ALAMEINBENCH: EVALUATING LONG-HORIZON STRATEGIC PLANNING OF LLM AGENTS IN A HISTORICAL WARGAME

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating LLM agents in strategic environments requires tasks that combine long-horizon planning, coordinated resource allocation, and adaptation to an opponent under explicit rules. We introduce AlameinBench, a benchmark built on a turn-based hex-grid wargame for evaluating these capabilities over complete campaigns. It comprises three scenarios centered on supplied advance, mine clearance, and timed withdrawal, each evaluated from two playable roles. The resulting six scenario–role conditions couple multi-unit coordination, logistics, phase-dependent actions, and delayed scoring within a shared game environment. AlameinBench combines a common observation and action interface, rule-validated execution, and an evaluation protocol covering terminal outcomes, scenario-specific progress, execution reliability, and computational cost. We also introduce Strategy–Allocation–Execution (SAE), a modular neuro-symbolic reference agent that connects strategic objectives to unit assignments and executable actions. Using this framework, we evaluate six LLM profiles alongside four LLM-free heuristic policies and a rule-based self-play reference. In the reported fixed-seed evaluation, performance varies across scenarios and roles, and a deterministic domain-informed heuristic matches or exceeds the best SAE profile in four of six conditions. Campaign traces further reveal differences in objective maintenance, task execution, and score sources. AlameinBench provides a structured testbed and reference baselines for studying long-horizon planning and strategic decision-making in LLM-based agents.

1 INTRODUCTION

LLM-based agents operating in interactive environments must make decisions whose consequences extend beyond the current step. In strategic tasks, they need to maintain objectives across turns, coordinate multiple units, allocate limited resources, and adapt to an opponent's actions. These requirements are closely coupled: an early movement can determine whether a later operation is feasible, a resource allocation can constrain future options, and an opponent's response can invalidate part of an otherwise reasonable plan. Evaluating long-horizon planning and strategic decision-making therefore requires sustained interaction in environments where such dependencies shape task outcomes.

Existing benchmarks have advanced the evaluation of strategic reasoning and interactive decision-making through wargame tasks, strategy games, and broader collections of game environments (Yin et al., 2025; Ma et al., 2024b; Qi et al., 2024; Paglieri et al., 2025). These settings provide complementary perspectives on agent capabilities. We focus on a combination of challenges in which heterogeneous units must coordinate under explicit logistics and action-phase constraints while pursuing objectives with delayed consequences. A shared environment that combines these mechanisms provides a controlled setting for examining how agents organize decisions across units, phases, and turns, and how their performance changes with the objective and role they are assigned.

We introduce AlameinBench, a benchmark built on a turn-based hex-grid wargame. It contains three scenarios centered on supplied advance, mine clearance, and timed withdrawal. Each scenario emphasizes a different set of dependencies: advancement must preserve scoring eligibility

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

through supply, clearance requires coordinated preparation and execution, and withdrawal requires positioning units before the relevant exit opportunities become available. Agents control either Axis or Allies against a common rule-based opponent, yielding six scenario–role conditions within one game. Role reversal changes the objectives and trade-offs that the agent must manage. Together, these conditions support the evaluation of multi-unit coordination, prerequisite planning, action timing, and adaptation over complete campaigns.

AlameinBench standardizes the interaction interface and evaluation protocol across agent systems. Agents receive structured observations and interact through a common set of actions and queries, while the game engine validates actions, updates states, and determines scenario scores. Evaluation combines rule-computed terminal outcomes with scenario-specific progress evidence, execution reliability, and computational cost. Complete campaign traces connect agent decisions to applied actions and subsequent state changes, allowing score components and task execution to be examined in context. This provides both a common basis for performance comparison and the evidence needed to investigate how different decision sequences produce their outcomes.

To establish reference results, we provide Strategy–Allocation–Execution (SAE), a modular neuro-symbolic agent architecture. SAE connects strategic objectives to unit assignments and executable actions through persistent task representations and local rule checks. We evaluate six LLM profiles within this architecture, alongside four LLM-free heuristic selection policies and a rule-based self-play reference. The heuristic policies use domain-specific candidate generation and scoring, providing a substantive comparison for LLM-based planning. These experiments characterize complete agent systems operating under the benchmark protocol, including their planning, checking, and execution components.

The reported fixed-seed evaluation reveals substantial variation across scenarios and roles. A deterministic domain-informed heuristic matches or exceeds the best SAE profile in four of the six conditions, demonstrating the competitiveness of structured domain knowledge in this environment. Campaign-level analysis further identifies differences in objective maintenance, completion of prerequisite-dependent tasks, and the sources of terminal scores. For example, extensive movement and replanning can leave a clearance objective unmet, while an improved defensive score can result from force preservation without reduced enemy withdrawal. These observations illustrate how AlameinBench supports both comparative evaluation and analysis of long-horizon agent behavior.

Our contributions are threefold:

- **A strategic decision-making benchmark.** We introduce three scenarios and two playable roles that combine multi-unit coordination, logistics, phase-dependent actions, and delayed scoring in a shared environment.
- **A standardized interaction and evaluation protocol.** We connect rule-validated execution and terminal outcomes with campaign traces, scenario-specific progress evidence, and computational cost.
- **Reference agents and an empirical evaluation.** We provide SAE and evaluate six LLM profiles alongside heuristic and rule-based references, establishing initial results and analyzing performance across scenarios and roles.

2 Related Work

Strategic and wargame evaluation. WGSR-Bench evaluates situation awareness, opponent-risk modeling, and policy generation, and includes an LLM wargame agent (Yin et al., 2025). TextWargame describes a textual observation/action interface and hierarchical action framework (Chen et al., 2026). Our emphasis is complementary: complete-campaign traces with explicit links among plans, rule outcomes, task evidence, and execution cost. Human–model comparisons in crisis simulations study behavioral differences and escalation rather than hex-level movement and supply (Lamparth et al., 2024). We do not claim to be the first wargame benchmark or infer real-world military ability from game scores.

Interactive strategy environments. TextWorld provides controllable text games (Côté et al., 2018), and ALFWorld aligns abstract text policies with embodied tasks (Shridhar et al., 2021). Crafter measures achievement-based progress (Hafner, 2022); the NetHack Learning Environment

supports exploration and long-horizon planning (Küttler et al., 2020). TextStarCraft II uses strategic summaries (Ma et al., 2024b), CivRealm supports Civilization agents (Qi et al., 2024), and BALROG evaluates agents across games (Paglieri et al., 2025). MACHIAVELLI examines reward–behavior trade-offs (Pan et al., 2023), SMACv2 studies cooperative combat (Ellis et al., 2023), and CICERO combines dialogue with Diplomacy strategy (Meta Fundamental AI Research Diplomacy Team (FAIR), 2022). LLM Skirmish is a public strategy-code project (LLM Skirmish Contributors, 2026). AlameinBench focuses on phase-constrained multi-unit coordination, supply, role-specific objectives, and delayed campaign scoring.

Interactive agent benchmarks. AgentBench spans diverse interactive environments (Liu et al., 2024b). MINT studies tool use and language feedback (Wang et al., 2024); τ-bench checks final states (Yao et al., 2024); ToolSandbox evaluates stateful dependencies and milestones (Lu et al., 2024). WebArena evaluates functional task completion on websites (Zhou et al., 2023), while OS-World uses execution-based evaluation in computer environments (Xie et al., 2024). AppWorld and SWE-bench assess executable coding outcomes (Trivedi et al., 2024; Jimenez et al., 2024). AlameinBench applies state-grounded evaluation to complete campaigns rather than web or software tasks.

Neuro-symbolic decision systems. Neuro-symbolic AI integrates neural learning with explicit knowledge representation and reasoning (De Raedt et al., 2020; Garcez & Lamb, 2023). SAE follows a modular form of this integration: pretrained LLMs generate structured proposals, while symbolic task predicates, dependency checks, and game rules govern progress assessment and execution. Its proposal–verification cycle is related to LLM-Modulo (Kambhampati et al., 2024); the present contribution is its organization and evaluation over complete, multi-unit campaigns.

Progress, cost, and evaluation scope. AgentBoard tracks intermediate progress through defined or annotated subgoals (Ma et al., 2024a). HELM evaluates multiple dimensions beyond accuracy (Liang et al., 2023), and AI Agents That Matter emphasizes joint assessment of effectiveness and cost (Kapoor et al., 2024). Building on these perspectives, our evaluation links fixed scenario scores, model-proposed task evidence, executed actions, and module costs. Our focus is linking these signals at the campaign level.

3 Evaluation Environment and Interaction

3.1 Battlefield, Information, and Control

AlameinBench uses a turn-based hex-grid game in which one controller commands multiple units on one side and a rule-based opponent commands the other (Figure 1). The irregular battlefield spans 49 numbered columns and includes open terrain, ridges, depressions, roads, and minefields. Mechanized and non-mechanized ground units differ in mobility and combat strength; engineers and mobile supply units provide specialized capabilities. The initial July, September, and October positions contain 32, 48, and 80 ground and engineer units across both sides, respectively, excluding supply units, mines, and administrative markers. Appendix A.1 gives the scenario inventory.

Let s_t contain the battlefield, unit attributes and supply state, active side, turn and phase, and scenario scoring state. The agent receives a structured observation $o_t = O(s_t, h_t)$ combining state-derived information with selected history and memory h_t. Both sides' state is available: this study does not simulate fog of war or claim an intrinsically partially observable task. Uncertainty instead includes the opponent's future decisions and stochastic rule outcomes.

During its side's phases, the controller chooses unit movements, attacks and their participants, eligible mine-clearance or exit actions, holds, and phase completion. Map, rule, supply, and combat queries support these choices. Each interaction step applies at most one game action after rule validation:

$$s_{t+1} = F(s_t, a_t, \xi_t), \qquad a_t \in \mathcal{A}_{\text{legal}}(s_t), \tag{1}$$

where ξ_t represents rule-governed random outcomes, including combat resolution. Invalid proposals do not change the game state. The engine, not the model or checker, determines legality and scoring; the same boundary also applies to the opponent.

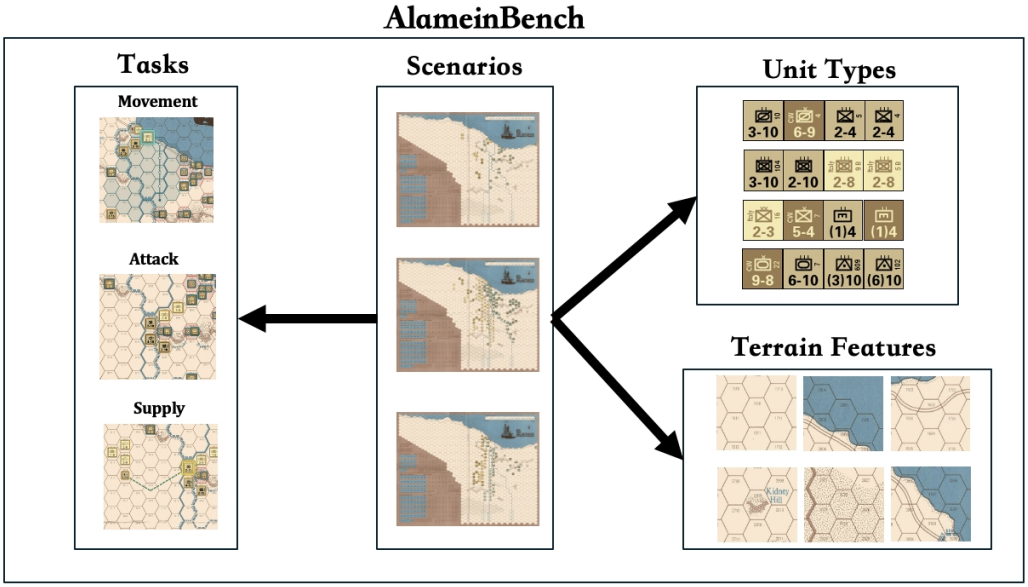


Figure 1: Overview of AlameinBench. *Tasks*: movement, attack, and supply actions on the hex map. *Scenarios*: initial deployments for July, September, and October. *Unit types*: representative Axis and Allied counters, each labeled with combat strength and movement allowance. *Terrain features*: open desert, coastline, roads, ridges, depressions, and coastal settlements.

3.2 Mechanics That Couple Decisions

Spatial coordination and logistics. Movement consumes terrain-dependent allowances and is constrained by enemy occupation, zones of control, and stacking limits. A route that is feasible for one unit may cease to be useful after another unit occupies its destination. Concentrating combat strength must therefore be balanced against congestion, exposed routes, and supply access. Supply is not merely descriptive: it affects unit condition and eligibility for scenario rewards. In July, reaching an eastern position without qualifying supply does not establish scoring progress. This couples advance, support, and preservation over multiple decisions.

Phase order and opportunity timing. Each side normally proceeds through initial movement, combat, mechanized movement, and supply movement before end-of-turn settlement; scenario exceptions modify this schedule. July and September begin with Axis, while October begins with Allies. Units cannot freely interchange these action opportunities. Preparatory movement must precede the relevant attack or clearance, and a later supply movement cannot retroactively change an earlier action. A hold ends a unit's current opportunity, not its strategic task; ending a phase can postpone an unfinished dependency until the next turn. Thus a seven- or fifteen-turn campaign comprises many more individual decisions than its turn count suggests.

Risk and delayed effects. Combat combines participant selection and terrain-dependent odds with a random result, which can cause retreat, elimination, or exchange. Legal execution and favorable outcomes are consequently different properties. A locally attractive attack can also disrupt a later route or expose a supply dependency. The opponent's intervening actions require plans to adapt without losing the campaign objective.

3.3 Three Scenarios, Two Roles

All scenarios start from fixed deployments and terminate at their scenario's final end-of-turn settlement. Intermediate VP can be inspected, but a temporary advantage is not terminal success. Table 1 and Figure 1 give their horizons, role objectives, and initial deployments. July awards three VP per column beyond column 34 reached by the farthest qualifying supplied or partially supplied Axis

Table 1: Three scenarios and two roles give six evaluation conditions within one game. Initial and final victory points (VP) are Axis-oriented. Objectives summarize the main scenario dependencies, not every scoring component.

Scenario	Turns	Initial VP	Axis objective	Allied objective
July	7	25	Advance the qualifying supplied combat frontier eastward.	Limit Axis scoring advance while preserving forces and supply.
September	7	35	Clear Allied minefields through coordinated approach and clearance.	Deny Axis clearance and limit combat and isolation losses.
October	15	−20	Prepare and complete eligible western exits after turn 10.	Interdict withdrawal without sacrificing friendly forces to isolation.

combat unit. September awards three VP per qualifying Allied mine marker cleared by Axis. October awards withdrawal points only for eligible Axis western exits after turn 10; merely moving west is not a successful withdrawal. Allied combat and isolation losses can also increase Axis VP. We therefore decompose the recorded score rather than infer its source from position or total VP alone.

Axis seeks higher VP and Allies lower VP. July Allied defense concerns the *Axis* scoring frontier, not the easternmost Allied unit. October defense requires separating denied enemy exits from preserved friendly forces: either can improve total VP, but they demonstrate different behaviors.

4 Trace-Based Evaluation

For condition c and seed z, let $V_{m,c,z}$ be the final Axis VP of system m and $B_{c,z}$ the rule-AI self-play reference. We report raw VP and role-adjusted improvement

$$\Delta_{m,c,z} = \sigma_c(V_{m,c,z} - B_{c,z}), \qquad \sigma_c = \begin{cases} +1, & \text{external side is Axis,} \\ -1, & \text{external side is Allies.} \end{cases} \tag{2}$$

Higher Δ is better for either role, but raw VP should not be averaged indiscriminately across scenarios with different scoring systems.

We evaluate outcomes, tactical progress, execution reliability, and computational cost from recorded traces. Beyond the metrics in Table 2, diagnostics include scoring components, accepted moves, combat odds/outcomes, scenario actions, task evidence, application-layer illegal actions, phase completion, and reported resource use. Unlike a shared annotated subgoal set such as AgentBoard's (Ma et al., 2024a), SAE's self-proposed task predicates diagnose a particular plan; scenario scoring remains the fixed outcome criterion.

Table 2 summarizes the five reported metrics: two rule-grounded primary metrics and three supplementary Judge metrics. Planning and strategy are assessed jointly through campaign outcomes and cross-turn evidence of preparation, objective maintenance, and adaptation; single-turn coherence alone does not measure long-term planning. Appendix B.2 specifies denominators, scales, and missing-data rules.

Table 2: Reported metrics. The first two are computed by the rule engine; the last three are supplementary LLM-Judge assessments.

Metric	Meaning
Final Axis VP	Rule-based terminal campaign score.
Rule rejection (%)	Rule-rejected / model-submitted action attempts.
Missed severity (0–3)	Mean Judge severity over confirmed opportunities.
Conversion (%)	Judge-converted / confirmed opportunities.
Plan coherence (1–5)	Mean Judge coherence per valid side-turn window.

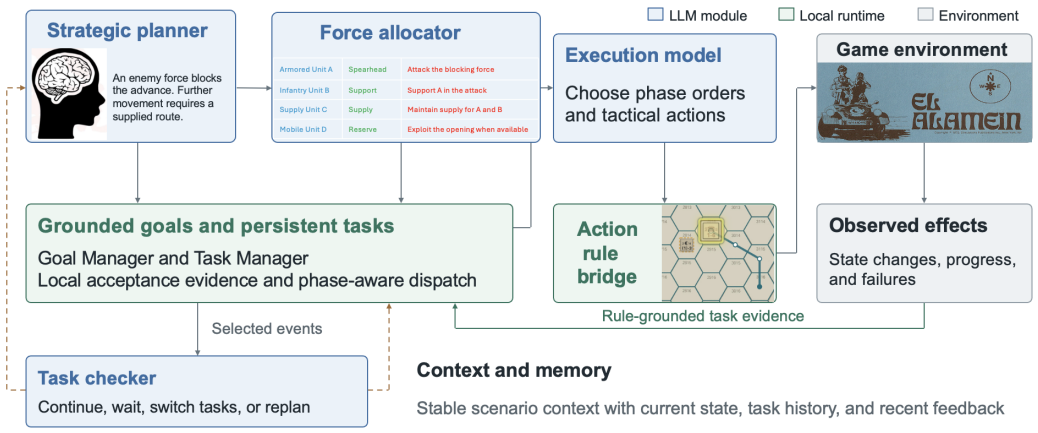


Figure 2: SAE architecture. LLM modules (blue) propose a strategy, unit allocations, and phase orders; the local runtime (green) grounds them as persistent tasks and applies one rule-verified action at a time. Observed state changes return as rule-grounded task evidence, and the task checker reviews selected events to continue, wait, switch tasks, or replan.

5 Reference Agents

SAE: a modular NeSy architecture. As shown in Figure 2, a strategic model proposes a campaign or operational objective and decomposes it into related tasks. Each objective is expressed as a testable condition specifying what is evaluated, the desired effect, and the scope of assessment. Task dependencies and unit assignments persist across decisions, while progress is assessed against observed state changes. The symbolic task layer maintains these conditions and dependencies explicitly; its state updates do not require additional learned planning components.

An allocation call assigns units to roles and tasks; local checks enforce ownership and remove conflicting or infeasible assignments. During eligible movement phases, a unit-plan call proposes orders that the runtime executes one action at a time, revalidating against the updated state. Combat and other action decisions use the enabled tools. Failures can enter bounded repair. Actual action effects feed back into the task state and subsequent decisions. A separate checker model reviews progress and risks at selected events; it cannot directly move units or change the rule engine's score. Strategy is revisited when tasks finish or fail, progress stalls, or supply and enemy conditions change.

The neural–symbolic interface is the structured task representation: LLM-generated objectives are normalized into predicates over units, regions, scoring metrics, and evaluation scopes, which symbolic checks evaluate against observed states. This modular coupling follows NeSy integration (De Raedt et al., 2020; Garcez & Lamb, 2023) and proposal–verification approaches (Ahn et al., 2022; Kambhampati et al., 2024), coupling pretrained LLMs with symbolic task checks without additional training (Appendix A.2).

Heuristic selection baselines. SR0–SR3 use a common complex-rule-AI candidate generator and raw scorer, without an LLM or SAE task system. SR0 takes the maximum-scoring candidate. SR1 samples uniformly among at most three candidates within one raw score point of the best. SR2 uses uniform exploration with probability 0.1 and otherwise chooses greedily. SR3 samples from a softmax with temperature 0.5. Candidate pruning and an 80-candidate cap make these *heuristic selection baselines*, not uniform random legal-action agents. Appendix A gives details.

Evaluation settings. Table 3 lists the evaluated systems. SAE is evaluated with six LLM profiles, SR0–SR3 cover the same six conditions without model calls, and rule-AI self-play supplies the per-scenario reference.

Table 3: Evaluated systems. All systems play the same scenarios against the same rule-based opponent at the same seed.

System	Scope
SAE model profiles	July, September, and October, both roles; shared checker.
SR0–SR3	Heuristic selection policies across scenario–role conditions.
Rule-AI self-play	Scenario-specific references matched to the SAE scenarios and random seed.

Table 4: Combined experimental results. Each scenario cell gives final Axis VP as external Axis / external Allies (higher / lower is better). Reject is the pooled rule-rejection percentage. Miss (0–3), Conv. (%), and Coh. (1–5) are supplementary Judge severity, opportunity conversion, and coherence. Lower Miss and higher Conv./Coh. are better. Blank cells are unavailable or inapplicable, not zeros.

		Final Axis VP (Axis / Allies)				Judge (row-level pooled)		
Model	Method	July	Sept.	Oct.	Reject (%)	Miss	Conv.	Coh.
DeepSeek V4 Pro	SAE	31 / 28	35 / 47	61 / 55	11.9	0.84	47.4	3.81
Claude Opus 5	SAE	31 / 30	35 / 48	43 / 55	7.5	1.14	40.0	3.16
GPT-6 Astra	SAE	31 / 37	38 / 53	65 / 55	0.9	0.69	46.2	3.41
Claude Sonnet 5	SAE	43 / 28	35 / 41	36 / 53	6.7	0.60	58.1	3.47
Kimi K3	SAE	25 / 38	35 / 50	17 / 55	2.6	1.00	44.2	3.02
GLM-5.3	SAE	25 / 34	35 / 50	-13 / 55	0.0	1.12	42.3	2.39
/	SR0	25 / 28	43 / 52	66 / 46		0.31	76.9	3.02
/	SR1	25 / 29	35 / 48	61 / 47		0.55	67.5	2.97
/	SR2	25 / 29	38 / 51	41 / 43		0.18	87.8	3.17
/	SR3	25 / 28	38 / 49	61 / 51		0.16	92.0	3.10
/	Rule AI	28 / 28	35 / 35	60 / 60		0.17	83.3	3.26

6 RESULTS

Table 4 pools non-score metrics over six games per SAE/SR row. SR and rule AI have no model-submission denominator; their retrospective Judge evaluations cover 232 and 58 windows, respectively. Rule AI repeats three self-play scores across roles, not six independent games; its Judge pools both role views. Appendix D reports coverage.

Legality is necessary but not sufficient. Every SAE and SR campaign completes with zero application-layer illegal actions, although SAE models still submit rejected proposals (0.0–11.9%). Outcomes nonetheless diverge sharply under this common legality: SAE July Axis ranges from 25 to 43 VP and October Axis from -13 to 65. Rule rejection also does not predict strategic success; GLM-5.3 has the lowest rejection rate and the lowest October Axis score.

Performance is scenario- and role-dependent. The SAE mean role-adjusted differences from rule-AI self-play are $+3.0$ and -4.5 for July Axis and Allies, $+0.5$ and -13.2 for September, and -25.2 and $+5.3$ for October. The distinction between aggregate point estimates and statistical evidence is emphasized by Agarwal et al. (2021); we do not interpret these means as significance tests. September Axis performance remains close to its initial 35 VP, while the evaluated September Allied systems allow more Axis VP than the self-play reference. The scenario–role pattern therefore matters more than a single pooled score.

LLM-free heuristics rival the SAE agents. The SR policies use no language model, yet the deterministic SR0 alone matches or exceeds the best SAE profile in four of six conditions: July Allies (28 versus 28 Axis VP), September Axis (43 versus 38), October Axis (66 versus 65), and October Allies (46 versus 53). In October Allies, every SR policy (43–51) outperforms every SAE profile (53–55) and the rule-AI reference (60). SAE leads in July Axis (Claude Sonnet 5, 43 versus 25) and September Allies (Claude Sonnet 5, 41 versus 48), but no evaluated system approaches the reference's 35 VP in September Allies. The SR policies achieve this with zero in-game model re-

Table 5: Model-level profiles across the six conditions. Δ is the role-adjusted score relative to rule-AI self-play (Equation 2), averaged over all conditions, the three Axis conditions, or the three Allied conditions. Best counts conditions in which the profile attains the highest Δ, including ties. Submitted counts model-submitted actions over six campaigns; output tokens are per external decision step.

Model	Mean Δ	Axis Δ	Allies Δ	Best	Submitted	Output tok./step
DeepSeek V4 Pro	-0.5	+1.3	-2.3	1	729	702
Claude Opus 5	-4.0	-4.7	-3.3	0	398	732
GPT-6 Astra	-1.8	+3.7	-7.3	2	454	630
Claude Sonnet 5	-1.3	-3.0	+0.3	4	479	2,673
Kimi K3	-11.0	-15.3	-6.7	0	191	1,831
GLM-5.3	-15.3	-25.3	-5.3	0	16	7,129

quests, compared with 4,001 SAE requests (Section 7). Because they inherit the rule AI's candidate generator and scorer, the comparison shows that domain-specific action proposals remain highly competitive with general-purpose LLM planning, not that the LLMs lack the relevant knowledge.

Randomized selection has no consistent observed advantage. Relative to deterministic SR0, randomized SR1–SR3 improve September Allied scores but score lower in September and October Axis; SR3 attains the highest Judge conversion (92.0%) yet improves only one of six scores (Appendix A).

Scoring components change the interpretation of defense. In September Allied evaluation, 98.7% of the aggregate Axis gain above initial VP comes from cleared mines. The primary observed failure is therefore scenario-objective denial, not simply excessive combat losses. In October, rule-AI self-play scores $-20 + 73 + 7 = 60$: withdrawal contributes 73 and Allied isolation 7. SAE Allied results decompose as $-20+75+0 = 55$, except Claude Sonnet 5 at $-20+73+0 = 53$. The evaluated profiles improve their role-adjusted final score, but do not reduce withdrawal points below the reference. The improvement comes from avoiding the isolation component, while withdrawal is unchanged or higher.

No model dominates, and roles induce specialization. Table 5 profiles each model. Claude Sonnet 5 attains the best role-adjusted score in four of six conditions, including the only large July Axis breakthrough ($+15$), whereas DeepSeek V4 Pro achieves the highest mean Δ (-0.5) through consistent, near-reference play. GPT-6 Astra is the strongest attacker (Axis $+3.7$, the only September mine clearance) but concedes the most in defense (Allies -7.3); Claude Sonnet 5 shows the reverse pattern and is the only profile with a positive Allied mean. Rankings transfer poorly across roles: in September, Axis and Allied Δ are negatively rank-correlated across models (Spearman $\rho = -0.66$).

Committed engagement, not deliberation volume, tracks outcomes. Across the six profiles, the number of model-submitted actions orders mean Δ exactly ($\rho = 1.0$), from 729 actions for DeepSeek V4 Pro to 16 for GLM-5.3. GLM-5.3 produces the most output per decision step (7,129 tokens) yet rarely commits an action, and in October Axis submitted actions track withdrawal VP ($\rho = 0.94$; 76 VP for GPT-6 Astra, 0 for GLM-5.3). Higher rule-rejection rates accompany better outcomes ($\rho = 0.71$), consistent with rejections arising from attempted play. Judge coherence orders the profiles exactly as mean Δ does ($\rho = 1.0$), so the supplementary Judge agrees with rule-computed outcomes at the model level.

7 Diagnosing the Decision Pipeline

7.1 Plans and Movement Can Fail to Produce Scenario Progress

The DeepSeek V4 Pro September Axis run illustrates the gap between action volume and useful progress. It records 162 moves, 28 passes, and nine replans, but ends at its initial 35 VP with no mine-clearance score. Its first engineer's end-of-turn positions are 3121, 3220, 3119, 3120, 3020,

3121, and 3019. The agent repeatedly changes position while its two-mine clearance objective remains unmet. Six replans are attributed to lack of task progress and three to enemy-state changes.

Counting replans or movements alone would mistake this activity for progress; the rule-grounded score and task evidence expose it.

July shows a related horizon problem. DeepSeek V4 Pro reaches scoring column 36 in turn 1, changes its operational target to column 37 in turn 2, but finishes at column 36 and 31 VP after six replans. Claude Sonnet 5, by contrast, pushes the qualifying frontier to column 40 for 43 VP.

7.2 Planning Style, Objective Maintenance, and Score Sources

Task-completion rates reflect each model's planning style. In July Axis, Kimi K3 and GLM-5.3 each propose a single measurable task and complete it while finishing at 25 VP, whereas Claude Sonnet 5 proposes 21 and completes two on the way to 43 VP. Ambitious plans with many subgoals therefore coexist with the strongest outcome, so we report task completion alongside its objective and terminal score rather than as a standalone ranking.

In DeepSeek V4 Pro October defense, the recorded withdrawal ceiling rises from zero initially to 2, 36, 56, and 73 in turns 12–15, while actual withdrawal finishes at 75 points. During turns 11–15 the external side makes no further moves or attacks. The progressively relaxed ceiling, rather than any interception effort, tracks the Axis withdrawal, illustrating late objective drift. Final scores likewise require action-level interpretation: GLM-5.3's October Axis trajectory contains no moves or withdrawals and scores -13 from the initial -20 plus seven points caused by opposing isolation. The score increase is therefore not evidence of successful Axis withdrawal.

7.3 Planning Dominates Attributed Request Time

Following the joint evaluation of effectiveness and cost in AI Agents That Matter (Kapoor et al., 2024), the SAE traces record 4,001 requests, 51,542,211 input tokens, and 5,963,800 output tokens (Appendix Table 8). Usage is reported for 3,061 requests (76.5%), with profile-level coverage from 58.7% to 97.4%. Checker calls account for 1,382 requests, or 34.5%, but only 5.3% of summed attributed request duration. Strategic planning accounts for 57.7% of that duration and unit planning for 26.8%. Tool execution accounts for 33.4% of reported tokens but 6.5% of attributed duration. Thus optimizing only checker frequency would not address the largest measured latency component.

7.4 LLM-Judge Scores as Supplementary Evidence

MT-Bench examines LLM-judge agreement and biases (Zheng et al., 2023); G-Eval uses structured assessment for generated text (Liu et al., 2023b); Prometheus learns rubric-conditioned evaluation (Kim et al., 2024). Agent-as-a-Judge extends evaluation to intermediate evidence in agentic task solving (Zhuge et al., 2024). Our Judge is a prompted trajectory evaluator that evaluates turn-level trajectory windows for missed opportunities and multi-step coherence over five opportunity categories: joint attack, objective capture, supply restoration, mine clearance, and breakthrough exploitation.

Because these scores assess local decision quality rather than campaign outcomes, they complement the rule-computed scores: SR3 attains the highest conversion rate yet improves only one of six scores over SR0 (Section 6). Appendix D gives full scores and coverage.

8 Conclusion

AlameinBench evaluates long-term planning and strategy through complete campaigns. Across six LLMs, legality is universal but effectiveness is not: a deterministic LLM-free heuristic matches or exceeds the best SAE agent in four of six conditions, no model dominates across roles, and committed engagement rather than deliberation volume orders model outcomes. Its diagnostic protocol supports testing whether future agents sustain effective strategies.

AI USE STATEMENT

Generative AI tools assisted with feedback on experimental design and methodological framing, implementation and debugging of the agent and evaluation code, processing and qualitative analysis of experiment logs, interpretation of results, and translation between English and Chinese. Additional uses included literature search and summarization, reference formatting, and drafting and editing manuscript text. Language models also serve as the evaluated agents and supplementary Judge, as described in Sections 5 and 7; these experimental roles are distinct from AI assistance in preparing the study. Campaign results are derived from recorded game executions, and Judge assessments are identified separately as model-generated evaluations. This work does not present mathematical theorems or proofs.

We have reviewed the AI-assisted work. Supporting automated checks compare reported table entries with stored experiment summaries and verify numerical consistency and bilingual alignment. Source records are retained to support verification of reported results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with generative AI assistance.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2108.13264.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*, 2022. URL https://arxiv.org/abs/2204.01691.

Zilie Chen, Haoran Wu, Shengqi Shen, and Qiyue Yin. TextWargame: A benchmark evaluating LLMs on strategic reasoning. In *International Conference on Computer and Communication Systems (ICCCS)*, 2026. doi: 10.1109/ICCCS69761.2026.11613233. URL https://ieeexplore.ieee.org/document/11613233.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. TextWorld: A Learning Environment for Text-based Games. *arXiv preprint arXiv:1806.11532*, 2018. URL https://arxiv.org/abs/1806.11532.

Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, and Giuseppe Marra. From statistical relational to neuro-symbolic artificial intelligence. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 4943–4950, 2020. doi: 10.24963/ijcai.2020/688. URL https://www.ijcai.org/proceedings/2020/688.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N. Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2212.07489.

Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: The 3rd Wave. *Artificial Intelligence Review*, 56:12387–12406, 2023. doi: 10.1007/s10462-023-10448-w. URL https://doi.org/10.1007/s10462-023-10448-w.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 2021. URL https://arxiv.org/abs/1803.09010.

Danijar Hafner. Benchmarking the Spectrum of Agent Capabilities. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2109.06780.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06770.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.01817.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI Agents That Matter. *arXiv preprint arXiv:2407.01502*, 2024. URL https://arxiv.org/abs/2407.01502.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.08491.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The NetHack Learning Environment. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.13760.

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. *arXiv preprint arXiv:2403.03407*, 2024. URL https://arxiv.org/abs/2403.03407.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. LLM+P: Empowering Large Language Models with Optimal Planning Proficiency. *arXiv preprint arXiv:2304.11477*, 2023a. URL https://arxiv.org/abs/2304.11477.

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*, 2024a. URL https://arxiv.org/abs/2307.03172.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2308.03688.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. URL https://arxiv.org/abs/2303.16634.

LLM Skirmish Contributors. LLM Skirmish. Public software project, 2026. URL https://github.com/llmskirmish/skirmish. Accessed September 2026; not cited as a peer-reviewed study.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. *arXiv preprint arXiv:2408.04682*, 2024. URL https://arxiv.org/abs/2408.04682.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. In *Advances in Neural Information Processing Systems*, 2024a. URL https://arxiv.org/abs/2401.13178.

Weiyu Ma, Qirui Mi, Yongcheng Zeng, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. Large language models play StarCraft II: Benchmarks and a chain of summarization approach. In *Advances in Neural Information Processing Systems*, 2024b. URL https://arxiv.org/abs/2312.11865.

Meta Fundamental AI Research Diplomacy Team (FAIR). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/10.1126/science.ade9097.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

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 *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.13543.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2304.03279.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint arXiv:2304.03442*, 2023. URL https://arxiv.org/abs/2304.03442.

Siyuan Qi, Shuo Chen, Yexin Li, Xiangyu Kong, Junqi Wang, Bangcheng Yang, Pring Wong, Yifan Zhong, Xiaoyuan Zhang, Zhaowei Zhang, Nian Liu, Wei Wang, Yaodong Yang, and Song-Chun Zhu. CivRealm: A learning and reasoning odyssey in civilization for decision-making agents. *arXiv preprint arXiv:2401.10568*, 2024. URL https://arxiv.org/abs/2401.10568.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023. URL https://arxiv.org/abs/2303.11366.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2407.18901.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2206.10498.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. URL https://arxiv.org/abs/2305.16291.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.10691.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. *arXiv preprint arXiv:2404.07972*, 2024. URL https://arxiv.org/abs/2404.07972.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023a. URL https://arxiv.org/abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

Qiyue Yin, Pei Xu, Qiaozhe Li, Shengda Liu, Shengqi Shen, Tong Wang, Yihong Han, Xiaonan Zhao, Likun Yang, Shiyue Cao, Shiyu Qiu, Yuxuan Liu, Shizhao Yu, Lei Cui, Chengxin Yan, Jie Sun, Xiangquan Tang, and Kaiqi Huang. WGSR-Bench: Wargame-based game-theoretic strategic reasoning benchmark for large language models. *arXiv preprint arXiv:2506.10264*, 2025. URL https://arxiv.org/abs/2506.10264.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. *arXiv preprint arXiv:2307.13854*, 2023. URL https://arxiv.org/abs/2307.13854.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-Judge: Evaluate Agents with Agents. *arXiv preprint arXiv:2410.10934*, 2024. URL https://arxiv.org/abs/2410.10934.

A Protocol and Evidence

A.1 Environment Details

Table 6: Initial scenario inventory. Each paired entry is Axis / Allies. Ground units, engineers, mobile supply units, and mine markers are counted separately. Supply-vanguard and other administrative markers are excluded from unit counts; these are initial forces, not counts of surviving units or available actions.

Scenario	Ground	Engineers	Mobile supply	Mine markers
July	14 / 18	0 / 0	2 / 2	0 / 10
September	25 / 21	2 / 0	3 / 3	18 / 31
October	38 / 36	0 / 6	4 / 4	70 / 24

The inventory is counted from the three initial scenario files whose hashes match the recorded experiment artifacts. The map's coordinate envelope spans columns 1–49 and rows 0–34; the playable boundary is irregular, so this envelope is not a count of playable hexes. Likewise, the terrain file's explicitly tagged hexes are not the complete map. The standard turn has four phases for each side followed by end-of-turn settlement. September skips the Allied initial-movement phase on turn 1. Campaign completion occurs at the final settlement (turn 7 for July/September and turn 15 for October), not when a model declares its tasks complete.

The shared movement configuration gives mechanized and non-mechanized units different terrain costs; stacking limits are four Axis or three Allied units per hex. Enemy occupation and zones of control further constrain routes. Combat uses ten odds columns, from 1:4 to 7:1, and six possible die values. In October, qualifying western exits after turn 10 contribute ten VP per supply unit and the scenario-defined combat-strength contribution for combat units. Scoring eligibility and realized rule events, rather than movement direction or a textual claim of success, determine these components.

A.2 Interaction and Reference-Agent Design

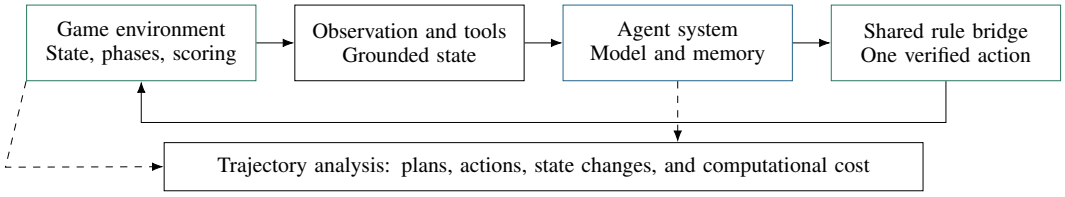


Figure 3: Benchmark boundary. The agent system may vary; the game and action-verification boundary are shared. Logs attribute model proposals and actual state changes separately. SAE is one agent system inside this interface.

Figure 3 summarizes the interaction boundary. To study how strategic intentions translate into outcomes, we link model plans to executed actions and the state changes they produce. Rule events allow final scores to be decomposed into their contributing actions and effects. Experimental conditions accompany the results so that observed behavior can be examined under the corresponding settings. This approach follows the emphasis on transparency in Datasheets for Datasets (Gebru et al., 2021) and supports interpretation and reproducibility.

SAE's action–feedback loop is related to ReAct's interleaving of reasoning and actions (Yao et al., 2023b) and Reflexion's use of verbal feedback in episodic memory (Shinn et al., 2023). Tree of Thoughts searches over reasoning branches (Yao et al., 2023a), while Voyager accumulates an executable skill library (Wang et al., 2023). SAE implements neither mechanism: it organizes model-proposed objectives through role-aware goal grounding, persistent task evidence, local rule validation, and execution tracing.

PlanBench distinguishes plausible plans from valid reasoning about action effects (Valmeekam et al., 2023). SayCan grounds skill selection in affordance values (Ahn et al., 2022); LLM+P delegates

formalized problems to classical planners (Liu et al., 2023a); LLM-Modulo couples proposals with external verifiers (Kambhampati et al., 2024). SAE shares the separation of proposals and verification and applies it to persistent, multi-unit campaign tasks.

Context combines a stable scenario/rules prefix and map reference with current state, tasks, and selected history. The evaluated policy retains full within-step context, cross-step summaries with bounded excerpts, and rejected options; phase changes or replanning reset selected recent context. Generative Agents uses retrieved experiences and reflection (Park et al., 2023), while MemGPT manages memory tiers (Packer et al., 2023). Lost in the Middle shows position-sensitive use of long context (Liu et al., 2024a). These works motivate SAE's explicit memory design.

A.3 Experimental Protocol

SAE configuration. SAE assesses progress through explicit task conditions and acts within a declared set of tools and computational budgets. The experiment records retain model choices, context policies, and checker settings for reproduction and comparison.

Artifact identity. Rule, scenario, map, opponent, replay, and source identifiers are retained in `data/evidence.json`. The results table groups runs by method, model profile, scenario, and external side; complete artifact and runtime hashes remain available for replay and audit.

Randomized action-selection study. SR0–SR3 share a candidate generator and raw scorer, use no language model or SAE task system, and play against the complex rule AI. SR0 is deterministic; SR1 samples among at most three top candidates within one point of the maximum, SR2 uses ϵ-greedy selection with $\epsilon = 0.1$, and SR3 uses softmax with temperature 0.5. The study changes the selection policy, not merely the seed. Each policy is evaluated on all six scenario–role conditions at seed 1942. All 24 campaigns complete with zero application-layer illegal actions and zero model API requests; no final-score entries are missing. SR0 is not the rule-AI self-play reference, whose action selection also applies a minimum-score gate.

Table 4 reports the final Axis VP for each condition. Relative to SR0, the numbers of conditions with better/tied/worse role-oriented scores are 1/1/4 for SR1, 2/1/3 for SR2, and 1/2/3 for SR3. These are score comparisons, not campaign win rates. All four policies score 25 in July Axis; benefits elsewhere depend on the scenario and role. The selection stream is separate from combat randomness, but different action histories can consume different combat draws, so a common seed does not match each battle outcome. In-game model requests are zero; retrospective Judge calls are accounted for separately.

Retrospective SR evaluation uses the archived SAE Judge prompt, configuration, model profile, and external-side turn windows. The adapter links 8,788 recorded decisions to their actual rule results and applies the shared first-six-candidate projection. Policy names and sampling probabilities are excluded; absent plans remain absent. State snapshots remain available for audit but are not additionally injected. All 232 windows returned valid scores; coverage is detailed in Appendix D.

Table 7: Descriptive SAE averages across the listed model profiles. Mean Δ is oriented to the external side relative to the matched rule-AI reference.

Scenario	Role	Mean Axis VP	Reference	Mean Δ
July	Axis	31.00	28	+3.0
July	Allies	32.50	28	-4.5
September	Axis	35.50	35	+0.5
September	Allies	48.17	35	-13.2
October	Axis	34.83	60	-25.2
October	Allies	54.67	60	+5.3

Table 8: SAE execution and reported token totals by model profile. Steps count external decisions, including phase completion. Tokens include all SAE modules and the checker; missing usage is not imputed.

Profile	Steps	Requests	Tokens (M)	Usage coverage
DeepSeek V4 Pro	831	696	10.83	97.4%
Claude Opus 5	628	622	7.66	59.0%
GPT-6 Astra	700	629	7.54	81.2%
Claude Sonnet 5	662	762	19.81	91.3%
Kimi K3	433	629	5.46	58.7%
GLM-5.3	269	663	6.20	66.4%

Table 9: Request attribution in the SAE traces. Token share uses reported input plus output usage. Time share uses summed attributed request duration, not elapsed batch time or server reasoning time.

Module	Requests	Request share	Token share	Time share
Strategy	839	21.0%	29.1%	57.7%
Allocation	204	5.1%	5.4%	3.2%
Unit plan	928	23.2%	21.6%	26.8%
Tool execution	581	14.5%	33.4%	6.5%
Repair	67	1.7%	0.6%	0.5%
Checker	1382	34.5%	9.9%	5.3%

Table 10: October Allied score decomposition. Every row has initial Axis VP -20. Lower final VP favors the external Allied side. Improved total scores do not imply reduced withdrawal.

External Allied controller	Axis withdrawal VP	Allied isolation VP	Final Axis VP
Rule AI self-play	73	7	60
DeepSeek V4 Pro	75	0	55
Claude Opus 5	75	0	55
GPT-6 Astra	75	0	55
Claude Sonnet 5	73	0	53
Kimi K3	75	0	55
GLM-5.3	75	0	55

Table 11: Reported July Axis task completion versus final score. Denominators count measurable tasks, not equal-difficulty benchmark objectives. Different task counts make a direct completion-rate ranking inappropriate.

SAE main profile	Final Axis VP	Completed/measurable	Rate
DeepSeek V4 Pro	31	5/7	71.4%
Claude Opus 5	31	2/3	66.7%
GPT-6 Astra	31	5/8	62.5%
Claude Sonnet 5	43	2/21	9.5%
Kimi K3	25	1/1	100.0%
GLM-5.3	25	1/1	100.0%

B Additional Quantitative Diagnostics

B.1 Completed-Run Diagnostics

B.2 Metric Definitions and Aggregation

Final victory points (VP). The rule engine computes terminal Axis VP from the scenario's initial score and recorded scoring events. Each scenario cell reports the score when the evaluated system controls Axis and when it controls Allies, respectively. Higher Axis VP favors Axis; lower Axis

VP favors Allies. VP is neither a win rate nor a percentage, and its scale varies by scenario. Only completed campaigns contribute terminal scores. The role-adjusted gain in Equation 2 compares a run with its matched rule-AI self-play reference; it does not measure improvement over the initial position.

Rule-rejection rate (Reject). Let S be the number of model-submitted action attempts and R the number rejected for violating game rules. We report $100R/S$ percent. Multiple attempts at one decision step are counted separately, including rejected attempts before a successful repair. Local candidate probes are not model submissions; strategic or advisory objections are not rule rejections when their type is recorded. This rate measures proposal legality, whereas application-layer illegal actions count invalid actions actually applied to the game. Rule AI and SR have no model-submission denominator, so their rejection cells are blank, not zero.

Missed-opportunity severity (Miss). For confirmed, non-abstained opportunity decisions O, let s_o be the Judge's severity score. Miss is $\sum_{o \in O} s_o/|O|$ on a 0–3 scale: 0 denotes conversion/no miss, and 1, 2, and 3 denote minor, clear, and critical misses. Lower is better. The denominator includes converted opportunities, not only missed ones; thus the main-table metric differs from severity conditional on a miss. The inherited opportunity categories are joint attack, objective capture, supply restoration, mine clearance, and breakthrough exploitation. This is a Judge-defined set of recognized opportunities, not an exhaustive enumeration of legal alternatives.

Opportunity conversion (Conv.). Let $C \subseteq O$ contain confirmed opportunities labeled converted by the Judge. We report $100|C|/|O|$ percent; higher is better. Conversion describes recognized opportunity uptake from the supplied action evidence, not a rule-verified favorable effect or campaign success. Unknown and abstained opportunities are excluded from both numerator and denominator, rather than counted as failures. A window with no confirmed opportunities adds no opportunity observation and is not assigned zero severity or a zero conversion rate.

Multi-step plan coherence (Coh.). Each window contains one side's recorded decisions within one game turn, spanning its action phases. For non-abstained, valid windows W, coherence is $\sum_{w \in W} c_w/|W|$. The 1–5 rubric ranges from conflicting actions (1), frequent plan changes (2), and a basic plan (3), to clear coordination (4) and a complete effective action chain (5). Higher is better. When no explicit plan is recorded, the Judge assesses the observed sequence without inventing a plan. A window can receive a coherence score even when it contains no confirmed opportunity.

Pooling, coverage, and auxiliary diagnostics. Within a table row, we sum rejection counts and submissions, opportunity severities and confirmed opportunities, and coherence scores and valid windows before division. We do not average per-game ratios with unequal denominators. Longer campaigns therefore contribute more turn windows; the pooled Judge score is descriptive, not an equal-scenario or independent-sample estimate. Failed judgments, abstentions, and missing observations do not become zeros; a zero denominator leaves a blank cell. Coverage is reported separately from performance, and Judge confidence is not used as a performance weight. For the self-play reference, two role views still come from one campaign. Task completion and rule-verified opportunity effects remain separate auxiliary diagnostics requiring observable predicates and linked state changes. Reported request usage and client-side durations describe computational cost, not gameplay quality.

C Qualitative Evidence Anchors

Case identifiers refer to the derived evidence table, not a new experiment. Each can be resolved to its original transcript through `data/evidence.json`.

Table 12: Selected audit anchors and the conclusions they support. These are deliberately narrower than causal claims about model internals.

Case	Observed evidence	Supported interpretation
A01: DeepSeek V4 Pro, July Axis	Qualifying frontier reaches 36 on turn 1; target changes to 37 on turn 2; final frontier remains 36.	Replanning does not establish further scoring progress.
A03: DeepSeek V4 Pro, September Axis	162 moves, 28 passes, nine replans; no clearance score; engineer revisits approach positions.	High legal action volume can fail to produce the scenario-specific outcome.
A06: DeepSeek V4 Pro, October Allies	No moves or combats in turns 11–15; withdrawal ceiling repeatedly relaxed; final Axis withdrawal 75 VP.	Late inactivity coincides with relaxed objectives and continued enemy withdrawal.
A19: Claude Sonnet 5, July Axis	At hex 4013 a key unit changes from unsupplied before settlement to partially supplied at game end; final frontier 40.	Terminal scoring evidence is valid but not equivalent to persistent full supply.
A35: GLM-5.3, October Axis	No moves or withdrawals; final $-13 = -20 + 7$ from opposing isolation.	The VP increase does not establish progress on the Axis withdrawal objective.

Attribution of combat and pass. Combat counts refer to external-side actions, not all battles in the games. An inability to find a legal attack may be a consequence of earlier movement, but the pass event alone cannot reveal whether the model considered and rejected an attack. The appropriate unit of diagnosis is the preceding state and action sequence.

D Supplementary Judge Results

Table 13: Aggregate Judge outputs. The opportunity and coherence aggregates use different eligible-window denominators, retained in the evidence package.

System/profile	Missed severity	Coherence
DeepSeek V4 Pro	0.84	3.81
Claude Opus 5	1.14	3.16
GPT-6 Astra	0.69	3.41
Claude Sonnet 5	0.60	3.47
Kimi K3	1.00	3.02
GLM-5.3	1.12	2.39

The Judge is separate from the shared DeepSeek V4 Pro SAE task checker. Missed-opportunity severity ranges from 0 (converted/no miss) to 3 (critical miss); coherence ranges from 1 (conflicting actions) to 5 (a complete effective operation chain). Its recorded opportunity categories are joint attack, objective capture, supply restoration, mine clearance, and breakthrough exploitation.

The existing missed-opportunity score and coherence score answer different questions. A missed-opportunity judgment requires a feasible alternative at that time, a clear benefit under the active scenario, and a specified evaluation horizon. Coherence requires a connection among plan, action sequence, and adaptation, but coherent pursuit of an ineffective plan is still poor gameplay. We interpret both scores alongside the rule-grounded outcomes and qualitative evidence anchors.

Randomized-policy Judge coverage. The 24 SR games contribute 232 successfully judged windows, with no failed windows. Opportunity metrics exclude abstentions and use confirmed opportunity counts; coherence uses valid turn windows. Absence of an explicit plan does not by itself imply

incoherence: the Judge evaluates observed action sequences. Table 14 supplies the denominators for Table 4.

Table 14: Heuristic-policy Judge coverage. Each SR policy has six games; Rule AI has three self-play games viewed from both roles. Each row has 58 expected windows. Opportunity abstentions are not zero-severity observations.

Policy	Judged/expected	Confirmed	Converted	Opp. abstain	Coh. valid	Coh. abstain
SR0	58/58	39	30	20	58	0
SR1	58/58	40	27	21	58	0
SR2	58/58	49	43	22	58	0
SR3	58/58	50	46	15	58	0
Rule AI	58/58	18	15	28	57	1

Rule-AI reference evaluation. Retrospective evaluation covers both sides of the three recorded self-play games using the same archived Judge prompt, configuration, and model profile as SAE/SR. The 1,765 recorded decisions produce 58 successfully judged windows, not six independent campaigns. The Judge confirms 18 opportunities, labels 15 converted, and abstains on 28 opportunity entries; severity is $3/18 = 0.17$ and conversion is $15/18 = 83.3\%$. Coherence averages $186/57 = 3.26$, excluding one abstained window. The adapter retains actual rule outcomes and recorded phase-ending decisions, uses the shared first-six-candidate projection, and does not invent plans.