Moreau Arena: A Controlled Strategy Benchmark for Evaluating Agentic AI

Victor Kuzmenko* Claude (Krakenshu)† Grok‡

February 2026

Abstract

Static benchmarks are failing to differentiate modern AI agents. MMLU is saturated, Chatbot Arena measures preference rather than capability, and game-based benchmarks introduce visual or mechanical confounds. More fundamentally, fixed benchmarks invite memorization: once the optimal strategy is known, the benchmark measures recall rather than reasoning. We present **Moreau Arena**, a *living* strategy benchmark where agents design creatures by allocating 20 stat points across four attributes and selecting from 14 animal species with 56 unique abilities. Creatures fight on an 8×8 grid under deterministic rules, isolating strategic reasoning from execution skill.

Moreau Arena introduces four innovations: (1) a **dynamic meta-balancing system** that continuously adjusts ability parameters via EMA-tracked win rates within a seasonal rotation framework, creating a benchmark that *evolves faster than agents can memorize it*—a direct countermeasure to the strategy collapse that plagues static evaluations; (2) a **draft-based evaluation protocol** with bans, counter-picking, and build adaptation that tests strategic depth beyond one-shot optimization; (3) an **offline patch generator** that produces hash-verified, diffable season configurations, ensuring tournament reproducibility while preserving meta evolution; and (4) a **multi-model tournament methodology** where frontier LLMs (Claude, GPT-5.2, Gemini, Grok) compete head-to-head under identical constraints, yielding Bradley–Terry rankings with bootstrap confidence intervals.

In two multi-model tournaments across 1,559 best-of-7 series with 13 agents (8 LLM + 5 baseline), we observe that prompt improvements—exact formulas, meta context, opponent build reveal—shift LLM agents from below-baseline (Tournament 001) to above-baseline performance (Tournament 002). Adaptation provides a measurable advantage for top models (+15% to +39% counter-pick win rate for Codex, Flash, and Grok). The benchmark processes 10,000+ matches per minute on consumer hardware at an estimated API cost under $2 for a full cross-model tournament.

1 Introduction

The rapid advancement of large language models has created an urgent need for benchmarks that evaluate agentic reasoning—the capacity to make consequential decisions under uncertainty, adapt to opponents, and optimize across complex trade-off spaces. Existing benchmarks increasingly fail at this task.

Static knowledge benchmarks like MMLU [Hendrycks et al., 2021] and GSM8K [Cobbe et al., 2021] have reached saturation: frontier models score above 90%, leaving insufficient headroom to

*Independent Research. Correspondence: supervitek@users.noreply.github.com
†Anthropic. AI collaborative partner under the Round Table methodology.
‡xAI. External reviewer and collaborative partner.

distinguish capability differences. These benchmarks test recall and pattern matching, not the strategic planning that characterizes agentic behavior.

Human preference benchmarks like Chatbot Arena [Chiang et al., 2024] capture subjective quality but conflate style with substance. A model that produces eloquent but strategically incoherent reasoning scores well on preference but poorly on tasks requiring actual optimization. The signal is further diluted by prompt sensitivity and evaluator variance.

Game-based benchmarks offer a promising direction—games provide natural adversarial settings with clear win conditions. However, existing approaches face critical limitations. Video game benchmarks like Orak (ICLR 2026) require visual processing that confounds measurement of strategic reasoning with perceptual ability. Complex strategy games like StarCraft [Vinyals et al., 2019] introduce thousands of interacting variables that make it impossible to isolate which cognitive capabilities drive performance. Game-theoretic benchmarks like GTBench [Duan et al., 2024] provide clean theoretical foundations but limited strategic depth, as optimal strategies for many included games are well-known.

We argue that an effective agentic AI benchmark must satisfy four properties:

1. **Controlled complexity.** The game must be complex enough to require genuine reasoning but simple enough that the full decision space is transparent to the agent.

2. **Deterministic reproducibility.** Given identical inputs, the benchmark must produce identical outputs, enabling rigorous comparison across models and configurations.

3. **Adversarial depth.** The benchmark must reward adaptation and counter-play, not just memorization of a single optimal strategy.

4. **Evolving meta.** The benchmark must resist strategy collapse—the phenomenon where a single dominant approach renders all alternatives irrelevant.

Moreau Arena is designed to satisfy all four properties. Agents face a constrained optimization problem (20 stat points, 14 animals, 56 abilities) with deterministic combat resolution, a draft system that rewards adaptation, and a dynamic meta-balancing engine that prevents any single build from permanently dominating.

1.1 Contributions

We make the following contributions:

1. **A living benchmark.** We present Moreau Arena, a strategy benchmark that isolates agentic reasoning from perceptual or execution confounds. Unlike static evaluations, the benchmark *evolves*: a dynamic meta-balancing system automatically adjusts ability parameters based on aggregate win rates within seasonal rotations, preventing any single strategy from permanently dominating. The design space spans 13,566 creature configurations and the engine processes 10,000+ matches per minute on consumer hardware.

2. **Offline patch generator architecture.** We introduce a reproducibility-preserving approach to benchmark evolution: an offline pure-function patch generator computes balance adjustments from match history, producing hash-verified season configurations. Tournaments run on frozen season snapshots (no parameter drift during execution), while the meta evolves *between* seasons. This resolves the tension between benchmark reproducibility and benchmark freshness (Section 3.6).

3. **Multi-model tournament methodology.** We define a head-to-head evaluation protocol where frontier LLMs compete under identical constraints (prompt, temperature, token budget, timeout) in best-of-7 draft series. Bradley–Terry scoring with bootstrap confidence intervals provides statistically grounded rankings. All matches produce JSONL provenance records enabling independent verification (Section 4.3).

4. **Cross-model empirical results.** In two tournaments (779 and 780 series respectively) with 13 agents each, we report comparative adaptation rates, counter-pick coherence, and series-level play patterns across Claude (Opus, Sonnet, Haiku), GPT-5.2, GPT-5.2-Codex, Gemini (Flash, Pro), and Grok. We observe that prompt engineering and structured output dramatically improve LLM performance, and that adaptation provides measurable advantages for top-tier models (Sections 7–7.2).

2 Related Work

2.1 Static Benchmarks

The dominant paradigm in LLM evaluation relies on static question-answer datasets. MMLU [Hendrycks et al., 2021] tests knowledge across 57 subjects; HumanEval [Chen et al., 2021] evaluates code generation; GSM8K [Cobbe et al., 2021] tests mathematical reasoning. While foundational, these benchmarks share a critical limitation: they are non-adversarial. The correct answer does not depend on an opponent's actions, and performance saturates as models improve. Frontier models now achieve 90%+ on MMLU, compressing the effective evaluation range.

SWE-bench [Jimenez et al., 2024] addresses some limitations by evaluating real-world software engineering tasks with measurable outcomes (resolved GitHub issues), but remains fundamentally static—the repository state does not adapt to the agent's approach.

2.2 Human Preference Benchmarks

Chatbot Arena [Chiang et al., 2024] introduced pairwise human evaluation at scale, providing a continuously-updated Elo leaderboard based on human preference. This captures aspects of model quality invisible to automated metrics. However, preference evaluation conflates multiple signals: instruction following, writing style, factual accuracy, and strategic reasoning are all collapsed into a single ranking. For agentic evaluation specifically, a model that produces a well-written but strategically suboptimal plan may score higher than one that produces a terse but correct strategy.

2.3 Game-Based Benchmarks

Games provide natural testbeds for agentic AI due to their adversarial structure, clear objectives, and measurable outcomes. The use of games for AI evaluation has a long history, from chess [Silver et al., 2018] to Go [Silver et al., 2016] to real-time strategy [Vinyals et al., 2019].

Orak (ICLR 2026) evaluates agents across 12 commercially available video games, measuring adaptability and generalization. While comprehensive, it introduces visual processing as a confounding variable—an agent may fail due to perceptual limitations rather than strategic deficiency.

GTBench [Duan et al., 2024] evaluates LLMs on game-theoretic reasoning across 10 classic games including prisoner's dilemma, auction mechanisms, and negotiation scenarios. While theoretically grounded, the strategic depth is limited—optimal strategies for most included games are well-known and can be pattern-matched rather than derived through genuine reasoning. Moreau

Arena's 13,566-configuration design space and non-transitive dynamics provide substantially greater optimization complexity.

Multi-agent strategic games. Superhuman AI systems have been developed for poker [Brown & Sandholm, 2019], Diplomacy [Bakhtin et al., 2022], and StarCraft II [Vinyals et al., 2019]. These demonstrate impressive capabilities but require specialized training (millions of self-play games) rather than evaluating general-purpose reasoning from a rules description. Moreau Arena tests whether an agent can reason about a novel game from its specification alone—a capability more relevant to general agentic deployment.

Procedural environment benchmarks. The NetHack Learning Environment [NLE; Küttler et al., 2020] provides a fast, procedurally-generated roguelike environment with rich long-horizon challenges. While strategically deep and free of visual confounds (symbolic ASCII state), NLE is fundamentally single-agent and designed for RL training loops rather than evaluating LLM reasoning from a rules description.

Real-time adversarial benchmarks. LM Fight Arena [Zheng et al., 2024] evaluates large multimodal models through head-to-head competition in Mortal Kombat II, testing real-time visual processing and low-latency tactical decisions. However, it requires visual input processing and evaluates tactical reflexes more than strategic planning.

Transparent reasoning benchmarks. GAMEBoT [Lin et al., 2025] evaluates LLM reasoning across 8 diverse games with chain-of-thought validation of intermediate reasoning steps. Its strength is transparency—measuring not just action quality but reasoning quality. However, opponents are rule-based rather than adaptive, and there is no evolving meta.

Social deduction benchmarks. AvalonBench [Light et al., 2023] and Werewolf-based benchmarks [Xu et al., 2023] test natural language persuasion and social deduction, evaluating a fundamentally different capability set than strategic optimization.

2.4 How Moreau Arena Differs

Moreau Arena occupies a distinctive niche in the benchmark landscape (Table 1).

Table 1: Comparison of AI evaluation benchmarks across key properties.

Property	MMLU	Chatbot Arena	SWE-bench	Orak	GTBench	NLE	LM Fight	GAMEBoT	GVGAI-LLM	AvalonBench	**Moreau**
Adversarial	No	No	No	Yes	Yes	No	Yes	Yes	No	Yes	**Yes**
Deterministic	Yes	No	No	No	Yes	No	Par	Yes	No	No	**Yes**
No visual conf.	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	**Yes**
Tests adaptation	No	No	No	Par	No	Par	Yes	Par	Par	Par	**Yes**
Self-evolving	No	No	No	No	No	No	No	No	Par	No	**Yes**
Strategic depth	Low	N/A	Hi	Hi	Med	VHi	Med	Hi	Med	Med	**Med-H**
Saturation risk	Hi	Low	Med	Low	Med	Med	Low	Med	Low	Low	**Low**
LLM-native	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	**Yes**

The combination of deterministic resolution, adversarial depth (via the draft system), and an evolving meta (via the auto-balancer) creates a benchmark that rewards genuine strategic reasoning while remaining fully reproducible and computationally efficient.

Unlike Orak's visual overhead or StarCraft's execution demands, Moreau Arena presents pure strategy: the agent's only input is the game rules and opponent information, and its only output is a stat allocation. This isolation is deliberate—it allows us to measure strategic reasoning without confounding it with other capabilities.

The self-evolving meta is, to our knowledge, a novel contribution with no direct analogue in existing benchmarks. By automatically adjusting ability parameters based on aggregate win rates within a seasonal framework, Moreau Arena ensures that memorizing the current optimal strategy provides only temporary advantage.

3 System Design

Moreau Arena is a turn-based auto-battler where two creatures fight on an 8×8 grid under fully deterministic rules. The agent's task is purely strategic: select an animal species and allocate stat points. Combat execution is handled entirely by the engine, eliminating any confound from motor skill, reaction time, or visual processing.

Formal problem definition. Let $\mathcal{A} = \{a_1, \ldots, a_{14}\}$ be the set of animal species and $\mathcal{S} = \{(h, k, s, w) \in \mathbb{Z}_+^4 \mid h + k + s + w = 20,\ \forall x \geq 1\}$ be the set of valid stat allocations. A *build* is a pair $b = (a, \mathbf{s}) \in \mathcal{A} \times \mathcal{S}$. The combat function $f : \mathcal{A} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{Z} \to \{\text{a}, \text{b}, \text{draw}\}$ maps two builds and a seed to a deterministic outcome. The agent's objective is to choose b^* that maximizes expected win rate against the distribution of opponent builds—a constrained optimization problem over a discrete space of $|\mathcal{A}| \times |\mathcal{S}| = 14 \times 969 = 13{,}566$ configurations.

3.1 Creature Design Space

Each creature is defined by two orthogonal choices:

Animal species. Agents choose from 14 animal species, each with a unique passive trait and two triggered abilities. For example, Bear carries Fury Protocol (Berserker Rage + Last Stand), Boar carries Charge (Stampede + Gore), and Tiger carries Ambush Wiring (Pounce + Hamstring). Species choice determines qualitative combat identity—the *type* of advantage a creature pursues.

Stat allocation. Agents distribute exactly 20 points across four attributes (HP, ATK, SPD, WIL), each with a minimum of 1. Stat allocation determines quantitative combat parameters—*how much* of each advantage the creature has. The total design space is $\binom{19}{3} = 969$ valid stat allocations per animal, yielding 13,566 possible creature configurations across all 14 species.

The interaction between species and stats is non-trivial. A Bear with 4/14/1/1 (offense-maximized) leverages Berserker Rage's ATK multiplier on a high base, while a Bear with 11/5/1/3 (defensive) wastes the synergy. Evaluating these interactions is the core reasoning challenge.

3.2 Stat Formulas

Each raw stat maps to derived combat parameters through explicit, deterministic formulas (Table 2).

Table 2: Stat-to-combat parameter derivation formulas.

Stat	Derived Value	Formula
HP	Max Hit Points	$50 + 10 \times \text{HP}$
ATK	Base Damage	$\lfloor 2 + 0.85 \times \text{ATK} \rfloor$
SPD	Movement Range	1 if SPD ≤ 3, 2 if SPD ≤ 6, 3 if SPD ≥ 7
SPD	Dodge Chance	$\min(0.30,\ 0.025 \times (\text{SPD} - 1))$
WIL	Ability Range	$\min(4,\ \lceil \text{WIL}/2 \rceil)$
WIL	Resist Chance	$\min(0.35,\ 0.03 \times (\text{WIL} - 1))$
WIL	Ability Power	$1.0 + 0.05 \times \text{WIL}$

These formulas are provided to agents in their prompt, making the optimization problem *transparent*. The key insight that ATK has super-linear returns (Berserker Rage multiplies base damage, which itself scales with ATK) is derivable from the formulas but requires multi-step reasoning to discover—analogous to the chain-of-thought reasoning studied by Wei et al. [2022].

Creature size is derived from HP + ATK sum: $\leq$10 yields 1×1, 11–12 yields 2×1, 13–17 yields 2×2, and 18+ yields 3×2. Larger creatures occupy more grid cells, granting a zone of control (free attacks on adjacent enemies) but limiting mobility.

3.3 Combat Resolution

Combat proceeds in discrete ticks on an 8×8 grid, with a hard cap of 60 ticks. Each tick follows a fixed phase order:

1. **Attack Phase.** Each creature attacks if adjacent to an opponent. Damage is computed as $\text{raw} = \lfloor \text{base\_dmg} \times \text{ability\_mod} \rfloor$, reduced by armor (capped at 50% of raw damage), then perturbed by a seeded $\epsilon \in [-0.05, +0.05]$.
2. **Buff Tick.** Active ability buffs decrement their duration timers.
3. **Ability Procs.** Each creature's two abilities roll independently for activation. Proc rates are split into two tiers: *strong* abilities (Last Stand, Berserker Rage, Gore, Mimic, Iron Will) at 3.5% per tick, and *standard* abilities at 4.5% per tick.
4. **Fury Check.** Bear's Fury Protocol activates below 50% HP, granting +20% damage. Non-stacking with Berserker Rage (takes maximum).
5. **DOT Tick.** Damage-over-time effects (Rend, Venom) apply their damage.
6. **Closing Ring.** After tick 30, the arena shrinks inward, dealing increasing unavoidable damage to creatures in the ring zone (1 dmg/tick at ticks 30–34, scaling to 3 dmg/tick at ticks 40+).
7. **Second Wind / Regen.** Healing effects resolve.

Determinism. All randomness derives from a single match seed via SHA-256 hash chains:

$$\text{tick\_seed} = \text{SHA256}(\text{match\_seed} \| \text{tick\_index}) \bmod 2^{32} \tag{1}$$

$$\text{hit\_seed} = \text{SHA256}(\text{match\_seed} \| \text{tick\_index} \| \text{attack\_index}) \bmod 2^{32} \tag{2}$$

Given identical inputs (builds + seed), the engine produces bit-identical outputs across runs and platforms.

Combat AI is rule-based and identical for all creatures, eliminating AI quality as a variable:

1. Ranged ability in range → use ability
2. Melee adjacent → attack
3. HP $< 25\%$ and escape route → retreat
4. Enemy has DoT and path exists → kite
5. In closing ring zone → move to center
6. Default → approach and attack

3.4 Ability System

Each animal species has exactly two triggered abilities—one offensive and one utility/defensive—drawn from a pool of 28 distinct ability types. Abilities proc stochastically each tick based on their proc rate.

The split proc rate system (3.5% for powerful abilities, 4.5% for standard abilities) is a deliberate balance mechanism. Strong abilities like Berserker Rage (+60% damage for 3 ticks) and Gore ($0.6\times$ damage that ignores dodge) have lower activation rates to compensate for their outsized impact. This creates a risk-reward tradeoff: builds relying on strong abilities have higher variance, while builds using standard abilities have more consistent output.

WIL provides a proc bonus of $+\text{WIL} \times 0.0008$ per tick per ability, and a resist chance of $\min(0.60,\ \text{WIL} \times 0.033)$ against enemy ability effects. This gives WIL-heavy builds a defensive identity—they reduce the impact of opponent abilities while increasing their own proc frequency, at the cost of raw damage output.

3.5 Draft System

The single-build evaluation protocol (Section 5) measures static optimization. The draft system extends this to sequential decision-making under adversarial pressure.

Format. Best-of-5 series. Before each game, both sides simultaneously ban one animal species. The banned animal is unavailable for that game only (bans do not accumulate across the series). After bans, both sides independently choose a build.

Winner locking. The winner of each game must use the same build in the next game. This prevents runaway dominance from a single build and forces the loser to find a counter.

Loser adaptation. The loser may change both animal and stat allocation for the next game, subject to a 4-point stat reallocation cap—they may shift at most 4 stat points from their previous build's allocation. This tests incremental adaptation rather than full redesign.

The draft system converts creature design from a one-shot optimization into a multi-round strategic game. Agents must reason about: which animals to ban (denying the opponent's strongest options), how to counter-pick after seeing the opponent's locked build, how to adapt within the 4-point reallocation constraint, and series-level strategy (sacrificing an early game to gather information).

3.6 Dynamic Meta-Balancing

Static benchmarks suffer from *strategy collapse*—once the optimal approach is known, the benchmark measures recall rather than reasoning. MMLU saturation illustrates this failure mode at scale [Hendrycks et al., 2021]. Moreau Arena addresses strategy collapse through a dynamic meta-balancing system that continuously adjusts ability parameters based on aggregate performance.

3.6.1 EMA Win Rate Tracking

The balancer maintains an exponential moving average (EMA) of win rates per animal species, with smoothing factor $\alpha = 0.25$:

$$\mathrm{WR}_t^{(a)} = \alpha \cdot \mathrm{outcome}_t + (1 - \alpha) \cdot \mathrm{WR}_{t-1}^{(a)} \tag{3}$$

where $\mathrm{outcome}_t \in \{0, 1\}$ and a indexes over all 14 animal species. The relatively high α ensures rapid response to meta shifts—after 10 consecutive wins, the EMA reaches approximately 0.94 from a 0.5 baseline.

3.6.2 Rebalance Trigger and Adjustment

Every $W = 500$ matches, the balancer evaluates all species. Any species whose EMA win rate deviates more than $\pm\tau$ from the target is flagged for adjustment, where $\tau = 0.10$ and $\mathrm{WR}_{\mathrm{target}} = 0.55$:

$$|\mathrm{WR}^{(a)} - \mathrm{WR}_{\mathrm{target}}| > \tau \tag{4}$$

The target of 0.55 (rather than 0.50) reflects a design choice: we allow slightly above-average animals to exist, as perfect 50% win rates across all species would indicate a lack of meaningful differentiation.

For flagged species, the adjustment magnitude is:

$$\delta = \mathrm{clamp}\left(\gamma \cdot (\mathrm{WR}^{(a)} - \mathrm{WR}_{\mathrm{target}}),\ -\delta_{\max},\ +\delta_{\max}\right) \tag{5}$$

where $\gamma = 0.075$ is a damping factor and $\delta_{\max} = 0.05$ is the per-cycle cap. The adjustment is applied as a *subtraction* to both the proc rate multiplier m_p and ability power multiplier m_w:

$$m_p^{(a)} \leftarrow m_p^{(a)} - \delta, \quad m_w^{(a)} \leftarrow m_w^{(a)} - \delta \tag{6}$$

A positive deviation (overpowered species) yields positive δ, which *decreases* the multipliers (nerf). A negative deviation (underpowered species) yields negative δ, which *increases* the multipliers (buff).

3.6.3 Convergence Properties

The system is designed to converge rather than oscillate, via three mechanisms:

1. **Damped adjustment.** The damping factor $\gamma = 0.075$ ensures that only 7.5% of the win rate deviation is corrected per cycle.
2. **Per-cycle cap.** The $\delta_{\max} = 0.05$ cap provides a hard bound on adjustment speed.
3. **Diminishing returns.** Proc rate reductions have sub-linear effects on win rate, providing implicit damping beyond the explicit γ factor.

In practice, the system approaches equilibrium within 3–5 adjustment cycles (1,500–2,500 matches).

3.6.4 Safety Bounds

Two mechanisms prevent the balancer from producing degenerate configurations:

Cooldown. Each animal has a cooldown of $C = 700$ matches between adjustments. This prevents rapid successive nerfs/buffs that could destabilize the meta through feedback loops.

Proc rate floor and ceiling. Effective proc rates are bounded regardless of accumulated multiplier adjustments:

$$0.025 \leq r_{\text{base}} \times m_p^{(a)} \leq 0.055 \tag{7}$$

where r_{base} is the ability's base proc rate (0.035 for strong abilities, 0.045 for standard abilities).

3.6.5 Offline Patch Generator

While the `AutoBalancer` operates online during gameplay, the **offline patch generator** (`propose_patch()`) provides a reproducibility-preserving alternative for tournament execution. The patch generator is a *pure function* with no side effects:

$$\texttt{propose\_patch} : \text{MatchHistory} \times \text{Params} \to \text{Patch} \tag{8}$$

Season configuration workflow:

1. **Generate patch.** `propose_patch(match_history, base_config_hash)` → `PATCH_S1`.
2. **Merge season.** `merge_season(MOREAU_CORE_v1, PATCH_S1)` → `SEASON_S1`.
3. **Hash and freeze.** The season config receives a fresh SHA-256 hash. The resulting `SEASON_S1.json` is immutable—any modification invalidates the hash.

This architecture resolves a fundamental tension in benchmark design: **reproducibility requires frozen parameters**, but **longevity requires evolving parameters**. By separating the evolution mechanism from the execution environment, we achieve both.

Table 3: Auto-balancer parameter summary.

Parameter	Symbol	Value	Rationale
EMA smoothing	α	0.25	Responsive to recent meta (eff. window $\sim$7)
Check window	W	500	Balance frequency vs. noise
Target win rate	$\text{WR}_{\text{target}}$	0.55	Allow slight imbalance
Deviation threshold	τ	0.10	Trigger on meaningful deviation
Damping factor	γ	0.075	Prevent over-correction
Per-cycle cap	$\delta_{\max}$	0.05	Hard bound on adjustment speed
Cooldown	C	700	Prevent cascade adjustments
Proc rate floor	—	0.025	No ability becomes inactive
Proc rate ceiling	—	0.055	No ability becomes dominant

4 Methodology

4.1 The Round Table Approach

Moreau Arena was developed through a collaborative methodology we term the *Round Table approach*: a human researcher working alongside multiple AI model instances as intellectual partners, each contributing distinct analytical perspectives to benchmark design.

The Round Table operates on three principles:

1. **Collective authorship.** All contributions are attributed to the collective, regardless of which participant (human or AI) originated them.

2. **Adversarial collaboration.** Partners are encouraged to critique proposals. A balance change that survives scrutiny from multiple independent reasoners is more robust than one approved by a single reviewer.

3. **Transparent methodology.** All design decisions, including rejected alternatives, are documented with their reasoning chains.

4.2 Multi-Model Review Process

A central mechanism in the Round Table methodology is the *Koordinatsionnyi Sovet* (Coordination Council, KC)—a structured multi-model review process for major design decisions. For each significant design choice (stat formula calibration, ability parameters, proc rate splits), a KC session consisting of proposal generation, independent analysis, external review, consensus scoring (CQ $\geq$ 8 to proceed), and synthesis is conducted.

This process was applied to three critical design decisions in Moreau Arena:

- **Variant C stat formula selection:** Three formula variants (A, B, C) were simulated across 3,600 matches each. Variant C (ATK coefficient 0.85, HP coefficient 10) achieved CQ 8 after demonstrating the best balance between offensive and defensive builds.

- **Split proc rate calibration:** The decision to differentiate strong abilities (3.5%) from standard abilities (4.5%) emerged from a KC session that identified single-rate systems as causing either ability-dominant or ability-irrelevant metas.

- **Auto-balancer parameters:** The EMA alpha (0.25), damping factor (0.075), and cooldown window (700) were tuned through KC review of simulation data showing convergence behavior.

4.3 Multi-Agent Evaluation Protocol

The evaluation methodology separates an agent's reasoning quality from its strategic prior. We instantiate multiple independent LLM agents, each receiving identical game rules but primed with a different strategic philosophy:

- **Offense-first:** "Maximize damage output. Identify the highest-DPS combination."

- **Defense-first:** "Maximize survivability. Identify the build that can outlast any opponent."

- **Speed-first:** "Maximize evasion and mobility. Identify the build that avoids damage while dealing consistent output."

Each agent produces a single build through one reasoning step—no iterative refinement, no access to simulation results. The builds then compete in a full round-robin tournament (100 matches per pair, deterministically seeded).

4.4 Benchmark Execution

Tournaments are executed on consumer hardware (NVIDIA RTX 3080 system) at throughput exceeding 10,000 matches per minute. Each match is a pure Python simulation requiring no GPU acceleration.

Elo calculation. We use standard Elo [Elo, 1978] with $K = 32$ and an initial rating of 1500:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{9}$$

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{10}$$

where $S_A \in \{1, 0.5, 0\}$ for win, draw, or loss.

Statistical significance. Each pair in a round-robin plays 100 matches across different seeds. With $n = 100$ matches, a win rate difference of 10% is significant at $p < 0.05$ (binomial test).

Reproducibility. All match results are fully reproducible given the same builds and seed sequence. The seeded random system (SHA-256 hash chains) ensures cross-platform determinism. We provide the complete seed list alongside results to enable independent verification. The benchmark is open-source with 940+ automated tests validating engine correctness.

5 Preliminary Results: 6-Agent Tournament

5.1 Experimental Setup

We evaluate six agents in a round-robin tournament: three Claude Sonnet 4.6 sub-agents with distinct strategic philosophies (Offense, Defense, Speed), two handcrafted baselines (GlassCannon: Bear 3/14/2/1; Greedy: Boar 8/8/3/1), and one random control. Each sub-agent receives identical game rules and must produce a single build through one reasoning step. The tournament consists of 100 matches per pair (1,500 total), all deterministically seeded. This experiment uses 6 of the 14 implemented animals (Bear, Tiger, Wolf, Monkey, Buffalo, Boar).

5.2 Results

Table 4: 6-agent round-robin results. Elo standard errors are ±28 points.

Rank	Agent	Build	Elo	WR	Type
1	Offense	Bear 4/14/1/1	1874	85.2% ± 3.6%	LLM (single-shot)
2	GlassCannon	Bear 3/14/2/1	1856	83.4% ± 3.7%	Empirical (3,600 sims)
3	Greedy	Boar 8/8/3/1	1677	67.0% ± 4.7%	Handcrafted
4	Defense	Buffalo 11/5/1/3	1501	36.8% ± 4.8%	LLM (single-shot)
5	Speed	Tiger 2/6/11/1	1379	27.4% ± 4.5%	LLM (single-shot)
6	Random	Raven 3/3/2/12	714	0.2% ± 0.4%	Control

The Offense agent independently identified what appears to be the dominant strategy—maximizing ATK on Bear for Berserker Rage synergy—matching the empirically-optimized baseline within 18 Elo points. Head-to-head between Offense and GlassCannon was 49–51%, effectively a coin flip between Bear 4/14/1/1 and Bear 3/14/2/1.

5.3 Key Finding: Reasoning Quality Correlates with Elo

The three LLM agents exhibited a clear hierarchy in reasoning quality:

- **Offense** correctly identified the damage formula (base_dmg $= \lfloor 2 + 0.85 \times \text{ATK} \rfloor$) as the dominant variable and maximized accordingly.
- **Defense** reasoned correctly within its framework (maximize HP for Iron Will scaling) but failed to recognize that ATK efficiency exceeds HP efficiency per stat point.
- **Speed** correctly identified dodge mechanics and Hamstring synergy but missed hard counters (Boar's Gore ignores dodge) and the dodge ceiling at 30%.

This suggests that Moreau Arena effectively discriminates between different *qualities* of strategic reasoning, not just different strategy choices.

5.4 Single-Shot LLM vs. Empirical Search

Table 5: Comparison of reasoning-based and search-based discovery of optimal builds.

Method	Build	Elo	Cost
Claude Sonnet 4.6 (1 API call)	Bear 4/14/1/1	1874	∼$0.01
Empirical sweep (3,600 sims)	Bear 3/14/2/1	1856	∼10 min compute

When game mechanics are transparent, LLM reasoning appears highly cost-effective. The critical factor is that Moreau's damage formula is explicitly provided—the agent can compute expected damage output directly.

6 Experimental Results: 10-Animal Arena

Section 5 reported preliminary results with 6 animals and single-build evaluation. Following the expansion to 10 animal species (adding Snake, Eagle, Fox, Scorpion—each with fully implemented ability effects), we conducted three experiments to evaluate Moreau Arena at increased strategic complexity.

6.1 Balance Tournament

Setup. 21 builds across 10 animals, round-robin format, 50 matches per pair, 10,500 total matches, deterministically seeded.

Table 6: Top 8 builds by Elo rating in the 10-animal balance tournament.

Rank	Build	Animal	Stats	Elo	WR
1	tiger_balanced	Tiger	5/6/5/4	1667	60.6% ± 6.9%
2	bear_tank	Bear	7/5/4/4	1623	59.1% ± 6.9%
3	eagle_balanced	Eagle	5/7/5/3	1622	51.8% ± 7.1%
4	bear_brawl	Bear	5/7/5/3	1620	59.1% ± 6.9%
5	boar_stampede	Boar	5/6/5/4	1604	54.4% ± 7.0%
6	eagle_glass	Eagle	3/9/5/3	1600	59.2% ± 6.9%
7	buffalo_balanced	Buffalo	6/5/5/4	1596	55.7% ± 7.0%
8	fox_trick	Fox	5/6/6/3	1524	55.9% ± 7.0%

The Elo spread from rank 1 to rank 21 is 367 points (1667 to 1300), indicating meaningful differentiation across the build space. The animal delta (best minus worst win rate) is 0.167, within our design target of ≤ 0.20. No single build exceeds 70% win rate.

Table 7: Aggregate per-animal win rates in the 10-animal arena.

Animal	WR	Δ from 50%
Bear	59.1%	+9.1%
Tiger	59.1%	+9.1%
Eagle	55.5%	+5.5%
Boar	53.9%	+3.9%
Scorpion	53.3%	+3.3%
Buffalo	45.6%	−4.4%
Fox	45.4%	−4.6%
Wolf	43.9%	−6.1%
Snake	43.8%	−6.2%
Monkey	42.4%	−7.6%

6.2 Draft System Evaluation

Setup. 4 agents (SmartDraft with counter-pick logic, StaticDraft, GreedyDraft, RandomDraft), 2 bans per side, blind pick, best-of-5, 20 series per pair, 120 total series across 10 animals.

Table 8: Draft tournament results.

Rank	Agent	Elo	Series W/L
1	SmartDraft	1608	38W/22L
2	StaticDraft	1561	37W/23L
3	GreedyDraft	1514	34W/26L
4	RandomDraft	1317	11W/49L

Table 9: Per-game win rates across a best-of-5 series, showing adaptation dynamics.

Agent	G1	G2	G3	G4	G5
SmartDraft	47%	63%	61%	65%	60%
StaticDraft	63%	55%	55%	52%	75%
GreedyDraft	67%	57%	55%	47%	44%
RandomDraft	23%	25%	29%	32%	25%

SmartDraft starts weak in G1 (47%) but improves to 60–65% in G2–G5, suggesting that the adaptation mechanic rewards agents that learn from opponent builds. GreedyDraft shows the opposite pattern—strong early (67% G1) but declining to 44% in G5 as opponents adapt to its predictable selections.

6.3 LLM Philosophy Agents—10-Animal Draft

Setup. Three Claude Sonnet 4.6 sub-agents with distinct philosophies, plus two baselines, competing in 10-animal draft format (1 ban each, blind pick, best-of-5, 200 total series).

Table 10: LLM philosophy agent draft results.

Rank	Agent	Elo	Series W/L
1	offense_claude	1685	52W/28L
2	greedy_draft	1607	58W/22L
3	defense_claude	1555	48W/32L
4	speed_claude	1391	32W/48L
5	random_draft	1264	10W/70L

Non-transitive dynamics. The head-to-head results reveal a non-transitive matchup structure:

Table 11: Pairwise win rates between philosophy agents, showing non-transitivity.

Matchup	Win Rate
offense_claude vs defense_claude	65%–35%
offense_claude vs speed_claude	**40%–60%**
defense_claude vs speed_claude	**90%–10%**

A non-transitive pattern emerges: offense beats defense, speed beats offense, and defense strongly counters speed. The 10-animal meta creates a rock-paper-scissors dynamic absent from the 6-animal version (where offense dominated all strategies).

6.4 Evolution from 6-Animal to 10-Animal Benchmark

Table 12: Comparison of benchmark properties at 6-animal and 10-animal scale.

Metric	6 Animals	10 Animals	Interpretation
Top build WR	85.2%	60.6%	Better balance
Elo spread (builds)	1160 pts	367 pts	More compressed
Elo spread (agents)	495 pts	421 pts	Still discriminates
Non-transitive	No	Yes	Richer landscape
Draft differentiation	61 pts	291 pts	Draft becomes meaningful
Adaptation signal	Weak	Strong	G1 $\rightarrow$ G5 divergence
Viable animals	2	7+	Diverse meta

7 Multi-Model Tournament Results

7.1 Tournament 001: Blind Pick (One-Shot)

Setup. 13 agents (8 LLM + 5 baseline), best-of-7 series, 10 series per pair, 779 series completed (1 error), blind pick format. LLM agents receive game rules and must select a build without seeing opponent information.

LLM models: Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, GPT-5.2, GPT-5.2-Codex, Gemini 3 Flash, Gemini 3.1 Pro, Grok 4-1 Fast Reasoning.

Baseline agents: SmartAgent (heuristic counter-pick), HighVarianceAgent, ConservativeAgent (Buffalo 10/8/1/1), GreedyAgent (Boar 8/8/3/1), RandomAgent.

Table 13: Tournament 001 Bradley–Terry rankings (95% CI).

Rank	Agent	BT Score	CI Lower	CI Upper	Games
1	SmartAgent	1.0000	0.6100	1.0000	120
2	HighVarianceAgent	0.9541	0.5546	1.0000	120
3	ConservativeAgent	0.6426	0.3718	0.9170	120
4	gemini-3-flash-preview	0.5028	0.2669	0.7714	120
5	gemini-3.1-pro-preview	0.4553	0.2465	0.6782	119
6	grok-4-1-fast-reasoning	0.4373	0.2298	0.6645	119
7	GreedyAgent	0.2388	0.1277	0.3456	120
8	gpt-5.2	0.2296	0.1217	0.3425	120
9	gpt-5.2-codex	0.1193	0.0665	0.1675	120
10	claude-sonnet-4-6	0.0656	0.0336	0.0973	120
11	claude-haiku-4-5	0.0564	0.0301	0.0805	120
12	claude-opus-4-6	0.0173	0.0094	0.0242	120
13	RandomAgent	0.0054	0.0030	0.0078	120

Key finding: Baselines dominated LLMs. In T001's blind-pick format, handcrafted baseline agents occupied the top 3 positions. All LLM agents ranked below ConservativeAgent (a fixed Buffalo 10/8/1/1 build). Claude Opus placed 12th out of 13—just above RandomAgent—due to locking onto a Wolf 3/8/1/8 build (WIL=8, SPD=1) that was the worst-performing build in the

tournament. This result reveals that the T001 prompt contained incorrect formulas for dodge and resist, causing LLMs to make suboptimal stat allocations.

7.2 Tournament 002: Adaptive Play

Setup. Same 13 agents, best-of-7 series, 10 series per pair, 780 series completed (0 errors), *adaptive* format: after each game, the loser sees the winner's build and can change their own build; the winner's build is locked.

Prompt improvements for T002:

- Exact combat formulas (corrected dodge and resist)
- Meta context: top-performing builds from T001
- Opponent build reveal for adaptation
- Structured JSON output (model-native schemas for each provider)

Table 14: Tournament 002 Bradley–Terry rankings (95% CI).

Rank	Agent	BT Score	CI Lower	CI Upper	Games
1	gpt-5.2-codex	1.0000	0.7925	1.0000	120
2	gemini-3-flash-preview	0.6793	0.3930	1.0000	120
3	grok-4-1-fast-reasoning	0.6496	0.3656	1.0000	120
4	claude-opus-4-6	0.2520	0.1429	0.3934	120
5	claude-sonnet-4-6	0.2520	0.1297	0.4266	120
6	gpt-5.2	0.2520	0.1398	0.4217	120
7	claude-haiku-4-5	0.2258	0.1269	0.3592	120
8	gemini-3.1-pro-preview	0.1882	0.1043	0.3108	120
9	SmartAgent	0.0926	0.0490	0.1465	120
10	HighVarianceAgent	0.0698	0.0379	0.1081	120
11	ConservativeAgent	0.0614	0.0308	0.1029	120
12	GreedyAgent	0.0426	0.0209	0.0761	120
13	RandomAgent	0.0066	0.0038	0.0101	120

Key finding: LLMs surpassed baselines. The T002 prompt improvements—exact formulas, meta context, and opponent build reveal—completely inverted the hierarchy. GPT-5.2-Codex rose from rank 9 to rank 1; Claude Opus rose from rank 12 to rank 4. All baseline agents dropped to ranks 9–12. The prompt engineering delta was dramatic: Codex moved from BT 0.12 to BT 1.00.

7.3 T001 → T002 Comparison

Table 15: Rank changes from Tournament 001 to Tournament 002.

Agent	T001 Rank	T002 Rank	Δ
gpt-5.2-codex	9	1	+8
claude-opus-4-6	12	4	+8
grok-4-1-fast	6	3	+3
claude-sonnet-4-6	10	5	+5
gpt-5.2	8	6	+2
claude-haiku-4-5	11	7	+4
gemini-3-flash	4	2	+2
gemini-3.1-pro	5	8	−3
SmartAgent	1	9	−8
HighVarianceAgent	2	10	−8
ConservativeAgent	3	11	−8
GreedyAgent	7	12	−5
RandomAgent	13	13	0

7.4 Adaptation Analysis

Tournament 002's adaptive format enables measurement of counter-pick behavior.

Table 16: Per-agent adaptation metrics in Tournament 002.

Agent	Times Lost	Adapted	Rate	Unique Builds
gpt-5.2-codex	150	137	91%	63
gpt-5.2	211	188	89%	45
claude-haiku-4-5	232	191	82%	32
grok-4-1-fast	174	130	75%	32
claude-sonnet-4-6	253	166	66%	29
claude-opus-4-6	219	116	53%	6
SmartAgent	275	110	40%	5
gemini-3.1-pro	201	79	39%	23
gemini-3-flash	151	58	38%	17

Table 17: Counter-pick success rate: win rate after adapting vs. not adapting.

Agent	Adapted WR	Not Adapted WR	Δ
gemini-3-flash	76%	37%	+39%
claude-opus-4-6	54%	31%	+23%
gemini-3.1-pro	35%	12%	+23%
grok-4-1-fast	64%	43%	+21%
gpt-5.2-codex	61%	46%	+15%
SmartAgent	39%	24%	+15%
gpt-5.2	44%	35%	+9%
claude-sonnet-4-6	63%	55%	+7%
claude-haiku-4-5	36%	46%	−10%

Adaptation provides the largest benefit to models with strong counter-pick reasoning (Gemini Flash: +39%, Grok: +21%) and smaller benefit to models that adapt frequently but less effectively (Haiku: −10%, suggesting over-adaptation).

7.4.1 WIL Trap Escape: Claude Opus

In T001, Claude Opus was locked on Wolf 3/8/1/8 (WIL=8, SPD=1)—the worst-performing build in the tournament. The T002 prompt improvements helped Opus escape this trap:

Table 18: Claude Opus build evolution from T001 to T002.

Metric	T001	T002
Primary Animal	wolf (100%)	bear (88%), buffalo (12%)
Avg WIL	7.9	1.0
Avg ATK	8.1	8.5
Unique Builds	2	6

Opus shifted from wolf to bear/buffalo with minimal WIL allocation, demonstrating that the corrected formulas in the T002 prompt were sufficient to overcome the reasoning failure that caused the WIL trap.

8 Discussion

8.1 What Moreau Arena Measures

Our results suggest that Moreau Arena measures a specific cognitive capability we term *strategic variable identification*—the ability to determine which variables in a multi-variable optimization problem contribute most to the objective function. In the 6-animal experiment (Section 5), the offense agent's critical insight was recognizing that ATK has super-linear returns through ability synergy, making it the dominant variable. In the 10-animal experiment, this capability extends to matchup reasoning: understanding not just which build is globally strongest, but which build counters a specific opponent.

The distinction between strategic variable identification and general knowledge recall is important. On MMLU, a model that has memorized the correct answer scores identically to one that

reasons about it. On Moreau Arena, a model that memorizes "Bear 4/14/1/1 is optimal" fails when the meta shifts via the auto-balancer or when facing a counter-pick in draft format. The benchmark rewards *reasoning about* strategy rather than *recalling* strategy.

8.2 The Non-Transitivity Result

The non-transitive matchup structure observed in the 10-animal experiment (offense beats defense 65–35, speed beats offense 60–40, defense beats speed 90–10) has important implications for benchmark design.

First, non-transitivity means that **no single agent can achieve 100% win rate**, even with perfect play. This provides natural headroom as models improve—unlike MMLU's 90%+ saturation, Moreau Arena's theoretical ceiling is bounded by the meta's rock-paper-scissors structure.

Second, non-transitivity supports the draft system as an important component. In single-build evaluation, the offense agent dominates. In draft format, speed's counter-pick advantage against offense becomes expressible, creating richer signal about adaptive reasoning.

Third, the 90–10 defense-speed matchup reveals that non-transitivity can be asymmetric, reminiscent of the payoff asymmetries observed in poker [Brown & Sandholm, 2019].

8.3 The Self-Evolving Meta as a Benchmark Property

If Moreau Arena's optimal build is discoverable (and our results show it is), then future agents can simply memorize it without engaging in genuine reasoning. The auto-balancer prevents this by shifting the meta after sufficient data accumulates. This creates a benchmark that functions like a *standardized test with rotating question pools* rather than a fixed exam.

8.4 Limitations

Philosophy priming confound. Our multi-agent protocol assigns strategic philosophies as prompts. This controls for reasoning direction but confounds innate model capability with prompt quality. Future work should include prompt-free evaluation.

Small sample sizes in draft experiments. The 20 series per pair provides sufficient signal for Elo estimation (±30 points) but limits ability to detect subtle adaptation patterns. Scaling to 100+ series per pair would enable per-game-number analysis with tighter confidence intervals.

10 of 14 animals experimentally validated. While 14 animals are fully implemented (with 940+ automated tests), the experiments cover 6 and 10 animals respectively. The remaining 4 animals (Crocodile, Raven, Shark, Owl) await inclusion.

Heuristic combat AI. The current combat AI uses a fixed priority-based decision system. This means the combat layer tests build design quality rather than in-combat tactical reasoning. A future iteration could expose tactical decisions to the agent's API.

8.5 Future Work

Frontier model comparison. The immediate next step is running the full 14-animal draft benchmark across all frontier models to produce cross-model Elo ratings.

Agent-controlled active abilities. Exposing combat-time ability decisions to the LLM agent via a turn-based API would test tactical reasoning alongside strategic build design.

Persistent memory evaluation. Evaluating whether LLM agents can maintain and leverage persistent memory across multiple series would test a capability increasingly important in agentic deployments [Park et al., 2023].

Cross-season evaluation. Running agents across multiple 14-day seasons would test whether agents can detect and adapt to meta shifts—a capability with no analogue in current static benchmarks.

Human baselines. Competitive strategy game players could provide calibration points, establishing where human strategic reasoning falls on the Elo scale relative to LLM agents.

We also plan to release a public challenge platform where users can submit custom agent configurations and compete on a live leaderboard, providing crowd-sourced data for future analysis of prompt sensitivity and strategic adaptation.

9 Conclusion

We presented Moreau Arena, a controlled strategy benchmark for evaluating agentic AI reasoning. The benchmark isolates strategic decision-making from perceptual or execution confounds by presenting agents with a transparent optimization problem: allocate 20 stat points across four attributes and select from 14 animal species with distinct ability sets, then let deterministic combat resolve the outcome.

Our key contributions are:

1. **A benchmark that measures strategic reasoning.** Moreau Arena evaluates *strategic variable identification*—the ability to determine which variables in a constrained optimization matter most. In our experiments, reasoning quality correlated directly with competitive Elo: the agent that correctly identified the damage formula as the dominant variable achieved Elo 1874, while agents constrained by suboptimal philosophies scored 373–495 points lower.

2. **A draft system that rewards adaptation.** The best-of-5 format with bans, blind picking, and counter-picking transforms creature design from a one-shot optimization into a sequential strategic game. Adaptive agents improve from 47% in game 1 to 65% in game 4, while static agents decline from 67% to 44%—to our knowledge, the first such adaptation measurement in an LLM benchmark.

3. **A self-evolving seasonal meta.** The auto-balancer monitors per-animal win rates via exponential moving averages and applies calibrated adjustments within 14-day seasons. This creates a benchmark that resists strategy memorization across evaluation epochs (Section 3.6).

4. **Non-transitive dynamics at scale.** The 10-animal arena produces emergent rock-paper-scissors matchups (offense $>$ defense $>$ speed $>$ offense), providing natural headroom beyond saturation. No single build exceeds 70% win rate in the balanced meta, and the animal delta stays within 0.20.

5. **Multi-model tournament validation.** Two tournaments with 13 agents across 1,559 series demonstrate that prompt engineering (exact formulas, meta context, structured output) shifts LLM agents from below-baseline to above-baseline performance, and that adaptation provides measurable competitive advantages for top-tier models.

6. **An AI-collaborative research methodology.** The Round Table approach—where human researchers and multiple AI instances contribute as intellectual partners—enabled rapid iteration from concept to a 940-test, 10,000+ match/minute benchmark.

Moreau Arena processes 10,000+ matches per minute on consumer hardware, runs deterministically across platforms, and is open-source. We release the simulator source code, agent implementations, tournament data (JSONL), and analysis code at https://github.com/supervitek/moreau-arena-paper to support reproducibility and independent verification. We believe it fills a needed gap in the AI evaluation landscape: a benchmark complex enough to require genuine strategic reasoning, simple enough to be fully transparent, and dynamic enough to resist strategy collapse as models improve.

10 Ethics and Broader Impact

10.1 Intended Use and Dual Use

Moreau Arena is designed as an evaluation tool for AI systems, not as a competitive game targeting human players. The benchmark measures strategic reasoning capabilities to inform AI development and safety research. The specific capability measured (stat allocation in a game context) is far removed from harmful applications. The transparent, deterministic design is deliberately chosen to *measure* rather than *train* adversarial reasoning.

10.2 Responsible Benchmarking

Evaluation-only design. Moreau Arena's combat engine is deterministic and seed-driven, meaning match outcomes are fully determined by build choices and the random seed. This makes the benchmark unsuitable as a training environment: there is no gradient signal to exploit, no reward shaping to game, and no environment to overfit to. Models interact through a single prompt-response cycle per decision.

Transparent scoring. All ranking computations (Bradley–Terry, Elo) use published formulas with fixed hyperparameters (Table 3). Bootstrap confidence intervals accompany every point estimate. The full JSONL match record is published, enabling independent verification and alternative ranking methodologies.

Baseline calibration. Every tournament includes deterministic baseline agents (RandomAgent, GreedyAgent, SmartAgent, ConservativeAgent, HighVarianceAgent) that establish floor and ceiling performance bounds. These baselines are not LLM-powered, so their performance is invariant across runs, providing stable reference points.

Minimal computational burden. At 10,000+ matches per minute on consumer hardware (CPU only, no GPU required), Moreau Arena has minimal environmental footprint. A complete cross-model evaluation requires approximately $2 in API costs.

10.3 Goodhart's Law Resistance

Moreau Arena incorporates three defenses against Goodhart's Law:

Dynamic meta-balancing. The auto-balancer monitors per-animal win rates and applies calibrated nerfs and buffs. If a model memorizes "Bear 4/14/1/1 is optimal," the auto-balancer will eventually reduce Bear's proc rates, making that cached solution suboptimal.

Seasonal rotation. The offline patch generator produces frozen season configurations that capture a snapshot of the meta at a point in time. A model trained on Season 1 data may find its cached strategies degraded in Season 2.

Non-transitive dynamics. The 14-animal roster produces emergent rock-paper-scissors matchup structures. No single build achieves universal dominance—the correct answer depends on the oppo-

nent's choice, making Goodhart optimization toward a single "best answer" fundamentally impossible.

10.4 Contamination Prevention

Novel game mechanics. All 14 animals, their abilities, stat formulas, and combat mechanics are original creations with no counterpart in existing games or datasets.

Seed-driven reproducibility without answer leakage. Match outcomes depend on the random seed. Publishing seeds enables reproduction but does not leak "answers"—the correct *build choice* is independent of the seed.

Rotating evaluation surfaces. The seasonal rotation system ensures that the optimization landscape changes between evaluation epochs.

Prompt-only interface. Models interact through a single text prompt. The prompt template is versioned and hashed, enabling detection of prompt-level contamination.

Config integrity verification. Every season config includes a SHA-256 hash computed over all game parameters. The `load_season()` function verifies this hash at load time.

10.5 Intent-Neutral Safety Observations

When evaluating LLM strategic reasoning, we adopt intent-neutral language throughout our analysis. We describe *observed behavioral patterns*—win rate trends, adaptation curves, build diversity metrics—without attributing cognitive states such as "the model understood," "the model intended," or "the model deceived." We report adaptation scores (Game 5–7 win rate minus Game 1–2 win rate) as quantitative measures without interpreting the underlying mechanism.

10.6 AI Collaboration and Data Ethics

The Round Table methodology treats AI models as collaborative intellectual partners. We document this collaboration transparently: all AI contributions are disclosed, and the paper is co-authored by human and AI participants. All match seeds, build configurations, and tournament results are published alongside the source code. No personal data is collected or processed. The benchmark contains no copyrighted game content—all game mechanics are original.

References

Bakhtin, A., Brown, N., Dinan, E., et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Brown, N. & Sandholm, T. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019.

Chen, M. et al. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374*, 2021.

Chiang, W.-L. et al. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *ICML 2024*, 2024.

Cobbe, K. et al. Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*, 2021.

Duan, H. et al. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. *NeurIPS 2024*, 2024.

Elo, A. E. *The Rating of Chessplayers, Past and Present.* Arco Publishing, 1978.

Hendrycks, D. et al. Measuring Massive Multitask Language Understanding. *ICLR 2021*, 2021.

Jimenez, C. E. et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *ICLR 2024*, 2024.

Küttler, H., Nardelli, N., Miller, A.H., et al. The NetHack Learning Environment. *NeurIPS 2020*, 2020.

Li, Y., Lin, C., Nasir, M.U., et al. GVGAI-LLM: Evaluating Large Language Model Agents with Infinite Games. *arXiv:2508.08501*, 2025.

Light, J. et al. AvalonBench: Evaluating LLMs Playing the Game of Avalon. *NeurIPS 2023 Workshop*, 2023.

Lin, W., Roberts, J., Yang, Y., et al. GAMEBoT: Transparent Assessment of LLM Reasoning in Games. *ACL 2025*, 2025.

Park, J. S. et al. Generative Agents: Interactive Simulacra of Human Behavior. *UIST 2023*, 2023.

Shinn, N. et al. Reflexion: Language Agents with Verbal Reinforcement Learning. *NeurIPS 2023*, 2023.

Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

Silver, D. et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Vinyals, O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019.

Wei, J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS 2022*, 2022.

Xu, Y. et al. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. *arXiv:2309.04658*, 2023.

Wang, G. et al. Voyager: An Open-Ended Embodied Agent with Large Language Models. *ICLR 2024*, 2024.

Yao, S. et al. ReAct: Synergizing Reasoning and Acting in Language Models. *ICLR 2023*, 2023.

Zheng, Y. et al. LM Fight Arena: Benchmarking Large Multimodal Models via Game Competition. *arXiv:2510.08928*, 2024.

Zhong, Y. et al. A Survey on Large Language Model-based Game Agents. *arXiv:2404.02039*, 2024.

A Complete Animal Roster

Table 19: All 14 animal species in Moreau Arena with passive traits and triggered abilities. Experiments in Sections 5–7 use the first 10 animals; the remaining 4 (marked with *) are fully implemented and tested but await benchmark evaluation.

Animal	Passive Trait	Ability 1	Ability 2	Proc Tier
Bear	Fury Protocol	Berserker Rage (+60% dmg)	Last Stand (2x at <15% HP)	Strong
Tiger	Ambush Wiring	Pounce (+70% dmg + stun)	Hamstring (−55% SPD)	Standard
Wolf	Pack Sense	Pack Howl (+ATK buff)	Rend (bleed DoT)	Standard
Monkey	Primate Cortex	Chaos Strike (random dmg)	Mimic (copy ability)	Strong
Buffalo	Thick Hide	Iron Will (dmg reduction)	Fortify (+armor)	Strong
Boar	Charge	Stampede (AoE charge)	Gore (ignore dodge)	Strong
Snake	Venom Glands	Venom (poison DoT)	Coil (constrict)	Standard
Eagle	Aerial Strike	Dive (ranged burst)	Keen Eye (+accuracy)	Standard
Fox	Cunning	Evasion (+dodge)	Trick (debuff)	Standard
Scorpion	Paralytic Sting	Sting (paralyze)	Exoskeleton (+armor)	Standard
Crocodile*	Death Roll	Death Roll (grapple)	Ambush (surprise atk)	Standard
Raven*	Omen	Shadow Clone (decoy)	Hex (curse)	Standard
Shark*	Blood Frenzy	Blood Frenzy (+dmg low HP)	Jaw Snap (execute)	Strong
Owl*	Night Vision	Foresight (predict)	Silent Strike (guaranteed hit)	Standard