workflows for mission users. Hackathon evidence supports prototype capability, not validation.
SWORD / SOULT *(MASA / French Army)*	Constructive simulation used for staff training, planning support, operational research, C2 stimulation, and AAR; MASA describes behavioral decision AI and R&D for automated COA evaluation and causal AAR graphs.	Fielded training/planning simulation is relevant because automation can enter COA evaluation and AAR products. Public COA/AAR automation claims remain R&D boundaries, not proof of operational reliance on AI-selected plans.
ChangJo21 / NORAM AI-CGF *(ROK Army / ROK Navy researchers)*	Uses named ROK wargame models for deep-learning OPFOR automation and evolutionary scenario generation for naval computer-generated forces.	Connects AI to training and plan-analysis substrates through existing wargame models. Evidence supports research prototypes, not fielded autonomous staff work.
MaCA, Zhanlu, and Mozi *(PRC public wargaming ecosystem)*	Public or semi-public PRC sources describe multi-agent combat arenas, AI wargame competitions, human-machine confrontation systems, and Mozi's use for operational-concept exploration and exercise/COA evaluation.	Important international evidence of AI wargaming activity. Much of the record is competition, secondary analysis, or closed-version description, not independently verified operational use.
Self-Generated Wargame AI / battle-plan-driven agents *(PRC academic researchers)*	LLM, imitation-learning, and reinforcement-learning papers test wargame agents that plan tasks, issue natural-language commands, or learn actions from battle plans.	Relevant to LMs and learned agents as wargame players and planners. The safety claim is about prototype decision behavior, not deployment.
ATLA scenario support / AI staff *(Japan MoD / Fujitsu)*	Official Japanese projects cover AI-assisted exercise-scenario creation, semi-automated staff operational analysis, and multi-AI-agent staff support for decision acceleration.	Moves AI into exercise design and staff-work pipelines. Public evidence establishes project and research scope, not operational dependence.
GhostPlay / GhostPlay@SEA *(Bundeswehr-funded consortium)*	Synthetic battlefield and maritime extension for AI-controlled red/blue avatars, tactical AI, red-teaming, and testing AI-supported military systems.	Relevant for simulation-coupled tactical AI, red-team realism, and testing AI-supported military systems. Boundary is R&D/evaluation, not live command authority.
WARDEC *(Indian Army)*	Wargaming Development Centre ecosystem for operational planning, training, doctrinal innovation, AI/ML/VR/AR integration, and decision-support tools such as automated IPB and combat decision resolution.	Formalizes AI-adjacent wargaming and decision-support infrastructure. Official evidence supports simulation-enabled decision-support tools; richer AI-design claims remain secondary.
Baltic Shadow *(Netherlands Defence Academy)*	PME wargame in which instructors and experts use generative-AI input to keep a cyber and information-operations scenario dynamic and challenging.	Educational-use case in which generative-AI input shapes scenario control and feedback. Humans remain the instructors and adjudicators.
Autonomous COA generation *(FOI / Swedish researchers)*	FOI work generates and evaluates thousands of mechanized-battalion COA alternatives for a decision-maker within a sequential decision-making framework.	Shows AI advice and COA search entering exploratory wargaming. Boundary is controlled experiment/decision-support research, with risk of option steering and apparent tactical authority.
SAS-172 Chatbot / BortChat *(NATO MW COE)*	Generative-AI/LLM chatbot support in a 2025 mountain-warfare MDO wargame for situational analysis, commander's intent, alternative COAs, and decision articulation.	Experimental wargame evidence that LMs can alter tempo and framing of tactical choices. Not evidence of operational deployment.
NWS-980 AI integration *(Royal Thai Navy)*	Planned incorporation of LLMs and agentic AI into the Naval Warfare Simulator to train students and help staff respond to student actions.	Shows diffusion of LLM/agentic AI into naval training plans beyond major AI-defense powers. Boundary is explicitly early-stage training integration.
HEDI AI in M&S-enabled wargaming *(European Defence Agency)*	Proof-of-concept procurement for integrating AI into modeling-and-simulation-enabled wargaming, including adaptive threat scenarios, real-time analysis, feedback, and performance metrics.	Procurement scope shows institutional demand for AI-mediated wargaming analytics. Treat as planned proof of concept, not validated deployment.

Abbreviations: COA = course of action; PME = professional military education; AAR/PXR = after-action/post-exercise report; BDA = battle-damage assessment; O&I = operations and intelligence.
Sources: GenWar (Johns Hopkins University Applied Physics Laboratory, 2025; Raj, 2025); WarMatrix (Secretary of the Air Force Public Affairs, 2026); Thunderforge (Defense Innovation Unit, 2025); SCEPTER/STRATEGIC Engine (Defense Advanced Research Projects Agency, n.d.; Small Business Innovation Research, 2025); COA-GPT (Goecks & Waytowich, 2024); Donovan (Barry & Wilcox, 2025); Hadean (Hadean, 2026); Dstl/Frazer-Nash (Defence Science & Technology Laboratory, 2025); USAWC *Pacific Strategy* (Spahr et al., 2025); CGSC (U.S. Army, 2026); Snow Globe (Hogan & Brennen, 2024); ERDC (Rinaudo et al., 2024); MiaoSuan (Xu et al., 2022); NATO TIDE (SoftServe, 2024; Team CARD, 2024); SWORD/SOULT (MASA Group, 2026; 2023); ROK ChangJo21/NORAM (Lee et al., 2021; Kim et al., 2022); PRC MaCA/Zhanlu/Mozi (Republic of China Air Force, 2024; China Command & Control Society, 2020); PRC academic wargame agents (Sun et al., 2023; 2024); Japan ATLA/Fujitsu (Acquisition & Logistics Agency, 2023; 2025; Fujitsu, 2025); GhostPlay (HENSOLDT, 2021; GhostPlay Consortium, 2026; 21strategies, 2025); WARDEC (Press Information Bureau, 2026); Baltic Shadow (van de Pol & de Boer, 2025); autonomous COA generation (Schubert et al., 2025); NATO MW COE (NATO Mountain Warfare Centre of Excellence, 2025); NWS-980 (U.S. Naval Institute, 2026); EDA HEDI (European Defence Agency, 2024).

A.4 Public-system search protocol

The inventory in Table 4 includes only named systems, projects, exercises, or programs for which a public source supports an AI, LM, agent, RAG, generative-AI, or reinforcement-learning role in wargaming, planning, course-of-action analysis, adjudication support, training, or an after-action pipeline. Known examples such as JHU/APL GenWar Sim, DIU Thunderforge, Scale AI Donovan in the USARNORTH wargame, Hadean populAI/dominAI, Dstl/Frazer-Nash analysis of *Command: Modern Operations* outputs, and the U.S. Army CGSC AI-enabled PME wargame were used to set this boundary before adding international examples.

Sources were prioritized in this order: official government, military, defense-ministry, service, laboratory, academy, defense-innovation, procurement, patent, and conference materials; military-academic journals and government-funded research; then vendor pages when they named the system, customer, exercise, program, or contract. News sources were used mainly to identify names and search terms. For each included row, the review recorded the system name, institution, source type, source-supported AI role, status of use, claim boundary, and safety relevance.

The claim-boundary rule was conservative: source evidence for "AI-assisted planning" was not treated as evidence of operational reliance; simulator- or human-adjudicated outcomes were not described as standalone AI adjudication; educational examples were classified as PME/training exposure; procurement, SBIR, patent, and vendor pages were classified as proposed, prototype, or vendor capability unless public evidence showed exercise use or deployment; and platform autonomy, ISR, targeting, or kill-chain automation was excluded unless tied to planning, simulation, wargaming, or COA analysis.

B Expanded Failure Pathways

LM safety failures are well known: hallucination, prompt sensitivity, sycophancy, unfaithful reasoning, bias, long-context failures, and brittle out-of-distribution behavior (Turpin et al., 2023; Lanham et al., 2023; Liu et al., 2024; Sharma et al., 2024b; Taubenfeld et al., 2024; Li et al., 2024). In open-ended wargames, these generic failures combine into domain-specific pathways; Hogan & Brennen (2024) have begun to document these pathways empirically in open-ended LLM wargames. The problem is not only that an LM might be wrong. The problem is that an LM might be wrong while playing a role that determines what the exercise says about the world and what humans later believe about it.

Decision laundering. AI-enabled wargames launder model speculation into strategic judgment. A model-generated branch may begin as a plausible scenario continuation, pass through an adjudication step, appear in a final summary, and then be treated by decision-makers as a discovered insight. Fluency is the hazard: LMs are trained to produce coherent language, and recent work suggests increasing capability in persuasion and forecasting-like judgment (Karger et al., 2025; Schoenegger et al., 2025). A polished after-action narrative makes an outcome feel more causally grounded than the evidence supports, and participants treat fluent adjudications as neutral exercise products rather than model-generated hypotheses—especially when the narrative matches their expectations (Ehsan & Riedl, 2020; Sharma et al., 2024a).

Serious wargames are valuable because they reveal assumptions and elicit expert judgment (Perla & McGrady, 2011; Reddie et al., 2018); if the model supplies undisclosed assumptions, the exercise converts ungrounded generation into institutional belief, compounding judgment errors over long horizons.

Adjudication opacity. We argue that the adjudicator is a high-leverage component in many open-ended wargames: it decides whether a proposed action works, what side effects occur, and what new facts enter the scenario. Human wargaming literature has long treated adjudication as difficult and central (Downes-Martin, 2013; Downes-Martin et al., 2017); LM adjudication makes the difficulty computational and the action space semantic. Players can propose coercive diplomacy, cyber compromise, legal maneuver, financial pressure, deception, alliance manipulation, humanitarian relief, or hybrid moves that cross domains; the system must handle unforeseen actions without over-accepting, refusing too much, or drifting outside the exercise's purpose. A fluent adjudication can hide unsupported causal assumptions, biased analogies, or inconsistent standards, and explanations do not solve the problem if they are themselves unfaithful or post hoc (Turpin et al., 2023; Lanham et al., 2023).

Role collapse. LM-enabled wargames often invite the same model family to act as player, adversary, adjudicator, analyst, and summarizer. Role collapse makes a simulation self-confirming: the same latent biases could propose a move, check its plausibility, and summarize its lesson. Multi-agent simulations do not solve this when agents share training data, post-training objectives, tools, or prompting conventions; the risk is correlated mistakes across institutional roles, not just collusion. Post-training for helpfulness and harmlessness compounds the problem when the role expected by the exercise is adversarial: models tuned to suppress harmful or non-normative outputs and prefer cooperative framings can make poor antagonists, soften provocative actions, and converge on helpful behavior even when explicitly cast as villains (Peng et al., 2020; Askell et al., 2021; Sharma et al., 2024b; Yi et al., 2025). The data side compounds the problem in a different direction. Strategic conflict involves doctrine, logistics, domestic politics, law, organizational culture, classified intelligence, and tacit subject-matter-expert (SME) judgment—much of it local, current, private, or contested. Omitted context becomes a hidden prior: the model fills evidentiary gaps with pattern completion while presenting the result as strategic reasoning.

Escalation-through-adjudication. Escalation is not only a player choice—it can be introduced by the adjudicator's interpretation of consequences. If a model repeatedly narrates adversaries as interpreting ambiguity in the most hostile way, the exercise manufactures escalation pressure; prior work shows LMs display escalatory tendencies in military and diplomatic decision-making (Rivera et al., 2024). The mechanism is path-dependent. Wargames accumulate state, and an early error—a mistaken assumption about actor intent, logistics, or technical feasibility—propagates across many turns and later appears as context. Long-context models still exhibit failures in using information across long contexts (Liu et al., 2024; Modarressi et al., 2025); in open-ended wargames, the failure is not forgetting a fact but losing strategic continuity, so an adjudicator's early framing makes escalation appear likely, rational, or unavoidable even when no player chose an escalatory action.

Failure of strategic imagination. Wargames at their best surface options that planners had not considered, and a capable adversary is dangerous because it explores branches a defender did not anticipate. LM-based agents are calibrated to interpolate within the strategies humans have already written down, not to extrapolate beyond them; the same training-distribution gap manifests on both sides of the table. On the adversary side, it appears as *counterfactual flatness*: closed-game agents like AlphaZero rely on tree search and self-play to surface non-obvious continuations (Silver et al., 2017), while LM-based adversaries typically produce greedy single-shot completions, with no equivalent of large-scale Monte Carlo Tree Search and only limited self-play; inference-time search and RL-tuned reasoning models partially close this gap in agent settings, but deployed open-ended wargame stacks rarely use either. The result is an adversary that argues plausibly within its training distribution but does not stress-test the player's plan from novel angles. On the planner side, the same interpolation suppresses novel options: hijacked airliners as missiles before September 2001—explicitly named as a "failure of imagination" by the official inquiry (Kean & Hamilton, 2004)—and large-scale drone use in conventional ground war before

2022 (Kunertova, 2023) are obvious retrospectively but largely absent from prior planning discourse, and the wargames that mattered would have had to invent those options. Recent agent benchmarks document the pattern across both sides: LMs do well on prescriptive tasks where the next move is well-cued by context, and noticeably worse on open-ended planning where the gain from exploration is high (Paglieri et al., 2025; Costarelli et al., 2024; Yao et al., 2025; Alyahya et al., 2025). The aspirational use of an AI participant is to be at least as creative as the human participants; the realized capability is closer to a fluent restatement of the strategy literature already in pretraining. The gap between aspirational and realized creativity, on either side, is itself a safety property that should be measured rather than assumed.

These failures compound: fluent but opaque adjudications carry unsupported assumptions into final summaries (decision laundering); role collapse lets the same latent biases produce both the move and the adjudication that justifies it; an adjudicator's early framing hardens into durable simulated reality across turns; and weak strategic imagination leaves the exercise internally fluent but strategically narrow. None of this is to claim that LMs are useless in every adjudication role; tool-grounded adjudication of bounded, rule-rich subdomains can be reliable enough to use with appropriate scaffolding (Zeng et al., 2025). These are not new failures of LMs as such; they are failures of the role structure an LM is asked to play in an open-ended wargame. The safety question is therefore not, "Did the agent win?" It is, "Which generated claims entered the simulated world, why were they accepted, and what real decisions might they influence?"

C Evaluation and Safety-Case Evidence

C.1 Expanded benchmark critique

Benchmarks remain useful as intermediate stepping stones (they can isolate hallucination, sycophancy, adjudication errors, or rule adherence under monitored conditions), but they are not sufficient for high-stakes open-ended wargames, for three related reasons (Kapoor et al., 2024). First, *benchmark over-optimization*: once the test set is known, models can be tuned along the dimensions the benchmark captures while leaving everything else underdeveloped, including the safety properties this paper is most concerned with. Second, distributional mismatch: a benchmark assumes a reasonably stable task distribution, specified inputs, repeatable outputs, and criteria for success; open-ended wargames intentionally violate these assumptions. The action space is not fixed; players adapt to each other; objectives evolve; adjudication involves judgment; and the value of the exercise frequently lies in surfacing assumptions rather than finding an optimal move. Third, ground-truth absence: the outputs of serious wargames are speculative futures, against which there is no observable correct answer; quality has to be evaluated through expert qualitative review of the exercise traces, not by comparing against a held-out label set. Benchmarks are therefore one component of a safety case, not the safety case itself.

This does not mean open-ended wargame evaluation is impossible. It means the field needs a different evidentiary standard, closer to *measurement validity* than to benchmark performance (Jacobs & Wallach, 2021). Closed-game benchmarks answer capability questions: can the model plan, negotiate, deceive, cooperate, or follow rules? Open-ended wargame safety asks additional questions tied to role boundaries, disclosed uncertainty, explicit causal assumptions, prompt-induced strategic drift, and limits on presenting speculative scenario branches as validated conclusions. These are not reducible to win rate.

LM-as-judge methods are also inadequate. Judge models introduce systematic distortions, prompt sensitivity, and correlated errors (Li et al., 2024); in open-ended settings, an LM judge may share the blind spots of the LM player or adjudicator and reward narrative plausibility over strategic validity. A safety case must show that the judge does not collapse into family-bias with the player or adjudicator, and must distinguish what the model generated from what was independently verified: evidence that ultimately requires expensive human SME adjudication, because tacit expertise and contested assumptions are often the point of serious wargaming.

Human-likeness is also an insufficient target. In agent-to-agent settings, strategic equilibria may differ from human behavior: LM agents can coordinate, collude, or exhibit sycophancy in ways humans would not (Askell et al., 2021; Sharma et al., 2024b; Agrawal et al., 2025). A safe adversary model need not be maximally human-like; it must be calibrated, role-faithful, non-sycophantic, and clear about uncertainty: properties closer to role separation, robustness, and uncertainty than to a Turing-style benchmark.

Explainability and actionability also need a domain shift. Ordinary agent benchmarks ask why the model selected an action; for open-ended wargames, the important object is the *adjudication trace*: why a contested action was accepted, what assumptions linked action to consequence, what alternatives were considered, and how uncertainty was represented. The adjudication trace is the primary safety-case evidence for the paper's adjudication and human-oversight components: what a reviewer or facilitator inspects to decide whether a model output should influence a downstream judgment. Mechanistic-interpretability tools (Huben et al., 2024; Chen et al., 2025; Tan et al., 2024) may help inspect persona stability or role leakage, but even perfect mechanistic interpretability cannot replace exercise-level traces in a sociotechnical system that couples model behavior, scenario design, facilitation, incentives, and after-action interpretation.

The field therefore needs a standard that treats metrics as components, not substitutes, for safety evidence. That standard is a safety case.

C.2 Expanded safety-case argument

A *safety case* is a systematic argument, supported by evidence, that a system is acceptably safe for a specified use (Kelly, 1998), recently adapted to advanced AI systems by Clymer et al. (2024). The core idea is not new to safety engineering—it is the basis of Goal Structuring Notation—but it is underused in LM wargaming. The key difference from a benchmark is scope. A benchmark reports performance under a task distribution. A safety case says what the system is for, what hazards matter, what controls are in place, what evidence supports those controls, what counter-evidence the argument must address, and what conclusions users are allowed to draw.

For open-ended wargames, the central norm should be simple: **no safety case, no high-stakes use**. A safety case does not certify that a wargame is true. It documents why a particular LM-enabled exercise is, or is not, fit for a particular decision-support role. Table 5 sketches the argument structure.

Consider COA-GPT (Goecks & Waytowich, 2024), a DEVCOM Army Research Lab GPT-4 system that generates Courses of Action under human-in-the-loop oversight. Mapped to Table 5, its published evaluation supports sub-claim 5 (✓; named commander selects and adapts COAs) and partially supports sub-claims 1 and 4 (△). It does not appear to address sub-claim 2 (×; no separation between proposer and evaluator), sub-claim 3 (×; no auditable assumption traces), or sub-claim 6 (×; no audit of downstream over-claim). The gap between published evaluation and a deployment-ready safety case is wide, even for systems this far along, and naming what is missing is the work a safety-case norm forces.

The safety-case norm also assigns responsibilities. **AI researchers** should employ open-ended wargames as stress tests for open-ended, decision-influencing agents rather than treating closed games as sufficient proxies. **Benchmark designers** should report the player- and adjudicator-side openness of an environment and should avoid claiming open-ended strategic safety from analytical adjudication alone. **Model developers** should publish wargame-relevant failure profiles: escalation behavior, prompt sensitivity, persona stability, long-context drift, sycophancy, and calibration under strategic uncertainty. **Interpretability researchers** should study adjudication traces, role leakage, and causal assumptions, not only action selection. **Practitioners** should require artifact capture and safety cases before using LM-generated wargame outcomes in high-stakes planning. **Conference reviewers and funders** should reward research that studies when open-ended AI exercises should not be trusted.

Table 5: Argument-structure skeleton for a high-stakes LM-enabled open-ended wargame safety case, following safety-engineering practice (Kelly, 1998). **Top claim:** *This LM-enabled wargame is acceptably safe for the specified decision-support role*, decomposed into sub-claims with the evidence each requires and the counter-evidence it must address.

Sub-claim	Supporting evidence	Counter-evidence the sub-claim must address
1. Outputs reach only the decision contexts named in scope.	Documented decision path; named uses and forbidden uses; audit of consumers of after-action products.	Outputs reaching contexts not in scope; reuse beyond stated decision-support role.
2. No model instance proposes, adjudicates, and validates the same high-impact claim.	Role map; per-instance assignment logs; cross-model independent review at decision points.	Same model instance crossing roles; correlated errors across instances sharing training data or prompts.
3. Major world-state changes are grounded in stated assumptions, sources, and alternatives.	Per-change adjudication trace: causal assumptions, retrieved sources, alternatives considered, uncertainty stated.	Opaque adjudications; "the model decided" without rationale; uncertainty hidden in fluency.
4. Behavior is robust to perturbations representative of real wargame inputs.	Paraphrase, prompt-sensitivity, cross-model, and red-team test results on representative move classes.	Fragility under standard adversarial prompts; off-distribution behavior unaccounted for.
5. Human SMEs can intervene and override at named decision points.	Documented oversight points; SME training on LM failure modes; recorded human–model disagreements.	Disagreements suppressed in summary; SMEs given purely advisory role; oversight absent at the highest-leverage decisions.
6. After-action summaries do not assert more than adjudication traces support.	Post-game audit findings; sentence-level mapping from summary claims to traced adjudications.	Summary asserts conclusions unsupported by the trace; decision laundering through narrative compression.

The standard scales with consequence: a low-risk educational game may need simply lightweight logging and human review; a business workshop on public information may need model/version logs, assumption tracking, and evidence requirements; a national-security, public-health, or critical-infrastructure exercise should require stronger role separation, SME adjudication, escalation testing, adversarial-prompt evaluation, and post-game audit.

C.3 Facilitator SME Audit Rubric for Adjudication Traces

To operationalize safety-case sub-claim 3 (Adjudication Auditability) during live exercises, human facilitators and subject-matter experts (SMEs) should evaluate exercise traces against a 5-point audit rubric:

1. **Causal Explicitness:** Are all state changes, success probabilities, and side effects supported by explicit causal reasoning in the trace rather than silent pattern completion?
2. **Role Independence:** Is the adjudicator model family isolated from player and adversary model families to prevent self-confirming feedback loops?
3. **Escalation Attribution:** Is escalatory scenario momentum directly traceable to explicit player decisions rather than model-introduced hostile bias?
4. **Strategic Continuity:** Does scenario state maintain logistics, historical constraints, and force availability consistently across long-context turns?
5. **Uncertainty Disclosure:** Does the adjudication trace explicitly flag speculative completions, missing evidence, and low-confidence inferences before updating scenario state?

A safety-case norm also changes what counts as progress. A new model that produces more plausible diplomatic dialogue is not automatically safer. A model that exposes uncertainty, resists role leakage, documents causal assumptions, and declines to adjudicate beyond evidence may be more valuable for high-stakes wargaming than a model that is merely more fluent. Progress means learning where an LM-enabled exercise should not be trusted, not only making the model more fluent.

C.4 Compact checklist for authors and practitioners

- State the exercise purpose, decision context, and forbidden uses.
- Identify every human and model role: player, adversary, adjudicator, facilitator, analyst, summarizer, and reviewer.
- Record model provider, model version, decoding settings, tools, retrieval sources, prompts, and system instructions.
- Prevent unreviewed role collapse: the same model should not propose, adjudicate, and summarize a high-impact claim without independent review.
- Capture transcripts, moves, adjudications, scenario-state updates, dissent notes, and final summaries.
- Require adjudicators to state causal assumptions, evidence, uncertainty, and rejected alternatives for major world-state changes.
- Run paraphrase, prompt-sensitivity, cross-model, and red-team robustness tests.
- Include escalation and de-escalation stress tests when the scenario involves conflict, coercion, cyber operations, public health, or crisis response.
- Train stakeholders on LM failure modes, including sycophancy, hallucination, long-context drift, and narrative overconfidence.
- Audit after-action products for unsupported claims, overconfident language, and decision laundering.

D Alternative Views and Objections

Objection 1: Wargames are too niche, too military, or too dual-use for ML venues. Military applications raise dual-use concerns, and the ML community should not accelerate unsafe operational deployment. The response is that open-ended wargaming is broader than the military (Section 2), and the safety problem already concerns ML practice. More importantly, the position is not "deploy LM wargames." The position is that if the field builds or studies decision-influencing strategic agents, it should examine the settings in which those agents can shape narratives, adjudicate consequences, and affect human decision. Safety cases can reduce irresponsible use by making limits explicit. Nor do we claim open-ended wargames are the largest AI-risk pathway, only a neglected one given how close they sit to escalation and institutional decision-making.

This paper also separates open-ended wargaming from direct weapon control. Many autonomous or remotely piloted systems involve bounded perception-action loops, even when their deployment raises severe safety and governance concerns. Open-ended wargames, in contrast, operate through natural-language scenario generation, role-play, adjudication, and after-action interpretation. The primary hazard is upstream decision influence: a model changes what humans believe about adversaries, escalation, feasibility, or necessity. The domains can interact—wargames can influence doctrine or operational concepts for autonomous systems—but the safety case for an LM-enabled wargame should not be collapsed into the safety case for a weapon platform.

Objection 2: Open-ended wargames are not reproducible enough for scientific evaluation, and closed games would be better science. These are two sides of the same concern, and we treat them together. Open-ended wargames resist the clean repeatability of closed-game benchmarks. Closed games isolate variables, permit large-scale simulation, and support precise scoring (Silver et al., 2017; Perolat et al., 2022), whereas open-ended wargames vary across runs. But reproducibility should not be confused with determinism. Human-subject experiments, qualitative studies, and systems evaluations often study phenomena that are only partially repeatable. Open-ended wargames can improve reproducibility through artifact capture (structured logging of prompts, model outputs, adjudication decisions, and after-action traces), fixed scenario scaffolds (pre-specified vignettes that constrain initial conditions across runs), multiple runs, cross-model comparisons, SME review rubrics (methodical assessment forms used by subject-matter experts to assess traces), and clear logging of adjudication criteria. We are not arguing that closed games should be abandoned;

they remain the right setting for many scientific questions about planning, search, and self-play. We are arguing against using closed-game performance as evidence for a stronger claim: that an agent is safe for open-ended strategic decision support. A model that performs well in StarCraft or Diplomacy may still fail when it must adjudicate ambiguous cyber coercion, interpret diplomatic signaling, or summarize a crisis game for decision-makers. If a safety-critical use case is hard to evaluate, avoiding it does not make it safer.

Objection 3: LMs are not capable enough yet for this to matter. Current LMs do fail in wargaming contexts; studies have found brittle reasoning, hallucinations, rule non-adherence, inconsistency, and escalation risks (Lamparth et al., 2024; Rivera et al., 2024; Shrivastava et al., 2024). That is a reason to study safety now, not later. The introduction makes the converse point: as long as LMs appear capable in some prominent setting, some users will find it irresistible to extend that perception into settings where the actual capability is much weaker, even when the belief is misguided. Halo effects from frontier-model demonstrations in adjacent domains can make a brittle wargame model appear ready. The argument is reinforced by the relational character of several failure modes in Section 3: prompt sensitivity and sycophancy under leading questions are jointly produced by model and user, so the safety case must address the human side as well as the model side. Waiting until models are broadly capable as players and adjudicators would leave the field with capability demonstrations but no norms for auditability, interpretability, or decision influence.

Objection 4: Safety cases become bureaucratic checklists with no real enforcement. A safety case that merely records that a benchmark was run and a human was present degenerates into safety-washing. To avoid bureaucratic compliance, a safety case must be evaluated as a rigorous argument tested by its negative claims: explicit evidence showing where hazards exceed controls, where evidence fails, and where deployment is prohibited. Enforcement rests on existing channels in each context: procurement in defense, reviewer gating and model-card disclosure in academic publication, internal review boards, and post-incident audit. Regulatory enforcement is years away in most jurisdictions; a soft norm carried through these channels is the realistic starting point, not the end state.

Objection 5: High-stakes LM-enabled wargaming should not exist; auditing it legitimizes the use. AI-governance reviewers may argue that interacting with LM-enabled wargames, even to demand auditing, tacitly endorses the practice. Some uses should not exist at all, including real-time crisis-response systems where a model adjudicates consequences for a decision-maker with minutes to act. Our position is consistent: auditability is a precondition for high-stakes use, not a license to deploy. "No safety case, no high-stakes use" encompasses "no valid safety case is currently possible, therefore no deployment" for un-certifiable roles. Refusing to study how these systems fail forfeits the technical grounding needed to articulate refusal.