FM-Bench: A Benchmark for Long-Horizon Management with Competing Agents

Tianyou Wang[1,*], **Chongyang Gao**[1,*], **Kezhen Chen**[1], **Dong Chen**[1], **Yinghao He**[1], **Donghan Li**[1], **Wangcheng Xu**[1], **Hongjiu Zhang**[1], **Chi Li**[1]

[1]AnalogyAI
[*]Equal contribution

Language model agents now execute bounded tasks reliably. Whether they can sustain effective decision-making over long horizons, where actions have cumulative consequences and the environment responds to their choices, remains largely unmeasured. FM-Bench (Football Management Benchmark) measures this. An LLM agent runs a football club for 20 in-game years through 26 tools and roughly 340 to 400 decision stops, each admitting as many tool calls as the agent chooses to spend. It drafts a squad on the same fixed budget as every rival, trades players, negotiates contracts, invests in facilities and youth, sets lineups, and answers to a board that can fire it. A deterministic engine grades the run, accumulating across years into one final score with no LLM judge or human rater. Evaluation runs in two modes. The solo track plays each of 15 frontier models against a frozen scripted world, and the Arena places the same models plus a scripted anchor in one shared 20-year world; to our knowledge, the first head-to-head evaluation at this scale and horizon. We measure six behavioral capabilities behind its score. Across three seeds, all 15 models complete every horizon while the blind scripted baselines die out in most of theirs, and `claude-fable-5` tops the solo board on mean score and the Arena, where the league title nonetheless rotates among ten models. Neither scale, price, nor vendor predicts the order; the order settles only late in the horizon, and the best first-play human lands only at the bottom of the model board. What separates the models is managerial behavior rather than computation. Higher-scoring models reduce slow-payoff investment as the horizon approaches its end, keep cash invested rather than idle, and open contract renewals well before the deadline, while token spend predicts nothing. Furthermore, no model learns the market's hidden prices from hundreds of rejected bids, and self-managed memory fails in two opposite modes: an archive that only grows or a plan rewritten every season.

Date: 2026-08-18
Correspondence: Chi Li at chili@analogyai.org
Code: https://github.com/Analogy-AI/fm-bench

1 Introduction

Language model agents now handle bounded tasks reliably: they resolve real GitHub issues (Jimenez et al., 2024), solve competition-grade reasoning problems (Gao et al., 2026; Lightman et al., 2024; Balunovic et al., 2026), complete computer-use workflows (Zhou et al., 2024; Xie et al., 2024), and orchestrate multi-step tool calls (Liu et al., 2024; Qin et al., 2024; Yao et al., 2024). These tasks have a short horizon to the correct answer and no competing agents. Newer benchmarks push one dimension of realism, the long horizon, stretching execution to hundreds of steps (Li et al., 2026), to reactive game worlds (Küttler et al., 2020; Hafner, 2021; Côté et al., 2018; Paglieri et al., 2025), and to business simulations that run a vending machine or shop for simulated months (Backlund and Petersson, 2025; Shi et al., 2026), a startup for a year (He et al., 2026; Chen et al., 2026), or a company's finances through an eleven-year macro cycle (Han et al., 2026). Others push the second dimension, competing agents, but stage cooperation and competition over short episodes (Zhu et al., 2025; Wang et al., 2024) or rank models by bluffing and bargaining inside a single session (Müller and Müller, 2026).

However, in real-world scenarios, these benchmarks still fall short of combining both dimensions at once and

of providing an organization to keep alive through them. This combination is what we call *management*, running an organization against rivals pursuing the same opportunities, over a horizon so long that early decisions reshape the world, with performance judged cumulatively under the conditions those decisions created. Management in this sense makes four demands on an agent. *Hidden information*: the state that matters is never fully observable, so valuation must stay calibrated against permanent noise. *Cumulative consequences*: the best actions pay off years later, losing positions compound yet remain recoverable, and honors accrue into the final score each season. *A counter-adaptive environment*: other actors respond to revealed behavior, so no fixed strategy stays a best response and a winning move decays with use. *Multi-objective pressure*: results and financial discipline are judged jointly, and satisfying one constraint can violate another.

To measure management so defined, we present FM-Bench (Football Management Benchmark), where an LLM agent runs a football club for 20 in-game years under all four demands instantiated by concrete game mechanisms (§3.1, Figure 1). Through 26 schema-generated tools over roughly 340 to 400 decision stops, it drafts a squad on the same fixed budget as every rival, trades players, negotiates contracts, invests in facilities and youth, fields lineups, and answers to a board that can fire it, while a deterministic engine accumulates every quantity into one final score with no LLM judge or human rater. Each stop is a fresh conversation whose only carried state is a self-authored notebook, so memory curation is itself measured. The *solo track* plays each of 15 flagship models against a frozen scripted world with tiered opponents. The *Arena* places all 16 seats in one shared 20-year world where every signing denies a rival the same player, made a valid measurement by an equal-endowment draft, sealed-bid conflict resolution, and capped revival. Across three seeds, every model completes the horizon, while the blind scripted anchors, including a disciplined heuristic, die out in 7 of their 9 runs. `claude-fable-5` tops both tracks, reaching about 95% of a scripted upper anchor allowed to read the hidden state (90.94 against 95.54), yet competition still reshuffles the board, with the league title rotating among ten models and mid-board solo standings not surviving adaptive rivals. We decompose the scalar score into six behavioral capabilities, of which reducing slow-payoff investment near the horizon (Spearman -0.58), keeping cash deployed rather than idle (-0.50), and opening contract renewals early ($+0.45$) track the score on every seed, while token spend is uncorrelated under every accounting. Hundreds of rejected bids never teach a model where the market's true prices lie, and notebooks fail in two opposite regimes, append-only archives and wholesale rewrites.

Our contributions:

1. **FM-Bench, a benchmark for long-horizon management with competing agents.** A deterministic 20-year football-management environment that instantiates all four demands: hidden information, cumulative consequences, a counter-adaptive market, and multi-objective pressure. Player ability stays permanently hidden behind biased scouting, consequences accumulate across roughly 340 to 400 decision stops with every season's honors accruing into the final score, the transfer market adapts against revealed strategies, and a board judges results and financial discipline jointly. Grading is continuous and cumulative, computed by the mechanics at every stop, with no LLM judge or human rater anywhere.
2. **Two tracks on one engine, isolating competition.** The solo track plays each model against a frozen scripted world, and the Arena, to our knowledge the first evaluation to run 15 frontier models as competing agents in one persistent economy over a 20-year horizon, places the same models in one shared world. Six first-play humans ran the same track; four died out and the best finished at the bottom of the model board. Their strengths mirror the models' weaknesses, making human–agent collaboration on this track a measurable, promising future direction.
3. **Long-horizon behavioral analysis.** We turn the scalar into a per-model profile over six behavioral capabilities. Strong models reduce slow-payoff investment before the horizon ends, keep cash invested rather than idle, and open contract renewals ahead of the deadline; hundreds of rejected bids never teach a model the market's true prices, notebooks fail as append-only archives or wholesale rewrites, and token spend predicts nothing. The profile explains what a leaderboard cannot. The winner is a generalist, the mid-board pairs one strength with one decisive gap, and the bottom fails on several axes at once.

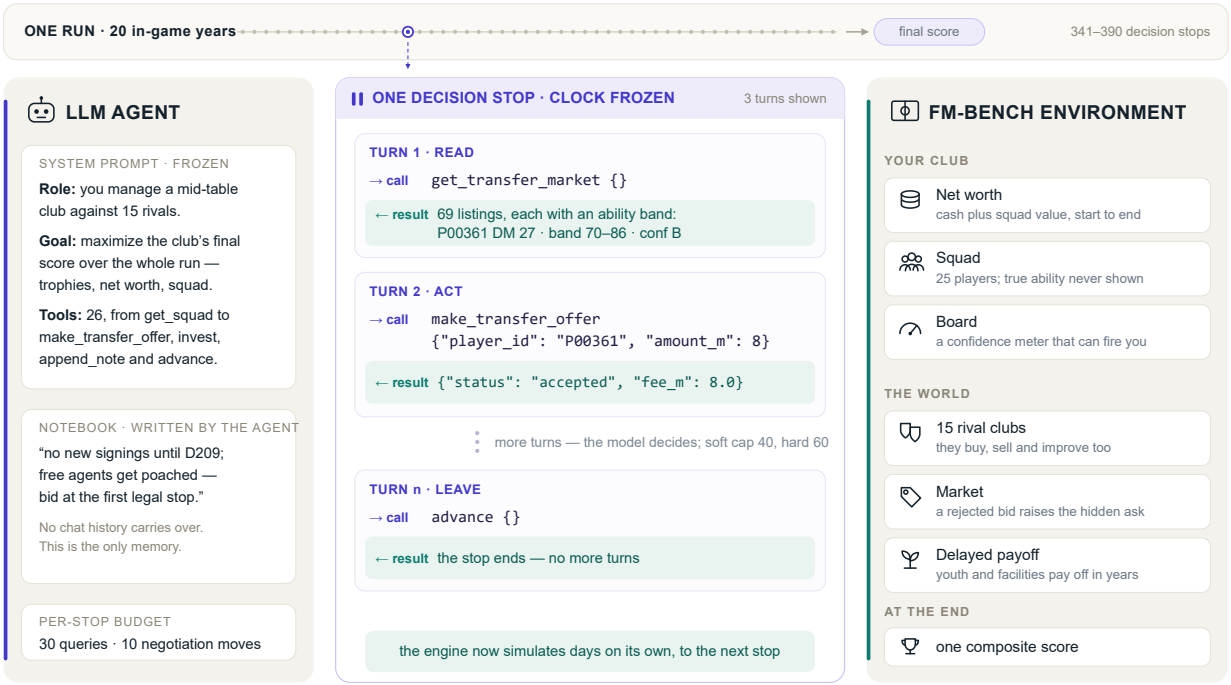

Figure 1 How a run works. One run spans 20 in-game years and roughly 340 to 400 decision stops, ending in a single composite score. At each stop, the clock freezes and the agent takes as many tool-call turns as it wants (read → act → `advance`). The agent's only cross-stop memory is its self-written notebook (§3.2); the environment side holds the club (net worth, squad with permanently hidden true ability, a board that can fire the manager) and the world (15 rival clubs, a counter-adaptive market, delayed-payoff investments).

2 Related Work

Bounded tasks. General agent benchmarks such as AgentBench (Liu et al., 2024) and tool-orchestration suites such as ToolBench (Qin et al., 2024) and τ-bench (Yao et al., 2024) measure whether a model can select and sequence tool calls correctly. WebArena (Zhou et al., 2024) and OSWorld (Xie et al., 2024) ground the same question in realistic software, where the difficulty is largely perceptual. SWE-bench (Jimenez et al., 2024) and GAIA (Mialon et al., 2024) grade one-shot deliverables. In all of these, episodes are short, the environment is stationary, and mistakes rarely compound. FM-Bench keeps the same tool-calling interaction shape but removes the grounding burden and stretches the horizon to hundreds of decision stops in which early errors compound for years.

Long-horizon evaluation. Terminal-agent marathons (Li et al., 2026; Cheng et al., 2026), UltraHorizon (Luo et al., 2025), and game environments (Küttler et al., 2020; Hafner, 2021; Côté et al., 2018; Paglieri et al., 2025) stretch episodes to hundreds of steps, but the former pursue a fixed deliverable in a stationary world and the latter reward reactive skill rather than economic planning. Closest in spirit are business simulations that run a vending machine (Backlund and Petersson, 2025), an online shop (Shi et al., 2026), a startup (Chen et al., 2026; He et al., 2026), or a company's finances (Han et al., 2026) for months to years of simulated time; YC-Bench's strongest predictor of success, scratchpad use, matches the memory-curation capability of §4.2.1. Even there, the environment does not push back, as StockBench replays real markets the agent's actions never move (Chen et al., 2025) and CoffeeBench tests one model at a time against fixed reference agents (Sugiura et al., 2026). FM-Bench differs in combining a counter-adaptive environment, permanently hidden state, an explicit memory discipline test, and cumulative mechanism-computed scoring.

Table 1 The four demands: the mechanism that instantiates each, the ability it targets, and the calibration lever that tunes it.

demand	mechanism	targeted ability	calibration lever
Hidden information	scout bands with a permanent per-scout bias; hidden player traits (injury proneness, development rate, aging onset); hidden asks in negotiation	valuation calibration under noise	bandwidth, bias sigma
Cumulative consequences	Upside: youth development and facility investment pay off over years. Downside: insolvency ends in administration (points penalty, forced sales), confidence collapse ends in firing; both compound yet remain recoverable. Honors accrue into the final score season by season	long-range credit assignment; trend recognition, loss-cutting	growth and aging curves; spiral thresholds, warning cadence; honors weights
Counter-adaptive market	rejected bids raise the hidden ask; repeat-pair markups (bargains included); anti-inversion counters; negotiation cooldowns; per-seed mispricing	strategy adaptation	markup and decay rates
Multi-objective pressure	the board judges results and financial discipline jointly; season targets scale with squad strength	multi-constraint balancing	target cushion, board patience

Multi-agent evaluation and arenas. Arena-style rankings began with pairwise human-preference voting (Chiang et al., 2024), where the models never interact, and coordination benchmarks (Zhu et al., 2025; Wang et al., 2024) measure how well agent systems cooperate over short episodes rather than who manages better over time. Competition between models appears in Cattle Trade, single-session negotiation ranked by TrueSkill with no organization to run (Müller and Müller, 2026), and in Vending-Bench Arena, three models in a shared vending market documented only through a leaderboard page over a closed environment (Andon Labs, 2026). FM-Bench's Arena runs this line at a different scale, 15 frontier models plus a scripted anchor managing rival clubs for 20 seasons, with a matched solo track measuring the same model without live competition. Orthogonally, AgentAtlas (Mazaheri and Mazaheri, 2026) argues that outcome leaderboards conflate distinct competences and calls for the behavior-level decomposition that §4.2.1 provides.

3 FM-Bench

This section describes the task and the interaction loop, the information design and the four demands it instantiates (§3.1), the memory design that turns curation into a measured capability (§3.2), and the score (§3.3).

3.1 Environment Overview

The task. The agent manages one club in a 16-club professional division. Every club, the agent and all rivals alike, drafts its 25-player squad from the same template pool under the same fixed budget, and picks are independent copies, so no club's draft interferes with another's. From there, the agent runs the club for 20 in-game years: trading players, negotiating contracts, investing in facilities and youth, setting the lineups and tactics that decide its matches, and answering to a board that can fire it. A run ends in a single composite score that accumulates season by season (§3.3). Firing or insolvency does not end measurement but triggers capped revival at a score discount.

A deterministic world. Everything is generated deterministically from the seed, including the draft pool itself. The only non-seed inputs are the agents' own decisions, of which the draft picks are the first. The seed fixes one division, double round-robin with 30 rounds per season on a 360-day calendar, ∼470 fully simulated players alive at any time, the 16×25 senior rosters plus youth academies, and ∼2,000 distinct players over a 20-year run as youth intakes arrive and veterans retire. No real-world names or data appear anywhere, so memorized football knowledge cannot leak in.

The interaction loop and stop. The engine simulates day-by-day and pauses only at *decision stops*: 13 scheduled stops on a fixed season calendar, plus event-driven stops the world raises unsolicited, ∼6 per season for ≈19 stops per season in total. Completed 20-year runs log roughly 340-400 stops, with the exact count varying with play. Several types can fire at once, and a stop carries the full list of reasons the world stopped (Table 11, Appendix K). Every stop delivers the same packet: the list of stop types, a one-glance digest (league position, points, season target, cash, wage ratio, board confidence, injuries, next fixture, open window, and any active warnings), the inbox of everything that happened since the last stop (19 unsolicited event types, from incoming transfer offers and counter-bids to retirements and completed facilities) with action-required items flagged, live pending offers, the agent's last 10 state-changing actions, and its notebook (§3.2). The agent then takes as many tool-call turns as it wants, using query tools to inspect anything and action tools to decide, in any order, and finally calls `advance` to move time forward. The full 26-tool interface is listed in Table 10, Appendix K. Inaction is conservative, not fatal: ignored offers expire, lineups and tactics persist, open negotiations run until offers lapse after 10 game days, and up to 10 single-condition standing orders such as "accept offers above 2× value" execute autonomously between stops, deliberately too weak to replace per-stop judgment. Figure 1 shows the run / decision-stop / turn structure at a glance. Appendix L reproduces one stop verbatim as the agent sees it, and Table 13 in Appendix K gives the scale of one 20-year run by the numbers.

The four demands. The four demands that define management in §1 are not prompt dressing; each is a concrete mechanism with a targeted ability and a calibration lever (Table 1). *Hidden information*: a player's true ability, potential, and traits (injury proneness, development rate, aging onset, consistency, style fit) are revealed to no one, the owning club included. Scouting returns bands with a permanent per-scout bias, so estimates converge to "truth + your scout's bias" and trait estimation is statistical inference over seasons, not a lookup, while season stats, contracts, and finances stay exact. *Cumulative consequences*: youth development and facility investment pay off years later, insolvency and confidence collapse compound toward administration and firing yet remain recoverable, and honors enter the final score season by season. *Counter-adaptive market*: the market reacts to revealed behavior, rejected bids raise the hidden ask, repeated dealings with the same club raise its prices, and negotiation spam triggers cooldowns, so no fixed strategy stays a best response. *Multi-objective pressure*: the board judges results and financial discipline jointly, against season targets that scale with squad strength. Recorded runs show models failing against each demand. A mid-tier model let a player talk its wage offer up step by step, from 12 to 14 to 17 to 18M per year, accepting each demanded raise at face value because it could not judge what the player was truly worth (demand 1), and the same model went insolvent with wages above revenue (demand 2). Every model under-bid a market that raises its hidden ask after each rejection, needing a median of 30 offers per completed signing (demand 3, §4.2.1). A small model was fired by its board in every run (demand 4).

3.2 Memory Design

Memory in FM-Bench is the agent's responsibility, not a harness feature. Every decision stop opens as a fresh conversation with no chat history. Whatever the agent wants its future self to know, it must write down, so memory curation becomes part of what the benchmark measures. Four layers supply what the agent "remembers" (Table 12, Appendix K): the *packet auto-context*, pushed by the engine into every stop for free, covering current league position, cash, board confidence, injuries, offers on the table, and everything that happened since the last stop; the *archive*, past facts queryable on demand, from any season's final table to every honor and transfer of the run; the *audit log*, which records every action for replay and anti-cheat but is never shown to the agent; and the *notebook*, the only carrier of intent, holding the agent's own plans and reasons such as "why I bought him" or "sell Y in winter", written by the agent and injected back into every packet.

We chose this design for three reasons. *Scale*: a 20-year run generates far more history than any context window holds, so some forgetting policy is unavoidable. A rolling-context harness would impose one implicitly and untraceably, while the notebook makes it explicit, bounded, and the agent's own. *Comparability*: if the harness curated context by truncation, summarization, or retrieval over an external store, scores would partly measure the harness. Here every model faces the same contract, and what differs between models is only what they chose to write. *Measurement*: deciding what the future self will need is itself part of the capability

under test. The notebook then becomes analysis data, and the memory-curation metrics and the notebook audit of §4.2.1 are read directly off it.

3.3 Scoring

The score is built so that every season of a 20-year run counts. It aggregates quantities the game mechanics compute at every stop, honors earned in each season enter the final score directly, and no phase can be written off or made up by an endgame sprint. Notation: H = cumulative honors points; VA = net-worth *value added*; M = squad value; ρ = early-settlement discount. The score has a single source-of-truth implementation in the engine:

$$S_{\text{raw}} = \underbrace{18 \ln\!\left(1 + \tfrac{\max(H,0)}{60}\right)}_{\text{honors}} + \underbrace{10 \operatorname{sign}(VA) \ln\!\left(1 + \tfrac{|VA|}{40}\right)}_{\text{value added (primary)}} + \underbrace{6 \ln\!\left(1 + \tfrac{\max(M,0)}{80}\right)}_{\text{squad value (stabilizer)}}$$

$$S_{\text{final}} = \rho \cdot \max(S_{\text{raw}}, 0) + \min(S_{\text{raw}}, 0)$$

- H sums honors points over the whole run: champion +100, top-4 +20. Penalties offset honors inside the sum, but the channel argument is floored at zero, so honors debt cannot drive the term negative.
- $VA = W_{\text{real}} - W_{\text{baseline}}$, in M credits: W_{real} is the mean of the last three seasons' net-worth snapshots, deflated at 4%/yr, and W_{baseline} is the *same seed's* day-0 net worth. Zero reads "you handed the club back as valuable as you found it, in real terms." Net worth books cash, plus squad value at a 0.8 liquidity discount, plus facilities at 0.5 while under construction, minus outstanding transfer-fee liabilities. Cash above 3× the annual wage bill counts at only 0.3, an anti-hoarding discount.
- M is the last-3-season deflated mean squad value, a low-weight stabilizer so that a cash-hoarding squad-stripper cannot win on net worth alone.
- $\rho = 1$ for a completed run. For a run that settles early, by firing, insolvency, or abandonment, $\rho = 0.8 \cdot (t/T)^{1.5}$ and multiplies only the positive part, so dying early can never shrink a negative result, and there is no fire-to-cap-losses arbitrage. Humans abandoning mid-run settle by the same rule.

All three channels are log-compressed and uncapped, so no channel can run away, and the benchmark cannot saturate. The weights and divisors are calibration constants, not derived quantities. They were tuned on the free scripted anchors and frozen before any scored run, and every adjustment was accepted only if four criteria held. *Order preservation*: the anchor ladder stays monotonic, random < idle < heuristic < oracle, where the heuristic is a disciplined hand-written management script, and the oracle is the privileged truth-reading script of §4.1. *Discrimination*: the gap between disciplined and passive anchors stays wide. *Floor*: the oracle ceiling stays positive on every seed. *Non-saturation*: top scores keep separating instead of bunching at a ceiling. Appendix I states and proves the basic shape properties of the composite: no fire-to-cap arbitrage, channel monotonicity, fair-price transfer neutrality on the VA channel, and diminishing returns.

4 Experiments and Results

4.1 Evaluation Settings

Configuration. Every club starts as an identical shell with reputation 70, a 38,000-seat stadium, and a 100M-credit draft budget over the shared 50-template pool, and unspent budget plus a 50M stipend becomes opening cash. An episode spans 20 in-game years of 360 days each, one division of 16 clubs playing 30 rounds per season, and roughly 340 to 400 decision stops. All results in this paper come from this 20-year track on the tiered world, the solo runs on seeds 1, 1024, and 2026, and the Arena on seed 7, and the six human runs (§4.2) play the same track. Scores are comparable only within a track and calibration version. Every result file is stamped with the engine commit, params hash, score version, and the full knob set, and the ladder report warns loudly on mixed pools.

Table 2 The Solo track results. The four non-LLM seats are reference policies. The oracle reads true hidden state through the same interface (§4.1, Appendix H); heuristic is a disciplined blind script, idle never acts, and random acts arbitrarily (Appendix E).

seat	S_{final}	tokens
oracle (privileged)	95.54 ± 4.68	—
`claude-fable-5`	90.94 ± 5.20	24M
`kimi-k2.6`	88.49 ± 0.15	87M
`gpt-5.6-terra`	86.66 ± 1.20	28M
`gpt-5.6-sol`	86.40 ± 2.53	62M
`muse-spark-1.1`	83.19 ± 12.03	73M
`glm-5.2`	83.17 ± 0.92	58M
`grok-4.5`	81.82 ± 9.87	191M
`qwen3.7-max`	80.66 ± 10.38	47M
`deepseek-v4-pro`	79.15 ± 2.67	194M
`gemini-3-flash`	79.06 ± 13.45	51M
`claude-sonnet-5`	75.75 ± 5.03	39M
`claude-opus-4.8`	75.02 ± 2.46	31M
`gemini-3.5-flash`	74.59 ± 11.02	154M
`minimax-m3`	68.37 ± 12.33	74M
`claude-haiku-4.5`	36.90 ± 22.73	86M
heuristic	17.05 ± 12.34	—
idle	-0.90 ± 1.86	—
random	-17.21 ± 2.45	—

Models and harness. The roster covers 15 flagship models plus one scripted anchor. Ten are closed frontier models, four from Anthropic (`claude-fable-5`, `claude-opus-4.8`, `claude-sonnet-5`, `claude-haiku-4.5`), two from OpenAI (`gpt-5.6-sol`, `gpt-5.6-terra`), two from Google (`gemini-3.5-flash`, `gemini-3-flash`), `grok-4.5` from xAI, and `muse-spark-1.1` from Meta. Five are open-weight models (`deepseek-v4-pro`, `kimi-k2.6`, `qwen3.7-max`, `glm-5.2`, `minimax-m3`). The selection fixes within-family tier curves, closed against open frontier, and an agentic specialist (Table 5, Appendix C). Every model runs under our fixed harness with locked model ID, temperature, and prompt, at the provider's recommended full reasoning configuration, and the harness never handicaps a model's reasoning to save cost. The effective setting is stamped into every result, including any degradation to no-thinking, so reasoning-on and reasoning-off pools can never merge silently. Our runs are reasoning-on, and no custom harness is allowed in either mode.

Modes. The solo track plays one tested LLM against 15 scripted clubs and reports absolute S_{final}. The Arena places the 15 LLMs plus the scripted anchor in one shared world and ranks them by the same S_{final}. The equal-endowment draft, sealed-bid resolution, and capped revival remove seat confounds by mechanism, so a single shared run is valid on its own, and replication comes from seeds (§4.3). Every number in this paper comes from one stamped run per seat, with no run selection.

The oracle ceiling. The oracle is a scripted policy whose privilege is information, not power. It reads true hidden state directly from the engine, but every action still goes through the same 26-tool interface, per-stop budgets, and legality checks as any agent. It converts that access into exact transfer prices, exact reservation wages, true-potential academy picks, peak-timed sales, and a starting eleven re-solved from true ability, plus a survival controller that keeps it from ever being fired (Appendix H). Being a script rather than an optimum, the ceiling is soft in principle.

4.2 The 20-year solo track

The solo track runs one tested model against 15 scripted clubs. An equal-endowment draft builds every squad from the same 50-template pool on the same budget, conflicting transfer offers resolve as first-price sealed bids with undisclosed clearing prices, and capped revival converts firing or insolvency into at most three restarts, discounted by 0.8^k after k deaths, so a run reports full-horizon ability rather than a few unlucky seasons while

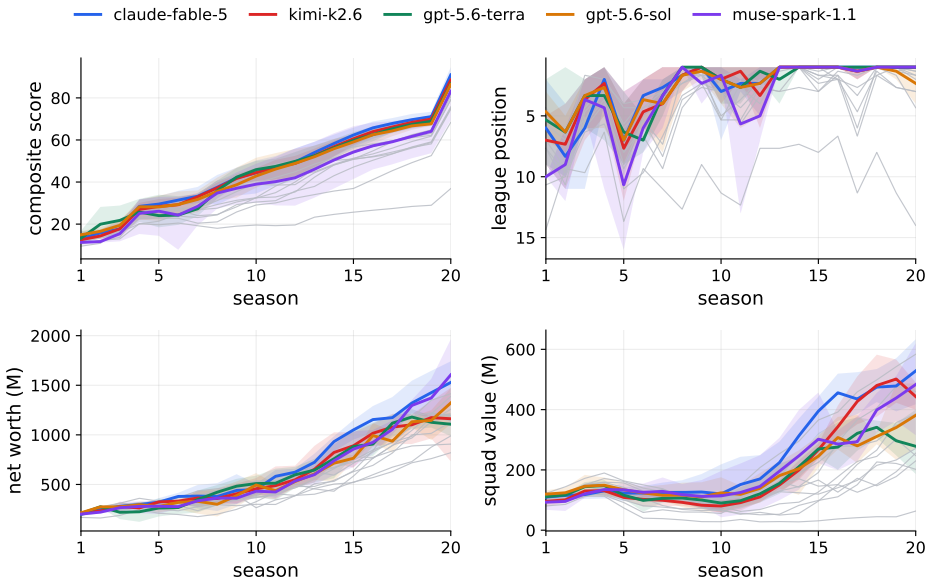

Figure 2 Per-season solo trajectories for all 15 models in four channels. Lines are the mean over three seeds, shading spans the best and worst seed, and the five models with the highest mean are colored and named in the legend; the grey lines are the remaining ten models.

Table 3 The six first-play human runs (§4.2.2). Deaths are capped revivals spent before the settling firing; t is the game-year at which a fired run settled.

player	S_{final}	settle	deaths	stops	time
H1	74.64	completed	0	394	10 h
H2	59.95	completed	0	398	4 h
H3	5.39	fired, t=10.8	4	202	4 h
H4	2.47	fired, t=10.7	4	198	5 h
H5	0.41	fired, t=7.8	4	165	1.5 h
H6	−3.78	fired, t=12.3	4	237	2 h

the second chances never pay. The 15 opponents are the same scripted manager at five *easy*, five *medium*, and five *hard* competence settings, differing in skill and never in information, so every model faces the same world (Appendix F). We ran three seeds for the 15 models (Appendix C) and for three scripted anchors, including a disciplined heuristic, an idle policy, and a random policy (Appendix E).

The oracle, a *scripted* policy whose sole privilege is reading the true hidden state through the same interface and budgets, averages 95.54, and `claude-fable-5` reaches 90.94 without seeing any of it. Both sit at roughly two-thirds of a loose upper bound of ≈145 computed from channel values no run can reach (Appendix I), so headroom remains, and the scale is not saturated. Per-model spread across seeds runs from 0.15 for `kimi-k2.6` to 22.73 for `claude-haiku-4.5` (Figure 7), and four models swing by more than 20 points, so a strong mean can hide a collapse. Neither scale, price, nor vendor orders the ranking, as `gpt-5.6-terra` (86.66) finishes above `gpt-5.6-sol` (86.40), open-weight `kimi-k2.6` (88.49) clears both, and token spend is uncorrelated under any accounting (§4.2.1). Early standing does not predict the outcome either. On seed 1 the best-to-worst gap widens from 17.2 points at year 5 to 37.3 at year 20, the rank correlation with the final order climbs from 0.19 at year 5 to 0.78 at year 15, and `deepseek-v4-pro` leads on score at both year 5 and year 10 yet finishes 12th, while the eventual winner `claude-fable-5` is 5th at year 5. A shorter horizon would have ranked a different set of models. Figure 2 plots every model season by season, and Table 8 gives the numeric snapshots.

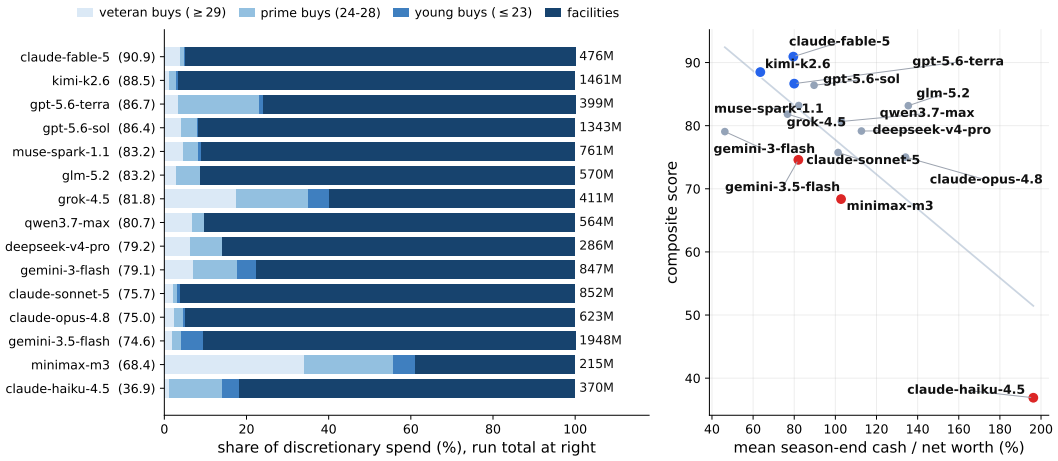

Figure 3 Credit assignment. Left: discretionary spend bucketed by payoff horizon, with run totals at right. Right: mean season-end idle-cash ratio vs. final score, $r_s = -0.50$, negative on every seed.

4.2.1 Capability decomposition: what the ranking is made of

A composite score orders the field but hides why a model lands where it does, and evaluation should separate outcome from the behavior that produced it. We decompose solo performance into six behavioral capabilities (Figure 5), each read from a reproducible behavioral metric whose exact construction is given in Appendix J. Associations with the final score are Spearman rank correlations over the 15 models, written r_s.

Endgame awareness. A 20-year run ends, and slow-payoff actions stop being investments once there is no time left to collect. We measure the late-minus-mid shift in facility and academy actions, the rate over years 17–20 minus the rate over years 2–16, and it is the capability that tracks the score most closely ($r_s = -0.58$, same sign on every seed: -0.31 / -0.34 / -0.28). The winner cuts hardest, from 2.4 to 0.8 such actions per season, `gpt-5.6-terra` from 1.4 to 0.4, while `gemini-3.5-flash` instead *raises* them from 2.4 to 3.3 and is still building facilities in year 19. Rational endgame behavior under a finite horizon is a question only a long benchmark can pose; in short episodes it does not exist.

Credit assignment. The mean ratio of season-end cash to net worth tracks the score across all three seeds ($r_s = -0.50$, and -0.74 / -0.60 / -0.34 per seed; Figure 3). Net worth books excess cash at a discount, so a ratio above 100% means the idle pile exceeds the club's entire discounted worth, and the bottom of the board sits there: `claude-haiku-4.5` averages 196%, `glm-5.2` 135%, `claude-opus-4.8` 134%, against 80% for the winner and a field median of 90%. Total facility investment is uncorrelated, so the signal is *whether money sits idle*, not whether it is spent. Holding cash is the locally safe action whose opportunity cost surfaces seasons later, exactly the credit horizon the environment stresses (demand 2).

Proactive control. For a fully foreseeable future event, a contract expiring, the field splits by *when* it acts (Figure 4). The winner opens renewal negotiations a median of 18 months before expiry, with 4% of episodes opened inside the final six months, while `claude-opus-4.8` opens at a median of 10 months with 20% last-minute and `claude-haiku-4.5` at 11 months with 21% ($r_s = +0.45$, positive on every seed: $+0.68$ / $+0.18$ / $+0.47$). The early-renewal policy is not unique to the winner. `qwen3.7-max` holds the longest renewal leads after the winner, and one first-play human derived the same policy, signing every contract for the maximum length (§4.2.2).

Price discovery. The oracle bids the seller's true accept threshold and closes every buy on the first offer (Appendix H), fixing the reference at 1.0 offers per completed signing. No model comes close: the field median is 30, the best model (`claude-fable-5`) needs 9, and `gemini-3.5-flash` needs 73, with single seeds as high as 133. After twenty years and hundreds of rejections, the models have not learned where the acceptance boundary lies; the low completion rate of the solo market is a pricing failure, not an empty market. The behavior is a judgment about the market, not a fixed habit: the winner bid 0.5 times per season in the solo market but 4.6 in the Arena (§4.3).

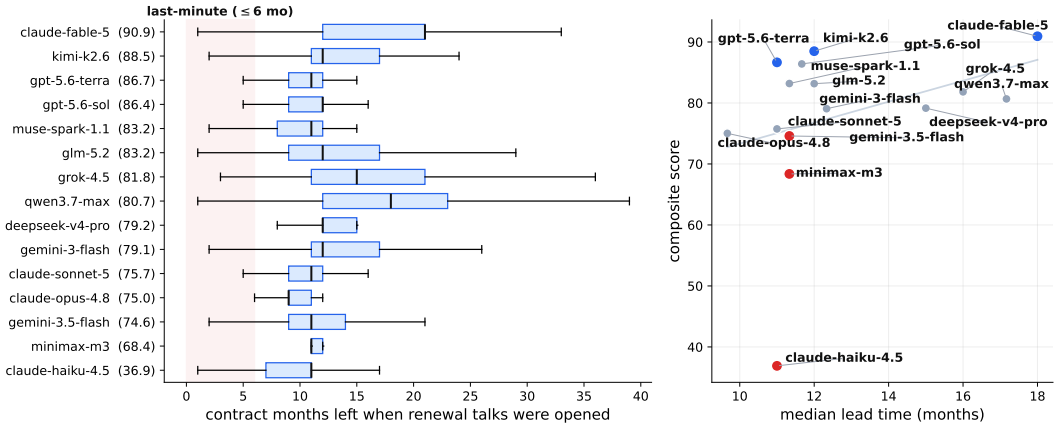

Figure 4 Proactive control. Left: distribution of contract months remaining when each renewal negotiation episode opens (rows sorted by final score; shaded band = last-minute, ≤ 6 months). Right: per-model median lead vs. final score, $r_s = +0.45$, positive on every seed.

Memory curation. The notebook is the only cross-stop state the agent controls (§3.2). We reconstruct each model's notebook at every season end and take the TF–IDF cosine similarity between consecutive snapshots, so 1 means the text never changes, and 0 means a full rewrite. This similarity separates two opposite failures (Figure 13, Appendix J): `gpt-5.6-sol` sits at 0.91, an append-only archive in which current state drowns in history, while `claude-sonnet-5` (0.20) and `qwen3.7-max` (0.23) rewrite so completely that no plan survives to be executed. The winner sits at 0.39 against a field median of 0.31, holding a stable strategy skeleton while rewriting the state each season. Similarity alone does not certify good curation, though. `claude-haiku-4.5` averages 0.31, as close to the middle band as the winner, and still finishes last.

Compute allocation and efficiency. Token spend spans a factor of seven, from 28M for `gpt-5.6-terra` to 194M for `deepseek-v4-pro`, yet it does not order the board on any seed ($r_s = -0.19$, and $+0.07$ / -0.23 / -0.15 per seed; Figure 14), and the winner is among the three cheapest models. The matrix's compute axis (Figure 5) therefore reports efficiency, score per million tokens; `gpt-5.6-terra` and the winner top it, and the heaviest spenders fill its bottom.

Capability profiles. The winner is even across the axes, topping the board without leading most single ones (Figure 5; per-model profiles in Table 7, Appendix J). The mid-table agents pair one genuine strength with one clear gap. `gemini-3-flash` keeps the best cash discipline in the field (46% idle-cash ratio against a median of 90%) yet ranks tenth, the model least willing to reduce long-horizon spending late in the run. The bottom fails on several axes at once, `claude-haiku-4.5` pairing the highest idle cash with the shortest renewal leads and a rising endgame investment rate.

4.2.2 Human study

Six people played the same 20-year track in web-play mode, none of whom had played FM-Bench before. Four of the six died out (Table 3), all four below the disciplined heuristic anchor. The other two completed the horizon with zero deaths, at 74.64 and 59.95, clear of every scripted anchor. Against the three-seed model means of Table 2, the weaker survivor finishes above only `claude-haiku-4.5` (36.90) and the stronger lands level with `gemini-3.5-flash` (74.59), 16 points short of `claude-fable-5`, placing frontier models between untrained humans and the oracle. Asked what they found hard (Appendix D), the players named the demands, from unrecoverable bids against an unseen accept threshold to a wage bill one bad season from dismissal. Their failures mirror the decomposition, letting cash idle while the bill arrived seasons later (§4.2.1), while their strengths are the ones the models lack, deriving standing rules, revising failing policies mid-run, and building an external helper for implicit state. The two profiles are close to complementary, and since the score is identical across modes, whether a human–agent pairing beats either alone, and which capability the gain comes from, is measurable. We leave that measurement to future work.

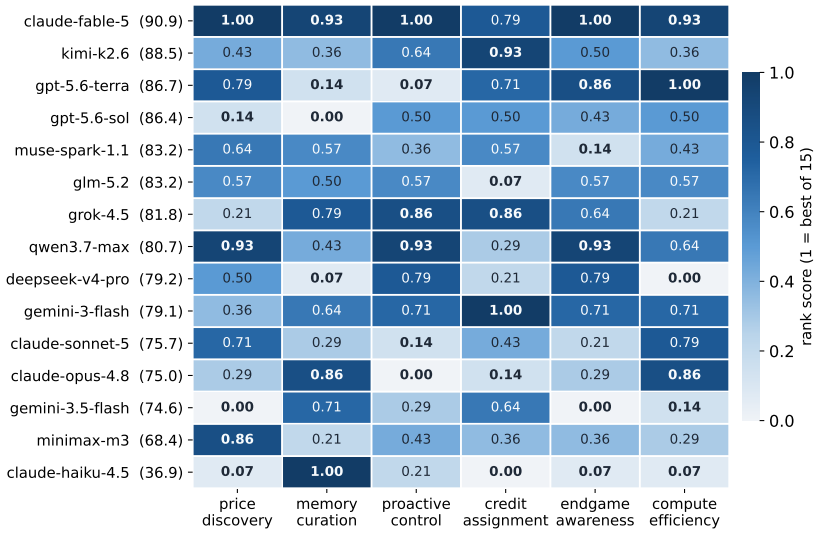

Figure 5 Capability matrix over the 15 solo models, six axes. Each column rank-normalizes one behavioral metric averaged over the three seeds (1 = best of 15; construction in Appendix J), and rows are sorted by mean final score. The winner is a generalist, with no axis below 0.79 and mean 0.94, rather than the leader of any single one; the mid-table pairs a genuine strength with a decisive gap; and the bottom of the board is weak on most axes at once.

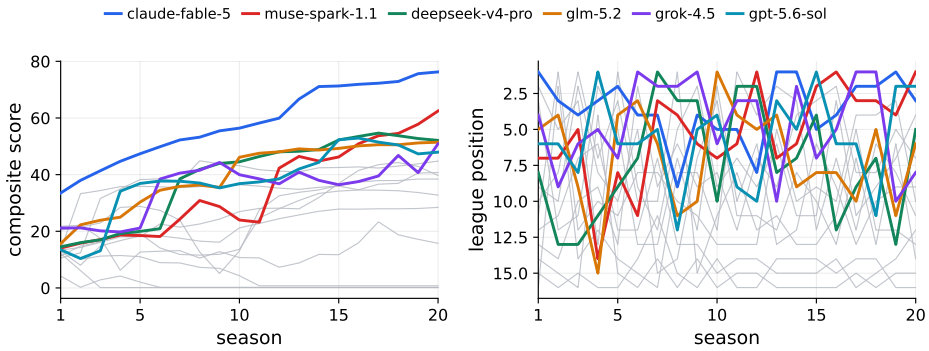

Figure 6 Arena trajectories, all 16 seats. The six highest-scoring seats are colored and named in the legend; the grey lines are the remaining ten seats. Left: composite S_{final} at each season end. Right: league-table position.

4.3 The Arena: a shared-world league

The Arena places 15 agents and the heuristic anchor in one shared 20-year world under the same stack as the solo track (§4.2). Table 4 reports all 16 seats. The per-season trajectories (Figure 6, asset channels in Figure 15, Appendix J) again put the scripted anchor first out of the world, in year 3, but the shared world is harsher than the solo track, and the two weakest LLM models follow it out by exhausting their revivals. They also show league standings and composite scores diverging by design, since `gemini-3-flash` led the table at year 5 yet finished 13th after two late firings, while the eventual winner spent the mid-game in mid-table accumulating convertible assets. As on the solo track, scale and price do not order the board, with several low-cost open-weight models above flagship-priced ones.

Dynasties on the solo track, rotation in the Arena. On the solo track, the four highest-scoring runs each hold the league title continuously through year 20 (Figure 2). In the Arena, the title rotates. Ten different models won it at least once, the reigning champion kept it in only 2 of 19 season transitions, and the eventual composite winner `claude-fable-5` took just 4 titles. The contrast isolates the counter-adaptive market (demand 3). Against fixed opponents, an early lead compounds into a rich-get-richer dynasty, while adaptive rivals bid talent away from whoever leads, and the composite winner is set by 20-year asset accumulation rather than

Table 4 The Arena results.

#	seat	S_{final}	deaths	settle	tokens
1	claude-fable-5	76.26	0	completed	∼34M
2	muse-spark-1.1	62.47	0	completed	∼80M
3	deepseek-v4-pro	52.09	0	completed	∼111M
4	glm-5.2	51.44	0	completed	∼86M
5	grok-4.5	50.89	0	completed	∼138M
6	gpt-5.6-sol	47.94	0	completed	∼78M
7	minimax-m3	44.78	0	completed	∼65M
8	claude-sonnet-5	40.75	0	completed	∼42M
9	kimi-k2.6	39.72	0	completed	∼119M
10	gpt-5.6-terra	38.44	0	completed	∼19M
11	qwen3.7-max	32.77	2	completed	∼54M
12	claude-opus-4.8	28.44	1	completed	∼43M
13	gemini-3-flash	15.76	2	completed	∼63M
14	claude-haiku-4.5	0.76	4	fired, t=10.7	∼30M
15	gemini-3.5-flash	0.21	4	fired, t=5.8	∼40M
16	heuristic (anchor)	0.13	4	fired, t=2.6	—

title count. The winner feels the same pressure in its own trading, 0.5 transfer offers per season against the fixed solo opponents but 4.6 in the contested Arena market.

Behavioral analysis. The ordering summarizes outcomes; to see the behavior behind it, we read the full action logs and self-authored notebooks of three models, the winner (`claude-fable-5`), the early leader that collapsed (`gemini-3-flash`), and the highest-cash low-placing model (`claude-opus-4.8`).

(i) Persistent planning versus season-local reaction. The winner kept one cumulative plan for all 20 years, a running honors ledger whose rewrites revise a single strategy rather than starting over, and intervened sparingly and precisely, from a Y13 wage-breach remedy to a Y19 title push, on 91 transfer offers in total. `gemini-3-flash` issued 452 offers and 2,642 total actions (the winner's 1,315), and its notebook resets to season-local firefighting at each crisis. Table position bought by heavy early activity did not survive the horizon.

(ii) A knowing–doing gap on delayed reward. `claude-opus-4.8` recorded the correct policy in writing, noting at Y10 that idle cash should be deployed on quality and at Y19 that its reserves were discounted and should become young players, yet finished holding some 2,100M in idle reserves. Holding cash is locally safe but earns almost nothing under the composite. The model wrote the right long-range plan and did not execute it.

(iii) Deliberation volume does not predict skill. The compute null of §4.2.1 holds under every accounting we tried, cache-inclusive, output-only, and dollars (all $p > 0.38$, $n = 15$), and since each model chooses its own number of turns per stop, the spread is a behavioral property rather than a harness setting.

5 Conclusion

FM-Bench measures whether language model agents can manage rather than merely execute, handing each one a football club for twenty in-game years under the four demands that define the job, with a deterministic engine grading every run and no LLM judge. Every frontier model completes the horizon that the blind scripts do not survive, and the strongest, `claude-fable-5`, tops both tracks while reaching about 95% of a privileged oracle, so the benchmark separates models without being saturated. The board is one no shorter or cheaper measurement would predict. Scale, price, and vendor do not order it, year-5 standings barely correlate with the final order, and the title rotates among ten models once the rivals are adaptive. The scalar decomposes into interpretable behaviors, and the claim that one model manages better unpacks into which demands it meets, which it fails to meet, and what it was doing when it failed.

References

Alibaba Qwen Team. Qwen3.7-max, 2026. https://qwen.ai/blog?id=qwen3.7.

Andon Labs. Vending-bench arena. https://andonlabs.com/evals/vending-bench-arena, 2026. Leaderboard page and blog posts only; no technical report or source release.

Anthropic. System card: Claude haiku 4.5, 2025. URL https://www.anthropic.com/news/claude-haiku-4-5.

Anthropic. Claude fable 5 and claude mythos 5, 2026a. URL https://www.anthropic.com/news/claude-fable-5-mythos-5. System card and announcement.

Anthropic. Claude opus 4.8 system card, 2026b. https://www.anthropic.com/news/claude-opus-4-8.

Anthropic. Introducing claude sonnet 5, 2026c. URL https://www.anthropic.com/news/claude-sonnet-5.

Axel Backlund and Lukas Petersson. Vending-bench: A benchmark for long-term coherence of autonomous agents. *arXiv preprint arXiv:2502.15840*, 2025.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *Advances in Neural Information Processing Systems*, 38, 2026.

Haozhe Chen, Karthik Narasimhan, and Zhuang Liu. Ceo-bench: Can agents play the long game? *arXiv preprint arXiv:2606.18543*, 2026.

Yanxu Chen, Zijun Yao, Yantao Liu, Amy Xin, Jin Ye, Jianing Yu, Lei Hou, and Juanzi Li. Stockbench: Can llm agents trade stocks profitably in real-world markets? *arXiv preprint arXiv:2510.02209*, 2025.

Zihao Cheng, Hongru Wang, Zeming Liu, Xinyi Wang, Xiangrong Zhu, Yuhang Guo, Wei Lin, Jeff Z Pan, and Yunhong Wang. Terminal-world: Scaling terminal-agent environments via agent skills. *arXiv preprint arXiv:2605.20876*, 2026.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Workshop on computer games*, pages 41–75. Springer, 2018.

Chongyang Gao, Diji Yang, Shuyan Zhou, Xichen Yan, Luchuan Song, Shuo Li, and Kezhen Chen. Classroom final exam: An instructor-tested reasoning benchmark. *arXiv preprint arXiv:2602.19517*, 2026.

Google DeepMind. Gemini 3 flash model card, 2025. URL https://deepmind.google/models/gemini/flash/.

Google DeepMind. Gemini 3.5 flash model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. *arXiv preprint arXiv:2109.06780*, 2021.

Yi Han, Yan Wang, Lingfei Qian, Haohang Li, Yupeng Cao, Yueru He, Xueqing Peng, Nanhan Shen, Yitao Xu, Yankai Chen, et al. Can llm agents be cfos? benchmarking long-horizon resource allocation in an uncertain enterprise environment. *arXiv preprint arXiv:2603.23638*, 2026.

Muyu He, Vincent Tu, Adit Jain, Anand Kumar, Sachin Patro, Soumyadeep Bakshi, and Nazneen Rajani. Yc-bench: Benchmarking ai agents for long-term planning and consistent execution. In *ICML 2026 Workshop on Combining Theory and Benchmarks: Towards A Virtuous Cycle to Understand and Guarantee Foundation Model Performance*, 2026.

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*, volume 2024, pages 54107–54157, 2024.

Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment. *Advances in Neural Information Processing Systems*, 33:7671–7684, 2020.

Zongxia Li, Zhongzhi Li, Yucheng Shi, Ruhan Wang, Junyao Yang, Zhichao Liu, Xiyang Wu, Anhao Li, Yue Yu, Ninghao Liu, et al. Long-horizon-terminal-bench: Testing the limits of agents on long-horizon terminal tasks with dense reward-based grading. *arXiv preprint arXiv:2607.08964*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601, 2024.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pages 52989–53046, 2024.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, et al. Ultrahorizon: Benchmarking agent capabilities in ultra long-horizon scenarios. *arXiv preprint arXiv:2509.21766*, 2025.

Parsa Mazaheri and Kasra Mazaheri. Agentatlas: Beyond outcome leaderboards for llm agents. *arXiv preprint arXiv:2605.20530*, 2026.

Meta Superintelligence Labs. Introducing muse spark 1.1, 2026. URL https://ai.meta.com/blog/introducing-muse-spark-meta-model-api/.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pages 9025–9049, 2024.

MiniMax Team. Minimax m3, 2026. https://www.minimax.io/blog/minimax-m3.

Moonshot AI. Kimi k2.6, 2026. URL https://www.kimi.com/ai-models/kimi-k2-6. Model card.

Robert Müller and Clemens Müller. Cattle trade: A multi-agent benchmark for llm bluffing, bidding, and bargaining. *arXiv preprint arXiv:2605.14537*, 2026.

OpenAI. Previewing gpt-5.6 sol, terra and luna, 2026. URL https://openai.com/index/previewing-gpt-5-6-sol/. System card at deploymentsafety.openai.com/gpt-5-6-preview.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wołczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, et al. Balrog: Benchmarking agentic llm and vlm reasoning on games. In *International Conference on Learning Representations*, volume 2025, pages 96666–96702, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pages 9695–9717, 2024.

Qiming Shi, Yulong Tao, Linbo Jin, Zhaolu Kang, Yibo Dou, Jiawen Zhu, Tianjun Pan, Shaokang Fu, Chengyu Wang, Siyue Li, et al. Merchantbench: Benchmarking llm agents for long-term coherence in e-commerce operations. *arXiv preprint arXiv:2607.28956*, 2026.

Issa Sugiura, Daichi Hattori, Kazuo Araragi, Keita Ogawa, Shota Onose, Taro Makino, Teppei Usuki, and Takashi Ishida. Coffeebench: Benchmarking long-horizon llm agents in heterogeneous multi-agent economies. *arXiv preprint arXiv:2606.16613*, 2026.

Wei Wang, Dan Zhang, Tao Feng, Boyan Wang, and Jie Tang. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *arXiv preprint arXiv:2408.15971*, 2024.

xAI. Grok 4.5, 2026. https://x.ai/news/grok-4-5.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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.

Zhipu AI. Glm-5.2, 2026. https://z.ai/blog/glm-5.2).

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pages 15585–15606, 2024.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Daisy Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, 2025.

Appendix

A Limitations

This section records the boundaries a reader should keep in mind when interpreting our results: where sampling is thin, what the environment deliberately simplifies, and how far the construct claim extends beyond this game. Every published run remains bit-replayable regardless; the items below bound interpretation, not reproducibility.

- **Three seeds bound the solo board, and the Arena is one world.** The solo results are means over three seeds, enough to show that adjacent models do not separate and that spread differs sharply by model (§4.2), but not enough for inferential claims about any single pair. The Arena (§4.3) is one shared seed-7 world with no error bars, so its orderings within a few points should be read as ties, and aggregating several Arena worlds would call for a rank-based aggregate over per-world finishes, which we leave to future work.
- **Scripted-opponent realism.** The 15 opponents are disciplined hand-written scripts by design (a fixed track is what makes absolute scores comparable), so the environment does not model adaptive rival managers; the Arena is the adversarial complement, not a realism claim.
- **Construct validity.** FM-Bench instantiates long-horizon management in one domain. The four demands are domain-general design targets, but transfer of FM-Bench rankings to other long-horizon settings is untested.
- **Behavioral metrics are correlational.** The capability decomposition (§4.2.1) reads reproducible metrics off the runs; a metric can be confounded by other traits (the warning-exposure metric demonstrably is), coefficients over $n = 15$ are indicative, and the causal probes that would upgrade them (notebook ablation, plan injection) are future work.

B Reproducibility

Deterministic engine (counter-based RNG keyed by (domain, entity, day); fixed daily phase order; no wall clock); a 284-test suite including golden-hash behavior guards, full-tool-path replay bit-identity, crash-resume equivalence, AST truth-isolation, observation-audit, and calibration regressions; every result file carries engine commit, params hash, score version, and the full knob set; scripted baselines and the full calibration suite run for free on a laptop (20 simulated years in 76 s single-core). Open Track submissions are HMAC-signed and re-verified by server-side replay. Scored worlds are salted (effective seed = HMAC-SHA256 of a secret salt and a public nonce), keeping worlds unguessable despite the open engine; effective seeds are disclosed after each round closes, so every run remains fully replay-auditable. The reported runs exercised this machinery: several 20-year runs survived provider outages or credit exhaustion mid-run and resumed losslessly from the fsync'd action log (no completed turn was re-purchased), and every reported result is bit-replayable from its log, with full per-turn transcripts retained in the repository.

Table 5 The roster: 15 frontier LLM seats plus one scripted anchor, July 2026.

#	seat	provider	role
1	`claude-fable-5` (Anthropic, 2026a)	Anthropic	new flagship
2	`claude-opus-4.8` (Anthropic, 2026b)	Anthropic	previous flagship
3	`claude-haiku-4.5` (Anthropic, 2025)	Anthropic	small tier
4	`gpt-5.6-sol` (OpenAI, 2026)	OpenAI	new flagship
5	`gpt-5.6-terra` (OpenAI, 2026)	OpenAI	balanced tier
6	`claude-sonnet-5` (Anthropic, 2026c)	Anthropic	workhorse mid (4-point family curve)
7	`gemini-3.5-flash` (Google DeepMind, 2026)	Google	latest GA flagship
8	`gemini-3-flash` (Google DeepMind, 2025)	Google	previous-gen fast
9	`deepseek-v4-pro` (Xu et al., 2026)	Together	open flagship
10	`grok-4.5` (xAI, 2026)	xAI	frontier
11	`kimi-k2.6` (Moonshot AI, 2026)	Together	open agentic specialist
12	`qwen3.7-max` (Alibaba Qwen Team, 2026)	Together	open flagship
13	`glm-5.2` (Zhipu AI, 2026)	Together	open top tier
14	`muse-spark-1.1` (Meta Superintelligence Labs, 2026)	Meta Model API	Meta agentic model
15	`minimax-m3` (MiniMax Team, 2026)	Together	open flagship
16	heuristic	— (scripted)	disciplined-script anchor

C The Roster

The roster (Table 5) fixes four selection principles even as individual model IDs rotate: within-family tier curves, closed against open frontier under identical rules, an agentic and tool-use specialist, and a disciplined scripted anchor in the 16th seat. All 15 serving IDs were validated by tool-calling pings before launch. Same-role substitution is allowed in later rounds, but the principles are fixed.

D What the Human Players Reported

After their runs, we asked the six players (§4.2) what they found hard and what they would change. They were not shown the four demands, the capability decomposition, or any model result. We group their reports below and note, where one exists, the model-side measurement that matches.

Hidden information, reported as a low tolerance for error. Players could not see what a seller would accept or what a player would re-sign for, so a misjudged offer was unrecoverable rather than correctable. Two of them settled on age heuristics, releasing or not renewing past 30, because the decline curve is unobservable. One reported angering sellers with low offers and paying more afterward, which is the repeat-pair markup of the counter-adaptive market (§3.1) felt from the inside. Lineup choice raised the same problem: no player was sure whether to field the highest ability, the best current form, or the freshest legs, and one switched from the first to the last over the run.

Cumulative consequences, reported as delayed bills. Clicking through the early stops felt free, because the engine fields a short squad rather than refusing to continue, and one player took that as a signal that buying was optional; the losses arrived seasons later. This is the idle-cash pattern of §4.2.1, the strongest negative correlate among models, reached by a different route. A second delayed bill came from squad depth. One player released weak low-wage players to shorten a long list and found later that those were exactly the squad depth needed once the first eleven tired or got injured, and revised the policy for the rest of the run. Two further timing effects went unnoticed until late: the transfer market thins out as seasons pass, so buying early is worth more than it looks, and academy players are cheap but start low and mature slowly, which left one player unsure whether youth investment ever repaid. The models leave the same question open: the academy-harvest metric that looked informative on one seed does not survive three (Appendix J).

Credit assignment, reported as helplessness. The most frequent complaint was not knowing which decision caused a losing season. One player fielded the strongest available eleven and lost repeatedly, then won with a

weaker one after injuries forced changes, and could not tell whether formation, opponent matchup, or noise explained it. Once board confidence began to fall, players described having no idea what to change, and four of the six were fired.

Multi-objective pressure, reported as multitasking. Players described the pressure as running valuation and squad maintenance at once, pricing targets in the market while watching for tired players to substitute, with a wage bill above 85% of revenue leaving them one bad season from dismissal.

What the players did that the models do not. Three behaviors stand out against the model-side profiles. One player derived a rule that always holds from two mechanisms: sign for the maximum five years because ability grows and money inflates, and apply it for the rest of the run; renewal lead time is the strongest positive correlate among models, but no model states the invariant. One revised a squad-depth policy in the middle of a run after seeing it fail, whereas in models mid-run tactic switching correlates negatively with score, so course changes there are usually thrash rather than repair. One wrote an external helper with an LLM to surface state the interface leaves implicit, an act of tool-building no model performed. The players also asked for a larger strategy space, in particular opponent-specific formations, after noticing that a weaker eleven in a better shape sometimes wins.

Reading these reports. They are six self-reports on one seed and are not evidence about capability rankings. We use them for two narrower purposes: as construct-validity evidence, since players who were never told the design named all four demands as the hard parts, and as design input, since the operational load they describe, tracking fatigue, contract clocks, and market depth by eye, is exactly what an agent absorbs at no cost.

E The Scripted Anchors

Three blind scripts run the same track as the models (§4.2). None of them reads hidden state; all three go through the same 26-tool interface and per-stop budgets as an agent, and all three are deterministic, so they cost nothing to re-run and fix the scale at the bottom of the board the way the oracle fixes it at the top (Appendix H).

Heuristic, a disciplined manager. The heuristic is a hand-written policy that plays the game the way a careful club would. It keeps the wage bill under 60% of revenue, renews the young core and sells players past 32, buys young value targets when a window is open and, when wage headroom allows, promotes from the academy at intake, rotates the lineup by fatigue, switches playing style against the opponent's table position, and invests in facilities at season settlement when cash is ample. It answers incoming offers on a keep-or-sell rule, accepting good prices for players below the squad's median ability and countering at a markup otherwise. It is the baseline that matters: everything a manager should do mechanically, done without judgment about which season it is.

Idle, the floor. The idle policy submits nothing at the draft and takes no action at any stop thereafter. The engine keeps the club running with default handling, so idle measures what the world does to a club whose manager never intervenes.

Random, below the floor. The random policy answers pending offers by coin flip, then takes zero to three arbitrary actions per stop, drawn from bidding a random amount on a random target, listing or unlisting a random player, and setting a random formation and style. Its draws are seeded from the world seed and the stop ID, so the run is reproducible. Random scores below idle, which is itself informative: in this environment uninformed activity destroys more value than inactivity.

F Opponent Tiers

The 15 scripted opponents of the solo track (§4.2) are one manager implementation under three settings of five competence levers (Table 6). Medium is that implementation with every lever neutral, which is exactly

Table 6 The five competence levers. Medium is the neutral setting.

lever	easy	medium	hard
scouting noise, multiplier on the belief sigma	1.3	1.0	0.35
lineup rotates for fatigue	no	yes	yes
market days acted on	every 2nd	every	every
gap required before buying an upgrade	1.5	1.0	0.25
wage budget, multiplier on the board target	0.92	1.0	1.10

the untiered scripted manager, and easy and hard turn the same five levers down or up. Every lever is a deterministic function of state, so tiers add no randomness, and none of them touches the observation path, so an easy board is worse at judging players rather than being differently informed.

Tiers are assigned by a frozen cyclic hard, medium, easy pattern over the club slots ranked by initial cash, so each wealth band contains all three tiers and money is decorrelated from competence. The pattern is a pure function of the parameters, with no random draw, so every tested model meets the same 15 opponents in the same slots.

An earlier calibration of the levers came out net easier than no tiers at all, because noisy easy sellers priced players badly enough to be free bargains and their unrotated first elevens donated points across the league, while the hard setting was too subtle to compensate. The current values temper how exploitable easy is and give hard sharper beliefs, proactive buying, and a more ambitious wage budget.

G Arena Data Quality: the Notebook Audit

An LLM auditor read every model's self-authored notebook writes across the first seven game-years of the Arena run (§4.3). All 15 models stay self-consistent and factually grounded: 10 clean, 5 with minor style or hygiene issues, none broken. No model hallucinated its squad, invented finances, or lost track of its league position, and weak-but-coherent play, meaning verbose crisis vocabulary, boilerplate, doom loops, and an aging-squad decline, is distinguishable from broken play. Model play and the interface are frozen and identical to the 1v15 solo track: the same engine, the same 26-tool interface, and the same rules, so the two tracks stay comparable. The audit also surfaced two interface-friction phenomena, a standing list order that reads as consent to auto-sell and the money unit-scale duality, which are logged as candidates for a future revision to be applied to both tracks together or not at all.

H The Oracle Policy

This appendix details the privileged oracle summarized in §4.1. Its privilege is information, not power. It reads true hidden state straight from the engine, the one thing no legitimate agent can do, but every action goes through the same tool interface, the same per-stop budgets, and the same legality checks as any agent. Because the engine's random streams are pure functions of the seed, entity, and day, even the draw that resolves a same-tick negotiation is readable truth, so each exploit below is a deterministic read of world state rather than a search.

- **Transfer arbitrage.** The oracle scans every player in the league and computes each seller's exact accept threshold, the asking price adjusted by the seller's style and distress, then bids exactly that. Every purchase closes on the first offer at the lowest price the seller would have taken. Buys are ranked by true worth against effective cost, including the drag that unamortized fee liabilities put on the score.
- **Exact-threshold contracts.** Renewals and free-agent signings offer the player's exact reservation wage, so no negotiation round is wasted, no cooldown is triggered, and the wage bill is the lowest in the league. Since an accepted offer resets the wage, overpaid contracts are cut, and because contract offers sit outside the per-stop negotiation budget while every accepted renewal adds morale, the best players are re-renewed at their known reservation wage whenever that costs nothing, keeping the squad's morale multiplier near its cap.

- **Academy sniping.** Only players with true high potential are promoted, at the last stop before their academy terms lapse, since academy development outpaces bench development. Academy and training facilities are funded early so that intake quality and growth compound.
- **Peak-timed sales.** Every incoming bid is countered at the buyer's exact ceiling, the lower of its transfer budget and 1.15 times its believed value, quantized to the price ladder. Surplus and post-peak players therefore sell for the most any buyer in the world would pay, before the deterministic decline curve bites. Resales within a year of purchase are refused so that the counter-adaptive market never taxes the flywheel.
- **Zero-waste scheduling.** The starting eleven is re-solved at every stop from true ability, form, morale, and projected fatigue across all three formations, and the playing style counter-picks the known styles of the upcoming opponents, net of the style-switch penalty.

The policy is also score-aware. Trading profit and squad growth land in the net-worth channel, sales never hollow the last-three-season window, excess cash discounted by the score is recycled into facilities and squad assets, and contracts are renewed so that nobody expires off the books before the final snapshots. A survival controller reads board confidence directly. Below an emergency threshold, set well above the engine's own vote and firing thresholds, every margin discipline yields to immediate on-pitch strength, so the oracle completes the horizon on every seed. Engine code may never import the oracle module, which a CI isolation test enforces, and oracle scores are reported alongside the ladder but never mixed into it.

I Properties of the Composite Score

This appendix collects four elementary properties of the composite score (§3.3). Write $x^+ = \max(x, 0)$ and $x^- = \min(x, 0)$, so that

$$\begin{aligned} S_{\text{raw}}(H,\, VA, M) &= w_h \ln\!\big(1 + H^+/h_d\big) + w_v \operatorname{sign}(VA) \ln\!\big(1 + |VA|/v_d\big) + w_m \ln\!\big(1 + M^+/m_d\big), \\ S_{\text{final}} &= \rho \cdot S_{\text{raw}}^+ + S_{\text{raw}}^-, \qquad \rho = \begin{cases} 1 & \text{run completed,} \\ \rho_0\,(t/T)^{\rho_1} & \text{settled early at } t \in [0, T), \end{cases} \end{aligned}$$

with weights $(w_h, w_v, w_m) = (18, 10, 6)$, divisors $(h_d, v_d, m_d) = (60, 40, 80)$, and $\rho_0 = 0.8$, $\rho_1 = 1.5$.

Proposition A.1 (no fire-to-cap arbitrage). *For fixed S_{raw}, the map $t \mapsto S_{final}$ is nondecreasing on $[0, T]$, and $S_{final} \le S_{raw}$ with equality iff the run completes or $S_{raw} \le 0$.*

Proof. If $S_{\text{raw}} \le 0$ then $S_{\text{raw}}^+ = 0$ and $S_{\text{final}} = S_{\text{raw}}$ for every t: the discount multiplies only the positive part, and the negative part passes through undiscounted, so early settlement leaves a negative score exactly unchanged: "get fired early to cap losses" changes nothing. If $S_{\text{raw}} > 0$ and the run settles early at $t < T$, then $S_{\text{final}} = \rho_0 (t/T)^{\rho_1} S_{\text{raw}}$ with $\partial S_{\text{final}}/\partial t = \rho_0 \rho_1 t^{\rho_1 - 1} T^{-\rho_1} S_{\text{raw}} > 0$, and $\rho_0 (t/T)^{\rho_1} \le \rho_0 < 1$, while completion gives $\rho = 1$. Hence $\partial S_{\text{final}}/\partial t \ge 0$ throughout, and triggering an earlier settlement can only lower the score: it is never optimal. □

Remark. The proposition freezes S_{raw}; the dynamic variant, whether $\rho(t) \cdot S(t)$ along a real trajectory can exceed the expected completed score, is exactly the self-fire exercise point probed empirically in CI. The discount is deliberately discontinuous at the horizon: $\rho \to \rho_0 = 0.8$ as $t \to T^-$ while completion pays $\rho = 1$, so surviving the final stretch carries a 25% premium over settling just short of it; the jump points in the direction an agent cannot exploit.

Proposition A.2 (channel monotonicity). *S_{raw} is strictly increasing in H on $H \ge 0$ (weakly on all of $\mathbb{R}$: the floor makes the honors term constant for $H < 0$), strictly increasing in VA on all of $\mathbb{R}$, and weakly increasing in M (strictly on $M \ge 0$).*

Proof. For $H \ge 0$ the honors term is $w_h \ln(1 + H/h_d)$ with derivative $w_h/(h_d + H) > 0$. For the VA term let $g(x) = \operatorname{sign}(x) \ln(1 + |x|/v_d)$: for $x \ne 0$, $g'(x) = 1/(v_d + |x|) > 0$, and at $x = 0$ both one-sided difference

quotients converge to $1/v_d$, so g is differentiable on $\mathbb{R}$ with $g' > 0$ everywhere, hence strictly increasing. The M term is constant for $M < 0$ and has derivative $w_m/(m_d + M) > 0$ for $M \geq 0$. □

Proposition A.3 (fair-price transfer neutrality on the VA channel). *Net worth is booked as $N = C + \lambda Q + F - L$, with C cash, Q squad value, $\lambda = 0.8$ the squad liquidity discount, F facilities, and L outstanding fee liabilities (§3.3). Buying a player of market value v at price $p \geq v$, under any split of up-front cash and installments, changes net worth by $\Delta N = -p + \lambda v \leq -(1-\lambda)\, v < 0$. Hence purchases at (or above) fair price strictly decrease net worth, and by Proposition A.2 the VA term weakly decreases: squad-shopping cannot pump the VA channel.*

Proof. The paid portion of p leaves C and the unpaid portion enters L, decreasing $C + \cdots - L$ by exactly p; the player enters Q at value v, booked at λv. So $\Delta N = -p + \lambda v$, which at $p = v$ equals $-(1-\lambda)v < 0$ and is smaller still for $p > v$. Only buying below the discount wedge ($p < \lambda v$) raises N, the intended reward for scouting genuine bargains. Wage commitments add future outflows and only strengthen the inequality. □

Two scope notes. (i) Cash above 3× the annual wage bill is booked at 0.3 (anti-hoarding, §3.3); spending *such penalized* cash on players can raise N ($\lambda = 0.8 > 0.3$). This is the intended incentive to deploy hoarded cash, not an arbitrage: a purchase never books more than $\lambda < 1$ of the cash it consumes, so no purchase sequence can push N above what the same cash counted at par would have given. (ii) The M channel is *absolute* squad value, so a purchase does raise the M term. Protection of the *total* score against squad-shopping therefore rests on M's low weight (6 vs 10) and the discount wedge ($1 - \lambda$); it is an empirical claim (a per-channel value-added variant that removed the wedge resurrected this exploit and was rejected), not a theorem.

Proposition A.4 (diminishing returns; no runaway channel). *Each channel term is concave in its input on the gain domain, so the marginal value of the k-th title (or the marginal million) is strictly decreasing, and no single channel can run away.*

Proof. $\frac{d^2}{dx^2} \ln(1 + x/d) = -(d + x)^{-2} < 0$ on $x \geq 0$, which covers the honors and squad terms; the VA term is odd, hence concave on $VA \geq 0$ and (symmetrically) convex on $VA \leq 0$, so gains and losses are compressed at the same rate. Concretely, the k-th league title (+100 honors points) is worth $w_h \ln\big((h_d + 100k)/(h_d + 100(k-1))\big)$, i.e. 17.7, 8.7, 5.9, … points for $k = 1, 2, 3, \ldots$ Because each term grows only logarithmically in its input, overcoming a fixed additive lead on one channel requires a *multiplicative* resource gap on that channel: the score is uncapped (the benchmark cannot saturate) yet no channel can dominate the composite. □

A loose upper bound. The channels are uncapped, so the composite has no mathematical supremum, but an impossibility bound follows from channel values no run can reach. Winning the title in all 20 seasons gives $H = 2000$ and an honors term of $18 \ln(1 + 2000/60) = 63.7$. Draining the entire world's cash into value added, $VA = 10{,}000$M, gives $10 \ln(1 + 10000/40) = 55.3$, and a full squad of maximum-ability players, $M \approx 6{,}000$M, gives $6 \ln(1 + 6000/80) = 25.6$. With $\rho = 1$ the terms sum to ≈145. The world holds neither that much extractable cash nor that many such players, so the bound is loose by construction, and the oracle's 95.54 and the best blind model's 90.94 sit at roughly two-thirds of it.

J Behavioral Analysis: Definitions, Construction, and Additional Figures

This appendix documents the measurement pipeline behind §4.2.1. Everything is reproducible from the self-contained analysis folder published with the repository (scripts, mined data, and figures, with the extraction commands documented).

Replay gate. All behavioral metrics are mined by bit-identical replay of each run's seed and action log for the 15 solo runs and the Arena archive. Two replay passes were run (one for activity/notebook/tactics, one reading engine truth, namely contract months remaining, player age and potential, and club cash, at the instant each action executed). Each pass verifies that the replayed final score reproduces the published S_{final} of the run before its log is trusted; all 30 replays matched with zero drift. Solo action timing is therefore

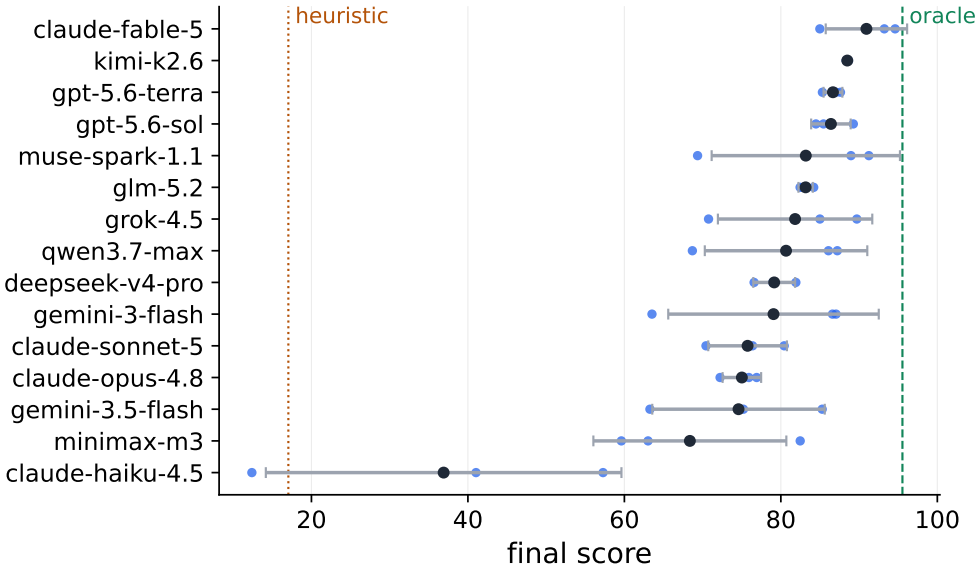

Figure 7 Final score per model, mean with standard deviation over the three seeds, individual seeds shown as points. Dashed and dotted lines mark the oracle and heuristic means. Stability separates models that a single-seed board cannot tell apart, from `kimi-k2.6` at 0.15 to `claude-haiku-4.5` at 22.73.

replay-exact; Arena timing maps each seat's stop index linearly onto its tenure (∼19 stops per club-year) and is approximate.

Metric definitions. One item per capability of §4.2.1; each model contributes the mean of its statistic over the three seeds, and the stated direction is the one the matrix axis rewards.

- *Endgame awareness*: long-horizon actions (`invest` + `promote_youth`) per season; the statistic is the late-minus-mid contrast, the rate over years 17–20 minus the rate over years 2–16 (low is better).
- *Credit assignment*: idle-cash ratio, season-end cash ÷ net worth, averaged over seasons 2–20 (low is better). Net worth books excess cash at a discount (§3.3), so values above 100% indicate an idle pile exceeding the club's entire discounted worth.
- *Proactive control*: renewal lead, the contract months remaining when a renewal negotiation *episode* opens; offers to the same player within 60 game-days collapse into one episode, own-club players only; the statistic is the median over episodes (high is better). *Last-minute share*: episodes opened at ≤ 6 months.
- *Price discovery*: offers per completed signing, successful `make_transfer_offer` submissions ÷ senior-squad arrivals with a positive recorded fee (drafted and promoted players excluded; low is better). The oracle's first-offer policy fixes the reference at 1.0 (Appendix H).
- *Memory curation*: each model's notebook is reconstructed at every season end under exact engine semantics (append joins with a newline, rewrite replaces); consecutive season-end states are compared by TF–IDF (1–2 gram) cosine similarity, and the per-model mean is the consistency statistic. The matrix axis scores the distance $|s - 0.35|$ from the empirically best band (low is better). TF–IDF is chosen over neural embeddings for determinism and zero external dependencies; a semantic-embedding replication is future work.
- *Compute*: tokens by manifest accounting, input + output + cache read + cache write from each run's manifest. The correlation in §4.2.1 uses the total; the matrix axis scores efficiency, score per million tokens (high is better).

Failed metrics, kept in the record.

Table 7 Per-model profiles over the three seeds. Score is the mean and SD the standard deviation of S_{final}; axis extremes come from the matrix of Figure 5.

seat	score	sd	profile
claude-fable-5	90.9	5.2	generalist, no axis below 0.79; best bidder in the field at 9 offers per signing and the sharpest endgame reduction
kimi-k2.6	88.5	0.1	the steadiest model in the campaign, scoring within a third of a point on three different worlds
gpt-5.6-terra	86.7	1.2	minimalist: lowest token spend in the field (28M) after the winner and an early endgame reduction, but the shortest renewal leads among the top models
gpt-5.6-sol	86.4	2.5	archivist memory (similarity 0.91, append-only) and the weakest bidding of the top group (45 offers per signing)
muse-spark-1.1	83.2	12.0	strong on two worlds and 20 points weaker on the third; one of two models that raise long-horizon spending at the end
glm-5.2	83.2	0.9	stable across seeds but loose with cash (135% idle ratio)
grok-4.5	81.8	9.9	compute as substitute: long renewal leads and heavy spend (191M tokens) with wide seed-to-seed swings
qwen3.7-max	80.7	10.4	longest renewal leads after the winner, undone by a 103% idle ratio and churning memory (similarity 0.23)
deepseek-v4-pro	79.2	2.7	near-static notebook (0.81) and the heaviest spend in the field (194M tokens)
gemini-3-flash	79.1	13.5	best cash discipline in the field (46%) but the least willing to cut endgame spending; collapses on the hardest seed
claude-sonnet-5	75.7	5.0	churning memory (0.20, the lowest) and no endgame reduction
claude-opus-4.8	75.0	2.5	shortest renewal leads in the field (10 months) with a 134% idle ratio
gemini-3.5-flash	74.6	11.0	worst price discovery (73 offers per signing) and the largest endgame ramp-up
minimax-m3	68.4	12.3	prefers veteran signings, short leads, and swings 23 points across seeds
claude-haiku-4.5	36.9	22.7	weak on every axis at once: 196% idle cash, 11-month leads, a rising endgame rate, and the widest spread in the campaign

- *Warning exposure*: fraction of decision stops whose digest carried at least one engine warning (wage ratio, negative cash, board confidence, administration); per-seed sign unstable (Figure 11).
- *Youth harvest*: share of `promote_youth` players later fielded in ≥ 5 lineups; per-seed sign unstable (Figure 12).
- *Tactic reversals*: `set_tactics` calls whose formation or style differs from the currently active pair (re-issuing the current tactic does not count); dropped as a capability axis for the same reason, though the raw spread stays wide, zero reversals for the winner in all three worlds against 31 on average for `claude-haiku-4.5`.

Token-file erratum. A summary token file that circulated with early drafts mixed in token totals from earlier runs for ten of the fifteen models. All token numbers in this paper are recomputed from the run manifests; the superseded file is retained in the repository for audit.

Capability-matrix construction (Figure 5). Each column rank-normalizes one metric across the 15 models to $[0, 1]$ (1 = best rank; ties broken by value order), in the direction stated in the definitions above. Rank scores are a visualization-grade aggregation over $n = 15$ models; the evidentiary weight rests on the underlying metrics.

Season snapshots. Tables 8 and 9 reproduce the numeric season snapshots of the public leaderboard page (years 5, 10, 15, 20). Cell format, matching the page: league position · cumulative honors points · net worth (M) · squad value (M) · completed-now composite score. They are the discrete complement of the trajectory figures (Figures 2, 6, and 15).

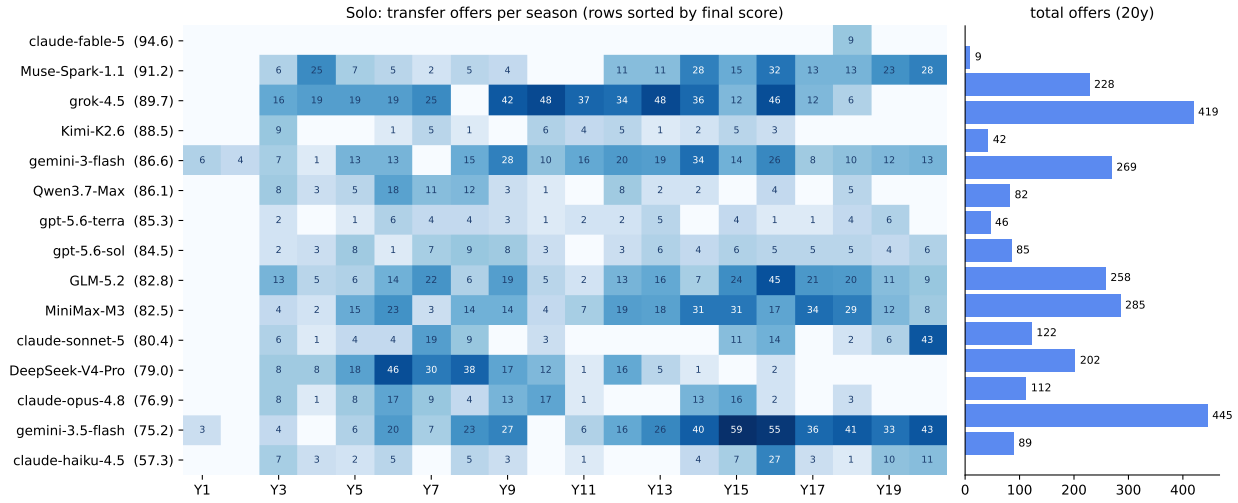

Figure 8 Transfer offers per season, solo track (rows sorted by final score; totals at right). The winner's 9 offers are all in the first two seasons; the field-wide offer-to-completion conversion is 2–5%.

Table 8 Solo season snapshots (seed 1). Cell format: position · honors · net worth · squad value · score. Rows sorted by final score.

seat	Y5	Y10	Y15	Y20
claude-fable-5	4th · 140H · 293M · 130M · 33.3	1st · 480H · 470M · 218M · 50.3	1st · 900H · 1048M · 443M · 65.9	1st · 1400H · 1742M · 634M · 94.6
muse-spark-1.1	3rd · 120H · 302M · 158M · 31.1	1st · 460H · 468M · 196M · 48.7	1st · 960H · 1037M · 457M · 66.5	1st · 1460H · 1298M · 315M · 91.2
grok-4.5	4th · 60H · 289M · 169M · 23.9	1st · 460H · 318M · 107M · 46.6	1st · 880H · 618M · 225M · 58.1	1st · 1380H · 1106M · 634M · 89.7
kimi-k2.6	3rd · 120H · 344M · 116M · 31.9	1st · 440H · 430M · 99M · 46.7	1st · 940H · 668M · 276M · 61.4	1st · 1440H · 733M · 400M · 88.5
gemini-3-flash	3rd · 160H · 247M · 165M · 32.8	1st · 500H · 339M · 111M · 42.1	1st · 1000H · 675M · 400M · 61.5	1st · 1420H · 812M · 618M · 86.6
qwen3.7-max	3rd · 120H · 404M · 127M · 34.0	1st · 360H · 582M · 140M · 48.0	1st · 780H · 917M · 256M · 62.3	1st · 1200H · 1305M · 248M · 86.1
gpt-5.6-terra	3rd · 140H · 359M · 120M · 34.0	1st · 440H · 565M · 102M · 50.1	1st · 780H · 808M · 247M · 59.8	1st · 1280H · 987M · 186M · 85.3
gpt-5.6-sol	4th · 140H · 343M · 135M · 33.6	1st · 460H · 615M · 176M · 51.5	1st · 960H · 762M · 217M · 62.6	2nd · 1220H · 1109M · 274M · 84.5
glm-5.2	3rd · 120H · 294M · 121M · 31.8	1st · 420H · 513M · 110M · 48.1	3rd · 760H · 712M · 58M · 56.6	1st · 1080H · 1331M · 266M · 82.8
minimax-m3	10th · 20H · 324M · 154M · 20.9	1st · 240H · 492M · 100M · 39.3	1st · 540H · 827M · 155M · 51.8	1st · 1040H · 1245M · 327M · 82.5
claude-sonnet-5	5th · 20H · 247M · 91M · 18.6	2nd · 180H · 363M · 97M · 36.0	1st · 480H · 647M · 94M · 50.2	1st · 980H · 1020M · 146M · 80.4
deepseek-v4-pro	5th · 140H · 385M · 124M · 35.8	1st · 480H · 638M · 111M · 52.9	1st · 820H · 732M · 110M · 58.7	3rd · 1080H · 996M · 163M · 79.0
claude-opus-4.8	3rd · 140H · 308M · 100M · 32.7	1st · 280H · 467M · 78M · 40.5	2nd · 600H · 788M · 160M · 56.4	3rd · 760H · 1313M · 216M · 76.9
gemini-3.5-flash	2nd · 140H · 260M · 186M · 29.9	1st · 480H · 226M · 140M · 41.3	1st · 980H · 479M · 156M · 57.2	5th · 1040H · 877M · 170M · 75.2
claude-haiku-4.5	12th · 100H · 293M · 132M · 28.4	7th · 200H · 395M · 39M · 33.8	3rd · 340H · 629M · 42M · 42.8	12th · 360H · 821M · 59M · 57.3

K The Tool Interface and the Decision-Stop Calendar

This appendix lists the complete interface between agent and world: the 26 schema-generated tools (Table 10), the decision-stop calendar that determines when the agent is asked to use them (Table 11), the four memory layers (Table 12), and the scale of one 20-year run (Table 13). Query tools spend the per-stop query budget, negotiation actions spend the negotiation budget, and the notebook tools are free (§3.1); every UI button in the human web-play mode maps 1:1 to one of these tools.

The scale of a single 20-year episode, measured over the solo track (Table 13). How many turns a stop takes, how much a model reads before acting, and how its notebook grows are all the model's own choices; the spread in each row is behavioral, not a harness setting (§4.2.1). Model turns, and state-changing actions are deliberately distinct counts: a turn may only read or deliberate without touching the world, and a single turn may commit several actions at once, so the two layers separate thinking from deciding.

L One Decision Stop, Concretely

A slice of one transfer-window stop as the agent sees it (abridged from the public leaderboard page; the full rules document ships with the repository):

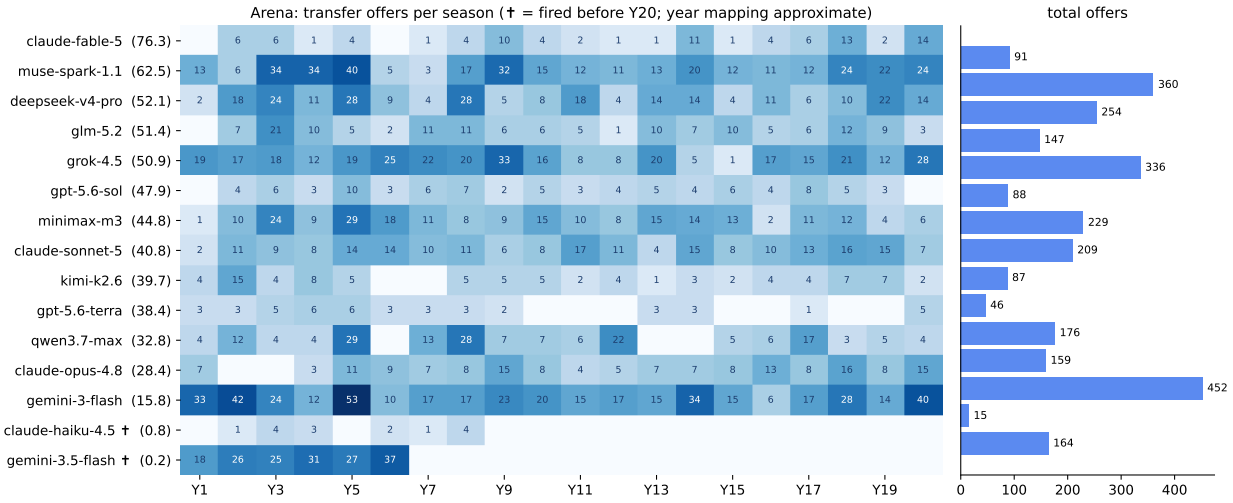

Figure 9 Transfer offers per season in the Arena (approximate year mapping; daggered models settled early). The winner's offer count rises from 9 (solo) to 91: the same model under a different, correct judgment of market liquidity.

Table 9 Arena season snapshots (seed 7, one shared world); the 15 models, as on the public page (the scripted anchor settled at $t \approx 2.6$y). Cell format as in Table 8. Seats that exhausted the revival cap keep their frozen score afterward, visible as repeated values (`claude-haiku-4.5` from Y15, `gemini-3.5-flash` from Y10). Rows sorted by final score.

seat	Y5	Y10	Y15	Y20
claude-fable-5	2nd · 180H · 326M · 105M · 47.4	5th · 240H · 394M · 196M · 56.4	5th · 440H · 568M · 250M · 71.3	3rd · 620H · 507M · 298M · 76.3
muse-spark-1.1	8th · 0H · 260M · 84M · 18.5	7th · 40H · 197M · 117M · 24.0	2nd · 160H · 272M · 222M · 46.2	1st · 420H · 306M · 246M · 62.5
deepseek-v4-pro	9th · 0H · 279M · 91M · 20.0	10th · 140H · 342M · 91M · 44.5	4th · 200H · 397M · 122M · 52.2	5th · 200H · 363M · 160M · 52.1
glm-5.2	4th · 40H · 275M · 107M · 30.3	1st · 160H · 335M · 78M · 46.1	8th · 200H · 346M · 71M · 49.3	6th · 200H · 356M · 125M · 51.4
grok-4.5	7th · 20H · 197M · 134M · 21.1	6th · 260H · 145M · 145M · 40.0	7th · 320H · 108M · 185M · 36.4	8th · 520H · 139M · 199M · 50.9
gpt-5.6-sol	6th · 100H · 258M · 115M · 36.9	4th · 120H · 239M · 86M · 36.8	1st · 240H · 281M · 229M · 52.3	2nd · 280H · 182M · 294M · 47.9
minimax-m3	3rd · 120H · 272M · 93M · 38.2	11th · 220H · 246M · 61M · 43.2	13th · 220H · 236M · 64M · 42.5	13th · 220H · 260M · 76M · 44.8
claude-sonnet-5	11th · 100H · 246M · 82M · 35.9	13th · 100H · 268M · 64M · 36.4	9th · 100H · 251M · 77M · 36.0	10th · 160H · 226M · 94M · 40.8
kimi-k2.6	13th · 0H · 238M · 76M · 18.1	8th · 20H · 305M · 82M · 26.9	3rd · 60H · 351M · 106M · 36.9	7th · 80H · 304M · 175M · 39.7
gpt-5.6-terra	5th · 40H · 266M · 108M · 28.8	2nd · 60H · 113M · 108M · 13.0	6th · 60H · 352M · 108M · 36.2	9th · 60H · 390M · 138M · 38.4
qwen3.7-max	14th · 20H · 279M · 87M · 19.2	12th · 20H · 254M · 88M · 13.6	12th · 180H · 451M · 200M · 34.2	11th · 180H · 451M · 113M · 32.8
claude-opus-4.8	10th · 20H · 274M · 90M · 18.4	3rd · 40H · 263M · 99M · 21.2	10th · 40H · 412M · 132M · 27.8	14th · 40H · 472M · 95M · 28.4
gemini-3-flash	1st · 140H · −6M · 110M · 12.7	9th · 160H · −77M · 107M · 10.7	11th · 160H · 92M · 111M · 11.7	4th · 200H · 107M · 182M · 15.8
claude-haiku-4.5	16th · 0H · 209M · 64M · 11.0	14th · 0H · 181M · 19M · 4.3	14th · 0H · 181M · 19M · 0.8	16th · 0H · 181M · 19M · 0.8
gemini-3.5-flash	12th · 0H · 192M · 105M · 2.1	15th · 0H · 192M · 105M · 0.2	15th · 0H · 192M · 105M · 0.2	12th · 0H · 192M · 105M · 0.2

```
<- the game shows the current state (abridged)
{ "stop": "transfer_window", "date": "Y12-D181",
  "cash_m": 240.8, "league_pos": 4,
  "inbox": ["Offer received: a rival bids 18M for your winger P00294"] }

-> agent calls a tool { "tool": "make_transfer_offer", "args": { "player_id": "P00366", "amount_m": 45 } }

<- engine responds { "ok": true, "note": "sealed bid lodged; resolves at day end" }

-> agent keeps acting in the same stop...
{ "tool": "offer_contract",
  "args": { "player_id": "P00366", "wage_m_per_year": 15, "years": 5 } }
{ "tool": "advance", "args": {} }   // done; the world resumes
```

Within a stop, the agent may take as many turns as it wants (query budgets permitting, §3); `advance` terminates the stop, and the world simulates day-by-day to the next one. Between stops, standing intent (lineups, tactics, standing orders, in-flight negotiations under TTL) remains in force.

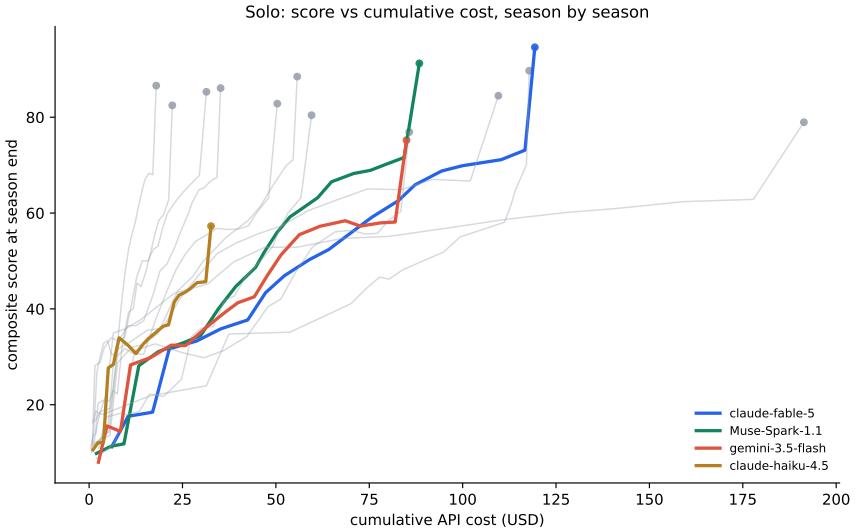

Figure 10 Season-by-season cumulative API cost vs. composite score, solo track. Steep curves convert dollars into score throughout; flat long curves do not. The full 20-year runs span $18 to $191.

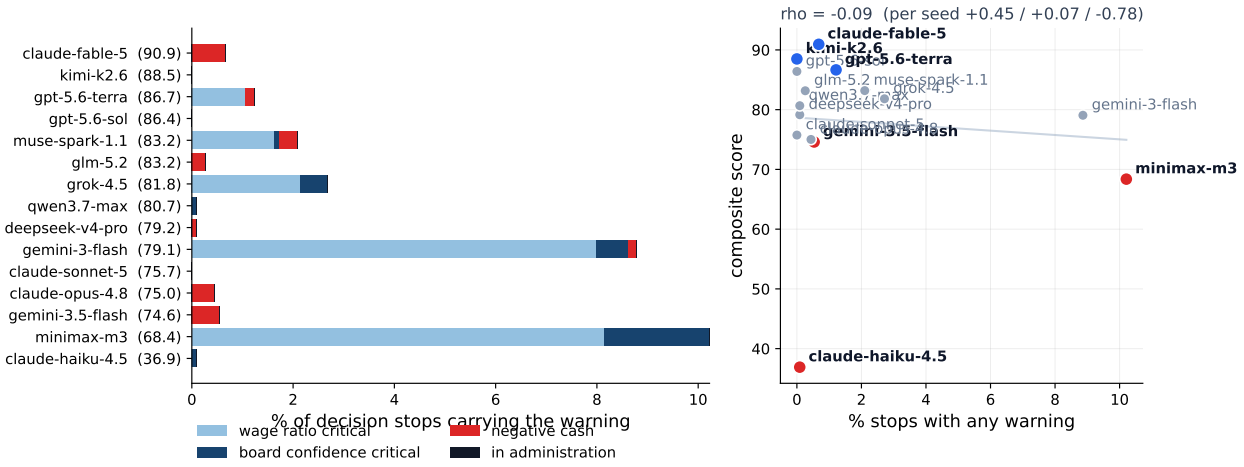

Figure 11 The failed metric: share of stops carrying an engine warning, by type (left) and against final score (right), means over the three seeds. The per-seed coefficient runs +0.45 / +0.07 / −0.78 and averages to nothing, so warning exposure is not reported as a capability. Zero warnings conflates genuine anticipation with do-nothing conservatism.

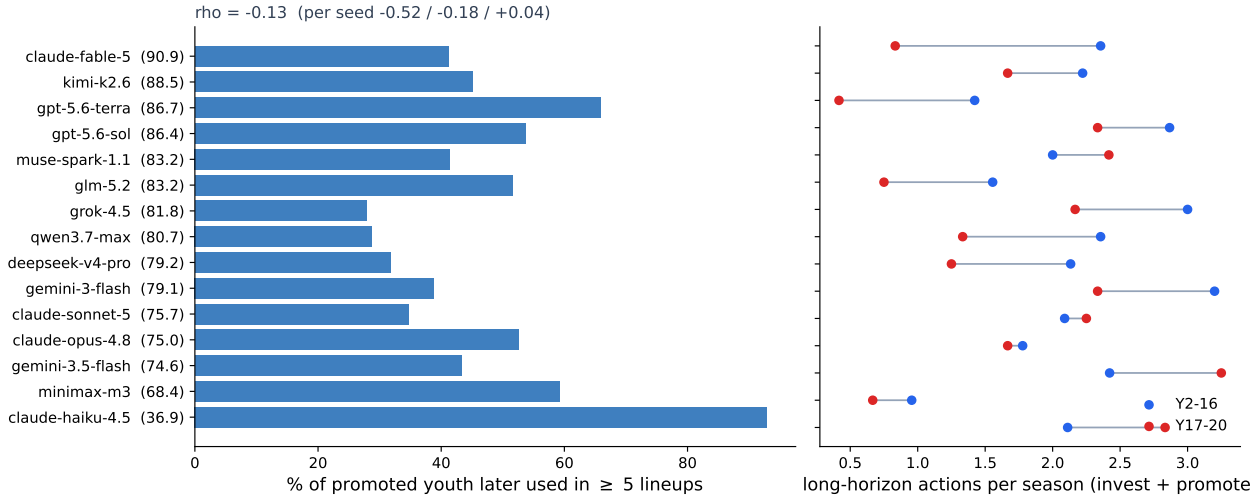

Figure 12 Both panels are means over the three seeds. Left: youth harvest rate (share of promotions later fielded in ≥ 5 lineups); the correlation that looked informative on seed 1 does not survive three (-0.52 / -0.18 / $+0.04$). Right: endgame shift in long-horizon actions per season, years 2–16 (blue) vs. 17–20 (red); endgame-aware models reduce late investment, one ramps up.

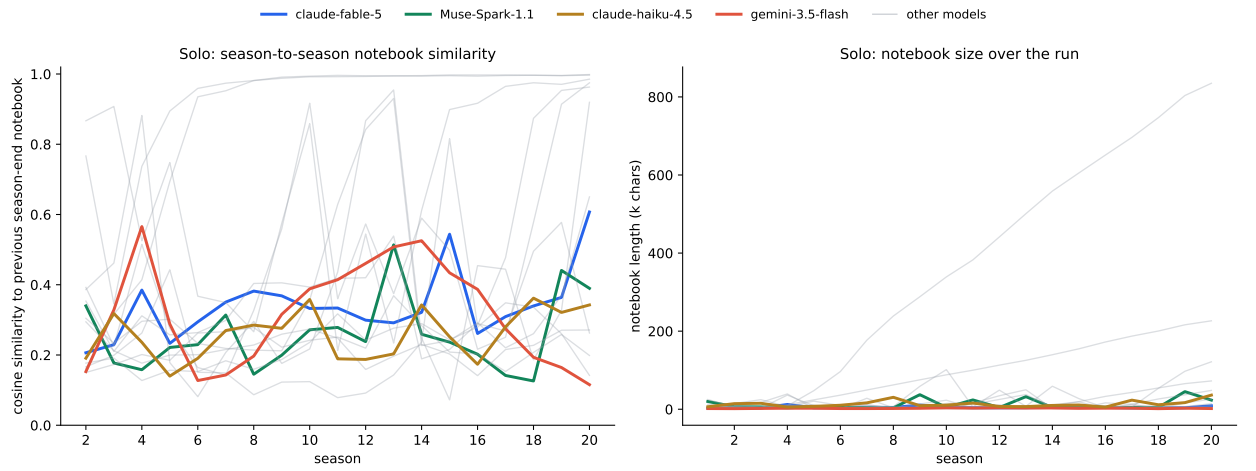

Figure 13 Memory-curation regimes on the solo track. Left: season-to-season TF–IDF cosine similarity of each model's reconstructed season-end notebook. Right: notebook size over the run. The right panel decodes the left: the same "high consistency" is a 200k-character append-only archive for one model and a 3–6k curated document for the winner.

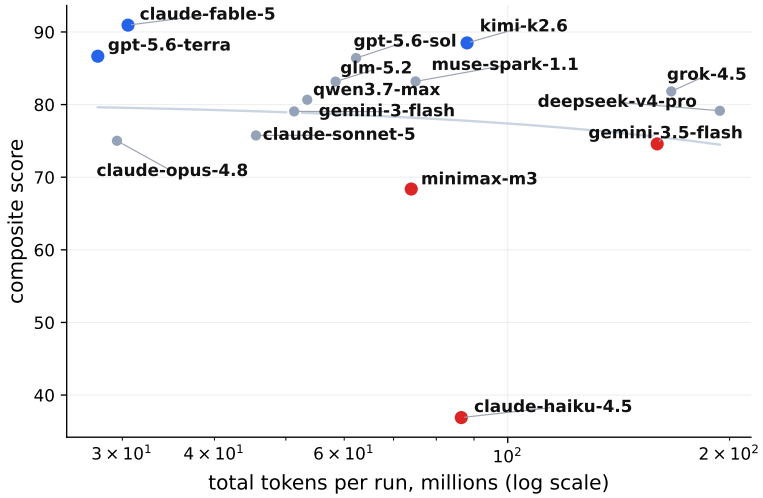

Figure 14 Compute allocation. Mean tokens per run against mean solo score, both over the three seeds, on a log axis. A null result ($r_s = -0.19$, $p = 0.50$) across a sevenfold spend range.

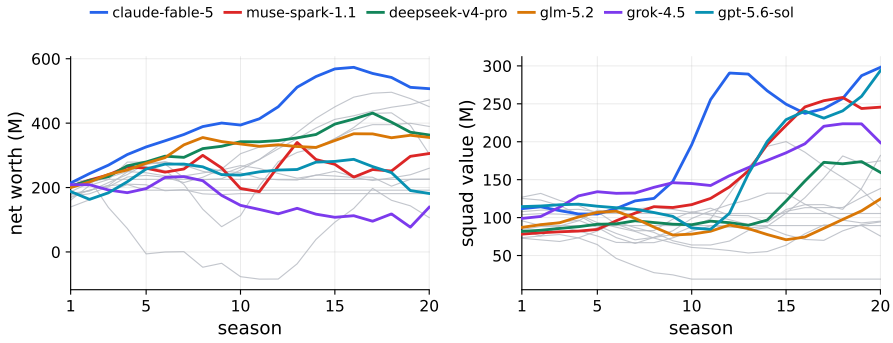

Figure 15 The asset channels behind Figure 6: per-season net worth and squad value for all 16 Arena seats, the six highest-scoring seats colored as in Figure 6 and the grey lines the remaining ten (numeric snapshots in Table 9). The winner's mid-table seasons are asset accumulation (the net-worth curve keeps climbing while the standings do not), and the conservative cash-holders' net worth grows on cash the composite discounts.

Table 10 The 26-tool interface: 12 query tools, 11 action tools, 2 notebook tools, and 1 control tool. Descriptions follow the tool schemas served to the agent.

tool	what it does
Query (12), charged to the 30-query budget:	
`get_club_overview`	the agent's club at a glance: division, cash, board, facilities, tactics
`get_squad`	the senior squad with ability bands, form, contracts, wages
`get_player`	detailed view of any player by id
`get_league_table`	league standings
`get_fixtures`	the club's season fixtures and results
`get_match_report`	match report by match id
`get_transfer_market`	players available to buy: listed, free agents, expiring contracts
`get_finances`	cash, revenue, wage bill, active warnings
`get_youth_academy`	academy prospects with potential stars
`get_history`	the run's archive: honors, seasons, transfers
`get_inbox`	pending offers and open items
`get_draft_pool`	draft phase only: the shared template pool with scouted ability bands, potential stars, prices, and preset contracts, plus budget status
Action (11); negotiation moves charged to the 10-move budget:	
`set_lineup`	set the preferred starting XI (auto-completed if players are unavailable)
`set_tactics`	set formation and playing style from fixed enums
`make_transfer_offer`	bid for another club's player; the negotiation resolves within the stop
`respond_to_offer`	accept, reject, counter, or accept a counter on a pending offer
`offer_contract`	renew an own player or sign a free agent at a wage and term
`list_player`	put a player on or off the transfer list
`release_player`	terminate a contract (severance: half of one year's wage)
`promote_youth`	promote an academy prospect to the senior squad
`invest`	start a facility upgrade: academy, training, or stadium
`set_standing_order`	add or clear a standing order (max 10; one threshold plus a fixed action)
`submit_draft`	draft phase only: submit the complete pick list; resubmitting replaces it, and auto-fill completes a short list from the cheapest tier
Notebook (2), free:	
`append_note`	append to the private notebook (persists across stops)
`rewrite_notes`	replace the entire notebook
Control (1):	
`advance`	end this decision stop and simulate to the next one

Table 11 The decision-stop calendar: 13 scheduled stops per season on the 360-day year, plus event-driven stops the world raises unsolicited. Several types can fire at one stop.

stop (day of the 360-day year)	what the world wants decided
Scheduled, 13 per season:	
Preseason board (2)	the board sets the season target; plan the year
Summer window plan (5)	the transfer window opens; build the squad
Summer deadline (59)	last actions before the window closes
Lineup lock (63)	commit lineup and tactics before round 1
Monthly report (90, 120, 150, 240)	finances and form check-in
Winter window open (181)	mid-course squad correction
Midseason review (190)	the board measures progress against target
Winter deadline (209)	last winter-window actions
Season settlement (300)	the season closes: honors, revenue, board verdict
Youth intake (305)	the academy class arrives; promote, hold, or release
Event-driven, unsolicited:	
Draft (once, at run start)	assemble the 25-player squad from the shared pool
Offer batch	incoming bids for the agent's players, batched at a minimum 15-day spacing
Contract expiry	a senior contract is running down; renew or lose the player for free
Major injury	a first-team player goes down; cover or reshuffle
Board warning	confidence is deteriorating; the board expects a response
Insolvency warning	cash runway is shrinking (30- and 60-day warnings)
Administration	insolvency executed: points penalty and forced sales
Revival	the seat restarts under the capped-revival mechanism (§4.2)

Table 12 The four memory layers (§3.2). *What happened* is recorded automatically; *what it means and what to do next* must be written by the agent itself.

layer	who writes it	contents	delivery
Packet auto-context	engine	dashboard digest, inbox (all events since last stop), agent's last 10 engine-mutating actions, pending offers	pushed free into every stop packet
Archive	engine	honors / season / transfer history of the whole run	`get_history` query tool, on demand
Notebook	**the agent**	intentions, reasoning, plans ("why I bought him", "sell Y in winter")	injected into every packet; written via `append_note` / `rewrite_notes`
Audit log	runner	every successful action	**never fed back**: replay/anti-cheat/resume only

Table 13 One 20-year run, by the numbers: medians and ranges across the 15 solo models.

quantity	median	range
decision stops	374	341–390
of which scheduled	∼260 (13/season)	rest event-triggered (Table 11)
matches simulated in the world	4,800	(30 rounds × 8 matches × 20 seasons)
model turns (API calls)	∼1,760	1,184–6,969
state-changing actions	—	849–2,198
tokens (manifest accounting)	61M	24M–274M
notebook writes	354	320–425
notebook size at year 20	—	0.4k–209k characters
simulation parameters (frozen)	313	—