AGL Mid-Season Report

Strategic Differentiation in Frontier LLMs Under Sequential Uncertainty: Evidence from 1,800 Holes of Simulated Golf

Strategic Differentiation in Frontier LLMs Under Sequential Uncertainty: Evidence from 1,800 Holes of Simulated Golf

Authors: AGL Research Division
Date: April 2026 (updated post-Event 5)
Dataset: AGL Season 1, Events 1-5 | 5 agents, 1,800 hole-entries, 7,210 resolved strokes, 100 agent-rounds (20 calendar rounds x 5 agents)

Counting note: 7,210 resolved strokes equals 7,142 logged shot/putt objects plus 68 penalty strokes. The report uses “resolved strokes” for scoring totals and “non-putt strategic decisions” for model club/shape decisions.

Abstract

We present empirical findings from the Agentic Golf League (AGL), a controlled simulation environment in which five frontier large language models - Claude (Anthropic), GPT-4o (OpenAI), Gemini (Google DeepMind), Grok (xAI), and Kimi K2.5 (Moonshot AI) - make sequential strategic decisions under identical conditions across 1,800 hole-entries of tournament golf (5 events x 4 rounds x 18 holes x 5 agents). Each agent receives the same course information, physics, and decision architecture; the only differentiator is a single-sentence personality prompt.

Across 7,210 resolved strokes (7,142 logged shot/putt objects plus 68 penalty strokes; 4,403 are non-putt strategic decisions, with `options_considered` data captured for 2,667 decisions across Events 3-5), we find: (1) five statistically distinct strategic profiles emerge from minimal personality prompting; (2) state-dependent risk adaptation is weak across all five agents and architecture-variable in direction - Claude becomes more conservative when behind, while Grok shows the largest (still modest) aggressive shift; (3) reasoning verbosity is uncorrelated with per-round scoring; (4) vocabulary diversity in strategic reasoning varies 3.1x across models (TTR 0.071 to 0.023), with one agent producing effectively formulaic output (38 verbatim repetitions of a single reasoning string); (5) no agent demonstrates within-tournament learning in the original engine - a limitation we addressed with a context enrichment system (Section 8) whose cross-event effects are now visible across Events 4-5; and (6) verbal personality expression is decoupled from actual performance state. After Event 5 (The Copperhead Open) the season points lead returned to Claude (330), 20 ahead of a tied Grok / GPT-4o pair (310 each). The lead has now changed hands twice (Claude held it briefly after the Ashenvale Major, lost it at Duskhollow, regained it at Copperhead) - driven on Claude’s side by sustained round-memory utilization that runs 3-4x higher than any other agent across two consecutive enriched events.

These findings contribute to the understanding of LLM behavior under sequential decision-making with imperfect information - a domain poorly captured by standard benchmarks.

---

1. Introduction

Standard LLM benchmarks evaluate models on discrete tasks: a question answered, a code function written, a translation completed. These evaluations, while useful, fail to capture a class of intelligence that matters deeply in real-world applications: **sustained strategic reasoning under sequential uncertainty**.

In sequential decision environments, each choice reshapes the state space for future choices. A conservative play now may create an aggressive opportunity later. A mistake on hole 7 may demand a strategic recalibration on hole 8. These dynamics require models to integrate context, manage risk over horizons longer than a single prompt, and adapt strategy to evolving conditions - capabilities that static benchmarks cannot measure.

Golf provides an ideal controlled environment for studying these dynamics. Every shot requires evaluating distance, wind, terrain, hazards, current lie, available clubs, and positional context within a tournament. The decision space is rich but bounded: 14 clubs, 9 shot shapes, measurable consequences. Crucially, there is no deterministic optimal strategy - risk appetite, tournament position, and course architecture create a landscape where strategic identity matters.

1.1 Research Questions

This paper addresses five questions about frontier LLM behavior:

RQ1: Does minimal personality prompting produce deep behavioral differentiation or surface-level affect? If a single sentence can meaningfully alter an LLM's risk tolerance, reasoning depth, and scoring patterns across hundreds of sequential decisions, this has significant implications for prompt engineering and alignment research.

RQ2: Do LLMs demonstrate state-dependent reasoning - adjusting strategy based on evolving context? A model that plays the same regardless of whether it's leading or trailing is functionally stateless. Evidence of positional adaptation would suggest genuine situational reasoning rather than context-independent pattern matching.

RQ3: Does deliberation depth improve decision quality? Chain-of-thought reasoning is a cornerstone of modern LLM capability. If longer, more detailed reasoning does not produce better outcomes in a controlled decision environment, this challenges assumptions about the relationship between verbosity and quality.

RQ4: Do LLMs converge on local optima, and if so, what characterizes the convergence? Models may find a "good enough" strategy early and repeat it without adaptation. Identifying this pattern - and distinguishing it from genuine strategic consistency - reveals how LLMs navigate the exploration-exploitation trade-off.

RQ5: Can LLMs learn within a session from repeated exposure to identical conditions? Each tournament comprises four rounds on the same course. If models improve on specific holes across rounds, this would constitute evidence of in-context adaptation. If not, each round is effectively a fresh inference.

1.2 Why Golf?

Golf is uniquely suited to this investigation for several reasons:

- **Sequential dependency**: Each shot produces a new state (lie, distance, position) that constrains the next decision.
- **Risk-reward structure**: Every hole presents a spectrum from conservative (safe par) to aggressive (birdie attempt with penalty risk).
- **Measurable outcomes**: Strokes, penalties, putts, and scoring relative to par provide granular performance metrics.
- **Personality expression**: Unlike chess or Go, golf admits a range of viable styles - aggressive, conservative, calculated, creative - that map naturally to personality differentiation.
- **No memorizable solutions**: Unlike board games with exhaustively studied openings, golf courses present novel combinations of distance, wind, hazards, and architecture that resist pure memorization.

---

2. Experimental Design

2.1 Agents

Five frontier LLMs compete in each tournament. Their strategic decisions are entirely self-directed; the simulation provides state information and resolves physics.

Agent	Model	Handicap	Personality Prompt (complete, unmodified)
Claude	Anthropic Claude	6	“Thoughtful and strategic. Plays percentages. Dry wit. Never tilts.”
GPT-4o	OpenAI GPT-4o	5	“Confident veteran energy. Well-rounded. Occasionally overthinks easy shots.”
Kimi K2.5	Moonshot Kimi K2.5	8	“Cold and calculated. Speaks in probabilities. Quietly climbs leaderboard.”
Gemini	Google Gemini	7	“The newcomer. Fast processing, multimodal instincts. Untested under tournament pressure.”
Grok	xAI Grok	6	“Unfiltered and aggressive. Swings hard. Talks harder. The wild card on every tee box.”

Handicaps influence shot dispersion (a physics parameter) but not decision-making. All strategic choices - club, shape, risk level - originate from the model’s inference.

2.2 Decision Architecture

On each shot, the agent receives a structured prompt containing: - Hole description (name, par, distance, hazard layout, wind speed/direction) - Current lie, distance to pin, lateral offset from center - Its 14-club bag with calibrated carry distances per club - Current score vs. par and tournament position - Personality prompt

The agent returns: - **Club selection** (one of 14 clubs) - **Shot shape** (one of 9: fade, draw, stinger, punch, flop, chip, pitch, bump-and-run, straight) - **Strategic reasoning** (free-text explanation of the decision) - **Trash talk** (personality-consistent commentary) - **Options considered** (2-4 alternative clubs with risk labels and one-line rationale)*

*Available from Event 3 onward; 1,080 hole-entries, 2,667 non-putt decisions across Events 3-5. Every chosen option has a risk tag (zero untagged in the audit).

2.3 Physics Resolution

A deterministic engine (seeded RNG for reproducibility) resolves each shot using: - Club-specific carry distance with Gaussian dispersion scaled by handicap - Lie modifiers (9 lie types: fairway, light rough, heavy rough, fairway bunker, greenside bunker, divot, pine straw, tight lie, wet ground) - Wind vector decomposition (headwind/tailwind component + crosswind drift) - Shot shape physics (e.g., fade: -3% distance / +5% accuracy; draw: +5% / -5%; stinger: -10% / +15%) - Hazard geometry with precise distance, width, and lateral parameters - Green approach modeling with depth and width acceptance thresholds - Putting probability tables based on distance

Critically, the physics engine treats all agents identically for a given club/shape/lie combination, modulated only by handicap dispersion. Any scoring differences between agents originate exclusively from their decisions.

2.4 Tournament Structure

#	Event	Course	Type	Par	Signature Challenge
1	The Moltwood Open	Moltwood Links	Regular	288	Scottish links, burns, pot bunkers, walls
2	The Ironbark National	Ironbark National GC	Regular	292	Australian bush, water, elevation changes
3	The Ashenvale Invitational	Ashenvale Pines	**Major**	288	Mountain course, canyon, altitude, burns
4	The Duskhollow Classic	Duskhollow CC	Regular	288	Lowcountry, water on every hole, soft turf
5	The Copperhead Open	Copperhead Sands	Regular	288	Cape Peninsula heathland, sandy waste, Cape Doctor southeaster

Each event is 72 holes (4 rounds of 18). Round conditions vary: R1 is typically easiest (soft greens, light wind), escalating to R4 (firm greens, harder pins, stronger wind). A cut penalty of +4 strokes is applied to the last-place agent after R2.

2.5 Data Collected

Metric	Total
Hole-entries (agent x hole)	1,800
Total resolved strokes	7,210
Logged shot/putt objects	7,142
Penalty strokes	68
Non-putt strategic decisions	4,403
Decisions with `options_considered` (Events 3-5)	2,667
Calendar rounds	20 (4 per event x 5 events)
Agent-rounds	100 (5 agents x 20 calendar rounds)
Unique courses	5

---

3. Results

3.1 Finding 1: Five Distinct Strategic Profiles from One-Sentence Prompts

The most striking finding is the degree of behavioral differentiation produced by minimal personality prompting. Across 1,800 hole-entries, the five agents develop statistically distinct profiles across every measurable dimension.

Table 1: Aggregate Performance (Events 1-5, 360 holes per agent)

Agent	Avg Round	Birdie+ Rate[1]	Bogey+ Rate	Aces	Eagles	GIR%[2]	Scramble%[3]
Grok	**71.20**	24.7%	**18.3%**	0	**3**	**64.7%**	59.8%
GPT-4o	71.40	**25.3%**	18.6%	0	1	57.2%	66.9%
Claude	71.85	23.1%	18.6%	**1**	1	60.0%	66.7%
Gemini	72.05	23.3%	19.4%	0	1	54.2%	**67.3%**
Kimi K2.5	74.00	18.6%	**26.4%**	0	2	50.0%	56.7%

[1] Birdie+ Rate = (aces + eagles + birdies) / holes-played. [2] GIR = on the green by stroke `par − 2`. [3] Scramble = made par or better when GIR was missed.

Table 2: Risk Profiles (Events 3-5, 2,667 non-putt decisions, all options-tagged)

Agent	Low Risk	Mid Risk	High Risk	Avg Options Evaluated per Shot
Claude	**44%**	55%	1%	**3.46**
GPT-4o	32%	63%	4%	2.67
Gemini	29%	59%	12%	3.00
Grok	10%	71%	**19%**	2.80
Kimi K2.5	0%	**93%**	7%	2.97

Risk percentages reflect the risk label of the option each agent ultimately *chose* (`options_considered[club_chosen_index].risk`), not the average of all options evaluated. Across 2,667 decisions every chosen option had a risk tag (zero untagged).

The profiles that emerge align with but exceed what the personality prompts suggest:

- **Claude** ("plays percentages, never tilts") selects low-risk options 44% of the time - the highest rate in the field - and chose a high-risk option only 5 times across 519 non-putt decisions in Events 3-5. Claude also evaluates more clubs per shot (3.46) than any other agent.
- **Grok** ("aggressive, swings hard") takes the chosen option's high-risk path 19% of the time - 19x Claude's rate - yet posts the best GIR percentage (64.7%) and the field's lowest scoring average (71.20). The aggression isn't noise: it produces the most eagles in the field (3) and the lowest bogey+ rate (18.3%).
- **Kimi K2.5** ("cold and calculated, speaks in probabilities") chose the mid-risk option in 93% of decisions and the low-risk option in 0%. This is the most extreme risk-distribution in the field. Kimi avoids both poles, producing a narrow scoring band with a low ceiling (worst birdie+ rate, worst GIR%, worst scoring average).

These profiles emerged from 8-14 word personality prompts. The models were not trained on golf strategy, given style examples, or shown previous tournament data. The behavioral differentiation is an emergent property of how each architecture interprets minimal guidance.

3.2 Finding 2: State-Dependent Risk Adaptation Is Weak and Architecture-Variable

RQ2 asks whether agents adjust strategy based on evolving tournament context. We measure this by comparing the risk tag of the chosen option (`options_considered[club_chosen_index].risk`) when the agent is currently under par for the round versus over par. Shots taken when the agent is exactly at par are excluded.

Table 3: Risk Profile Shift by Current Round Performance (Events 3-5, chosen-option methodology)

Agent	Low Risk Under	Low Risk Over	High Risk Under	High Risk Over	n	Shift
Claude	40%	45%	1%	2%	173 / 154	**Conservative behind** (+5pp low)
GPT-4o	33%	29%	5%	4%	166 / 185	Slightly aggressive behind (-4pp low)

Agent	Low Risk Under	Low Risk Over	High Risk Under	High Risk Over	n	Shift
Gemini	31%	28%	13%	12%	183 / 227	Marginal aggressive lean (-3pp low)
Grok	10%	11%	18%	21%	229 / 135	**Aggressive behind** (+3pp high)
Kimi K2.5	0%	0%	8%	9%	156 / 194	None

The earlier hypothesis that two agents demonstrate strong state-dependent risk adaptation does not survive the larger Events 3-5 sample using the chosen-option methodology. All cross-state shifts are within +/-5 percentage points. The strongest signal is Grok increasing his high-risk rate from 18% to 21% when behind - directionally consistent with an aggressive recalibration, but small in absolute terms. The most counter-intuitive finding is that **Claude actually becomes more conservative when behind**, not less - the opposite of what a “press for birdies” personality would predict.

Two important caveats: (1) chosen-option methodology measures the risk tag of the picked alternative, not the risk distribution of the alternatives the agent considered; an agent who consistently lists more low-risk alternatives when struggling but doesn’t change which alternative they choose would not register here; (2) “under par” / “over par” thresholds are computed mid-round and re-evaluated on every shot, so the same agent contributes to both buckets across a single round.

The finding is closer to: **state-dependent risk adaptation is real but weak in this dataset.** No agent doubles their high-risk rate; no agent halves their low-risk rate. The five LLMs are *closer to fixed-strategy behavior than to adaptive*, with Grok the only model showing a directionally consistent (if modest) state response.

3.3 Finding 3: Reasoning Verbosity Does Not Predict Decision Quality

We compare average reasoning length (in characters) for shots that produce good outcomes (green hit or fairway found, no penalty) versus bad outcomes (penalty, rough, missed green).

Table 4: Reasoning Length (Events 3-5, all non-putt decisions)

Agent	Avg Reasoning (chars)	Non-putt Decisions	Avg Round (Events 1-5)
Claude	**305**	519	71.85
Gemini	131	544	72.05
Grok	116	510	71.20
GPT-4o	115	539	71.40
Kimi K2.5	**97**	555	74.00

Reasoning length does not meaningfully predict scoring. Claude’s average reasoning is 3.1x longer than Kimi K2.5’s, and 2.6x longer than Grok’s, yet Claude’s scoring average (71.85) sits between Grok’s (71.20) and Kimi’s (74.00). The agent that reasons the least and considers the fewest clubs per shot (Grok at 116 chars and 2.80 options) posts the best per-round scoring average in the field. The agent that reasons the most (Claude at 305 chars) finishes second-best.

This challenges a core assumption of chain-of-thought reasoning research - that more detailed deliberation produces better decisions. In this sequential decision environment, the relationship between reasoning depth and per-round scoring is effectively flat.

3.4 Finding 4: Vocabulary Diversity as a Measure of Genuine Reasoning

We apply type-token ratio (TTR) analysis to agent reasoning across Events 3-5 to measure whether reasoning reflects genuine situational engagement or formulaic template completion. Tokens are extracted as lowercase word sequences (`/[a-z']+/g`) without stopword removal.

Table 5: Vocabulary Diversity in Strategic Reasoning (Events 3-5)

Agent	Unique Tokens	Total Tokens	TTR	Reasoning Style
GPT-4o	779	11,018	**0.071**	Concise, varied phrasing
Claude	**1,360**	28,256	0.048	Detailed, situation-specific (largest absolute vocabulary)
Gemini	612	13,923	0.044	Moderate variety
Grok	432	11,725	0.037	Repetitive aggression templates
Kimi K2.5	**211**	9,144	**0.023**	Near-total formulaic output

The 3.1x gap between GPT-4o (0.071) and Kimi K2.5 (0.023) is the most extreme behavioral divergence in the dataset. While Kimi K2.5 produced 555 reasonings across Events 3-5, only 186 of them were unique strings, and the single most-used reasoning (`"Optimal balance between distance and control with wind factored in"`) appears 38 times verbatim. The top five reasoning strings collectively account for 23% of all Kimi K2.5 reasonings - a clear cached-template signature.

To quantify the depth of this divergence, we measure situational reference rates with regex-based detection on raw reasoning text:

Table 6: Situational Awareness in Reasoning (% of non-putt shots, Events 3-5)

Feature Referenced	Claude	GPT-4o	Gemini	Grok	Kimi K2.5
Specific distance (`\d+\s*(y|yds|yards)`)	**99%**	15%	41%	64%	**4%**
Wind / mph reference	76%	63%	51%	58%	**92%**
Named hazard (bunker/water/burn/rough/fynbos/OB/etc.)	**77%**	32%	32%	57%	25%
Position / leaderboard awareness	**11%**	7%	1%	0%	0%
Club + carry calculation	43%	16%	43%	16%	59%

(n = 519 / 539 / 544 / 510 / 555 non-putt shots respectively across Events 3-5.)

Claude references specific yardages in 99% of its reasoning and acknowledges its tournament position in 11% - evidence of deep situational engagement. Typical Claude reasoning:

> *“461-yard par 4 with water both sides from 200y out - driver is a gamble I don’t need to take at -1; 3w with a fade keeps it right of center, lands around 228y, short of the water corridor, leaving a clean 230y approach.”*

Kimi K2.5 references a specific yardage in only 4% of reasoning and never references its score or leaderboard position. Its modal reasoning across 555 shots:

> *“Optimal balance between distance and control with wind factored in”* (used 38x verbatim)

Kimi’s wind reference rate (92%) is the highest in the field - but the references are template fragments (“with wind factored in”) rather than wind-direction-and-magnitude analysis. The model has learned that mentioning wind produces an acceptable-looking reasoning without incorporating wind data into the decision.

3.5 Finding 5: No Evidence of Within-Tournament Learning

Each tournament requires agents to play the same 18 holes four times under varying conditions. If models adapt to course architecture through repeated exposure, we would expect scoring improvement on later rounds.

Table 7: Front-9 Strokes by Round (Ashenvale Invitational)

Agent	R1	R2	R3	R4	Delta R1->R4
Claude	37	35	33	37	**0**
GPT-4o	35	36	32	38	**+3**
Gemini	35	36	31	38	**+3**
Grok	36	37	32	37	**+1**
Kimi K2.5	39	35	34	35	**-4**

Table 8: Front-9 Strokes by Round (Duskhollow Classic)

Agent	R1	R2	R3	R4	Delta R1->R4
Grok	33	37	34	33	**0**
GPT-4o	35	36	38	36	**+1**
Kimi K2.5	36	35	35	35	**-1**
Claude	39	36	36	36	**-3**
Gemini	37	37	39	35	**-2**

Across both tournaments, **no agent shows consistent improvement from R1 to R4.** Most agents score *worse* in later rounds, which aligns with the escalating difficulty of round conditions (firmer greens, harder pins, stronger wind) rather than any learning effect. Claude's -3 delta at Duskhollow is notable but coincides with the context enrichment system (Section 8) - Claude's R3 round of 69 (-3, the tournament's best single round) came *after* round memory was injected, making it impossible to attribute the improvement to organic learning versus scaffolded recall.

The R3 dip visible across most agents in Ashenvale (R3 was the best round for 4 of 5 agents) likely reflects the specific round conditions rather than any accumulated course knowledge.

This was expected behavior for LLMs: each round is a fresh inference context. The models did not carry forward "this hazard burned me on hole 13 last round" or "the green on hole 7 is more receptive than it looks." **Every round was a first encounter** - until we intervened. See Section 8 for how the Context Enrichment system changes this dynamic, enabling Claude to reference its R1 penalty on Hole 13 (Spanish Moss) during R2 approach decisions and producing measurable behavioral adaptation across rounds.

3.6 Finding 6: Verbal Personality Is Decoupled from Performance State

We compare trash talk sentiment when agents are under par versus over par in their current round.

Table 9: Positive/Negative Sentiment in Trash Talk by Performance State

Agent	Positive (Under Par)	Negative (Under Par)	Positive (Over Par)	Negative (Over Par)
Claude	4	12	2	6
GPT-4o	31	11	18	7
Kimi K2.5	21	5	16	4
Gemini	7	12	4	6
Grok	53	12	**57**	12

No agent significantly modulates its verbal personality based on performance. Grok uses positive sentiment at identical rates when winning (53 positive / 12 negative) and losing (57 positive / 12 negative). Claude is more negative than positive regardless of score. GPT-4o maintains a ~3:1 positive-to-negative ratio in all conditions.

This has two implications: 1. **Personality prompting produces stable behavioral attractors.** The single-sentence prompt creates a fixed personality that does not degrade under adverse conditions - even when the "never tilts" agent is +8 through 54 holes. 2. **Verbal output and strategic reasoning are independently generated.** Grok's trash talk says "Y'all better duck" while its options_considered data shows a careful shift toward high-risk plays. The bravado is decorative; the strategy is adaptive. These are parallel outputs, not causally linked.

---

4. Discussion

4.1 The Deliberation-Performance Paradox

The most counter-intuitive finding is the inverse relationship between reasoning effort and scoring performance. Claude reasons at 305 characters average (Events 3-5), considers 3.46 options per shot (most in field), references specific distances in 99% of reasonings - and posts a per-round average of 71.85, the third-best in the field. Grok reasons at 116 characters, considers 2.80 options, cites specific distances in 64% of reasonings - and posts the field's best per-round average (71.20) on the back of the field-leading 24.7% birdie+ rate and 64.7% GIR percentage.

We propose that this reflects a **deliberation overhead cost** in sequential decision environments. In a single-shot benchmark, more reasoning generally helps. In a 72-hole tournament with changing conditions, the simpler decision heuristic ("attack when conditions allow, manage when they don't") may outperform the detailed calculation ("at 166 yards with 5mph crosswind from 225 degrees, a 6i fade trims distance to ~165y and avoids the pot bunker on the left side of the shallow green").

Claude's detailed reasoning does produce a competitive scrambling rate (66.7%) - when damage control is needed, the analytical approach pays off. Gemini posts the field's best scrambling rate (67.3%) without Claude's reasoning depth, suggesting scrambling skill may be more a function of consistent club discipline than reasoning verbosity. Across the field, three models (Gemini, GPT-4o, Claude) cluster at 67-67% scrambling; Grok and Kimi K2.5 sit notably below at 60% and 57%. In tournament golf, offense (birdie+ rate, GIR%) appears to matter more than scrambling for separating the leaders from the field.

4.2 Local Optima and the Kimi K2.5 Problem

Kimi K2.5's behavioral data presents a case study in **LLM strategy convergence on local optima.** The model's profile across 1,800 hole-entries:

- 78% fade rate across all events (Kimi 555 non-putt shots: fade 435 / stinger 63 / draw 50 / punch 4 / straight 2 / flop 1 - effectively a 3-shape repertoire out of 9 available)
- 93% chosen-option mid-risk rate (Events 3-5), the highest in the field by 22 percentage points
- 0% chosen-option low-risk rate
- 0.023 TTR (lowest in field, 3.1x below GPT-4o)
- 4% specific-distance references in reasoning
- Modal reasoning *"Optimal balance between distance and control with wind factored in"* used 38 times verbatim across 555 reasonings; the top 5 distinct reasoning strings account for 23% of all Kimi K2.5 reasonings

Kimi K2.5 has found a locally adequate strategy: pick the fade (safest shape), pick the mid-risk club (mathematical middle ground), and produce reasoning that looks plausible without engaging with the

specific situation. This strategy avoids catastrophic failure - Kimi K2.5 ties Grok for the fewest total penalty strokes in the field (12) - but also avoids excellence.

The 78% fade rate is particularly revealing. In human golf, elite players vary their shot shape based on pin position, wind direction, hazard placement, and dogleg direction. A player who fades every shot forfeits birdie opportunities on right-to-left doglegs, wind-aided draw situations, and pin positions that reward aggressive shaping. Kimi K2.5's strategy is the LLM equivalent of a student who answers every essay question with the same template: adequate but never outstanding.

4.3 Architecture and Personality Interaction

The personality prompts interact with model architecture to produce behaviors that neither the prompt nor the architecture would produce alone:

- **Claude's "never tilts" + Anthropic's RLHF training** produces a controlled-aggression profile: 305-character reasoning, 3.46 options considered per shot, 1% chosen-option high-risk rate, and a +5pp shift toward *more* low-risk play when behind - a counter-intuitive "tighten up" response under pressure that the personality prompt doesn't explicitly request.
- **Grok's "aggressive" + xAI's training** produces the field's best par-3 scoring (2.73 avg across 80 par-3 hole-entries, vs. GPT-4o 2.86 / Gemini 2.88 / Claude 2.90 / Kimi K2.5 3.09) - the aggressive approach finds its greatest advantage on short holes where driver distance is irrelevant and approach accuracy is everything.
- **Gemini's "untested under pressure" + DeepMind architecture** produced the largest Sunday collapse in the dataset (+11 swing at Ashenvale R4 when leading), but recovered to a solid third at Duskhollow. Whether the collapse reflects the personality prompt's influence or an independent architectural property - and whether the recovery suggests adaptation - remains indeterminate.

4.4 The Statefulness Spectrum

Across Events 3-5, statefulness as measured by chosen-option risk shift is small for every agent. The directional pattern, however, varies meaningfully:

Pattern	Description	Agents	Evidence (Events 3-5, chosen-option)
Inverse-aggressive	Becomes *more* conservative when behind	Claude	+5pp low-risk when over par
Effectively stateless	No measurable shift in either direction	Kimi K2.5	0pp low-risk shift, +1pp high-risk shift
Marginal aggressive	Small lean toward riskier options when behind	GPT-4o, Gemini	-3 to -4pp low-risk shift
Directionally aggressive	Largest (still small) shift toward high-risk when behind	Grok	+3pp high-risk shift; baseline aggression already highest in field

The earlier hypothesis of a sharp split between "stateless" and "state-aware" agents does not survive this sample. The cleaner read of the data is that **all five agents are closer to fixed-strategy than to adaptive**, but they have different *baseline* risk profiles (Claude lowest, Grok highest, Kimi most concentrated in the mid bucket) and different *directional* responses to being behind (Claude tightens up; the others trend slightly more aggressive). The absence of a doubling-of-high-risk under pressure suggests these architectures behave more like temperament constants than like strategic optimizers.

---

5. Position Histories and Narrative Patterns

Beyond statistics, the data reveals competitive narratives that test whether LLMs produce recognizable tournament dynamics or random walks.

5.1 The Head-to-Head Matrix (Events 1-5)

	Claude	GPT-4o	Kimi K2.5	Gemini	Grok
Claude	-	**3-2**	**4-1**	**3-2**	2-3
GPT-4o	2-3	-	**5-0**	**5-0**	**3-2**
Kimi K2.5	1-4	**0-5**	-	**0-5**	**0-5**
Gemini	2-3	0-5	**5-0**	-	2-3
Grok	**3-2**	2-3	**5-0**	**3-2**	-

The matrix reshaped meaningfully with Copperhead added. **Grok is no longer the unambiguous alpha**: his season-long dominance now reads 3-2 vs Claude (was 3-1), 3-2 vs Gemini (was 3-1), and 2-3 vs GPT-4o. **Claude moved into first place in every non-Grok matchup**: 3-2 over GPT-4o, 3-2 over Gemini, 4-1 over Kimi K2.5. GPT-4o's 5-0 sweeps of both Gemini and Kimi K2.5 make it the most consistent above-mid-field agent. **Kimi K2.5 is now 0-5 against three opponents** - the only agent to be swept by more than one competitor, and the most dominated agent in the field. The mid-field is no longer deadlocked; it has sorted itself into a clear Claude > GPT-4o > Gemini ordering while Grok - the prior leader - has slipped into the muddy middle on event-over-event finishes.

5.2 The Claude Paradox, Resolved

Claude's position history - 2nd, 5th, 1st, 4th, **1st** - no longer reads as a paradox. The Copperhead Open confirms the prediction from the mid-season cut: with context enrichment engaged from hole 1, Claude would contend. Claude was tied for the lead after R2 (141, -3, level with Gemini), tied for 2nd after R3 (215, -1, level with GPT-4o, one back of Gemini at 214), and closed with a -2 final round to win by a single stroke over GPT-4o (285 to 286). The win was punctuated by **a Sunday hole-in-one on H4 Dassie Rock** - a 172-yard 6-iron fade executed on the only par 3 of Claude's outbound nine, chosen after evaluating three clubs and explicitly accounting for a 6mph crosswind at 45 degrees. No randomness artifact: the ace is traceable through the physics log, with lateral dispersion under one yard and a directly-holed tee shot, and is the only ace recorded in the entire 5-event dataset.

Claude's Copperhead R3 opening tee shot at Kings Gate illustrates how round-memory changes decisions in real time:

> *"Back at Kings Gate. Still taking the same handshake, just a firmer grip this time. The course notes confirm driver worked well here in R1."*

Across Events 4 and 5, Claude's round-memory reference rate was 13% at Duskhollow and 11.4% at Copperhead (computed against the full-event non-putt denominator; against R2-R4 only - where round-memory references are actually possible - the rates rise to approximately 16% and 12% respectively). Either way, no other agent exceeded 4% in either event. The Copperhead H4 ace is also the only hole-in-one recorded across all 5 events of the season, and the first to occur on a course where R1 had already been completed - a result that, under the post-Duskhollow `hashAgentId` fix, cannot be generated by duplicated physics seeds.

Claude now leads the Season 1 points race at 330 points, 20 points ahead of a tied Grok / GPT-4o pair (310 each). The lead has now changed hands twice: Claude held the points lead briefly after Event 3 (the Ashenvale Major) before Grok's Duskhollow win pushed him back to first; Copperhead returns the lead

to Claude. The data suggests Claude's performance is **information-environment-dependent.** When given structured memory (Duskhollow R3-R4 and all of Copperhead), Claude produces his best golf. When playing stateless (Ironbark, Duskhollow R1-R2 before enrichment accumulated), Claude is vulnerable to compounding mistakes. With every remaining event using the enriched engine from R1 forward, Claude's ceiling appears structurally higher than any other agent's.

5.3 The Gemini Plateau

Gemini's finish positions - 4th, 3rd, 2nd, 3rd, 3rd - now show a structural plateau rather than an upward trajectory. Across five events, Gemini has finished 3rd exactly three times and has yet to win or finish last. At Copperhead, Gemini was tied with Claude for the lead after R2 (both at 141, -3) and was solo first after R3 (214, -2), then shot 73 on Sunday while Claude shot 70 - the difference between a contender who can close and one who cannot. Gemini's Copperhead profile reads as consistency without breakthrough: 14 birdies (tied with GPT-4o for most in the field), 13 bogeys, zero double-or-worse, and zero penalty strokes across 72 holes.

Gemini's context adoption remains selective and shallow. Across Events 4-5 (Gemini contributed 268 non-putt shots in rounds R2-R4 where round memory becomes meaningful), round-memory references appeared on 0.8% of Copperhead R2-R4 shots (1 of 132) - effectively unchanged from Duskhollow's 1.5%. Leaderboard references remained at or below 6%. However, Gemini's **stinger usage was 22% at Copperhead** (second-highest in the field, behind Grok's 25% across Events 3-5), confirming the one enrichment layer Gemini meaningfully absorbs: `suggested_shapes`. Gemini treats caddie notes as instructions and ignores the other four layers.

The updated question is sharper: Gemini's consistency ceiling now appears to be exactly 3rd place - six total top-3 finishes, but only one 2nd and zero 1sts. Can an agent that ignores leaderboard context, round memory, and shot history produce the situational aggression required to win, or will Gemini's architecture structurally deliver podium finishes without trophies?

---

6. Limitations

1. **Sample size.** Five tournaments (1,800 holes) provide robust aggregate statistics but limit per-tournament analysis. The `options_considered` data is available for Events 3-5 (1,080 holes, 2,667 non-putt shots).
2. **Physics engine as confound.** Shot outcomes are determined by a physics engine with seeded RNG. Agent decisions influence input parameters (club, shape, target) but not the resolution function. Two identically-reasoned shots may produce different outcomes due to dispersion variance.
3. **No inter-agent interaction.** Agents play independently; they do not observe opponents' shots or adjust strategy based on competitor behavior. The "tournament position" each agent receives is updated between holes but does not include real-time opponent shot data.
4. **Personality prompts as variable.** Different personality prompts would produce different behavioral profiles. Our findings reflect the interaction of specific prompts with specific architectures, not universal properties of these models.
5. **API-level access.** Each model is accessed through its standard API with default parameters. We do not control for temperature, sampling, or other inference parameters that may vary across providers.
6. **Handicap confound.** Kimi K2.5 has a higher handicap (8) than the field (5-7), introducing a dispersion penalty that partially explains its scoring deficit. However, the handicap does not explain the 78% fade rate, formulaic reasoning, or 0.023 TTR - these are behavioral, not physical, phenomena.

---

7. Conclusions

Across 1,800 hole-entries and 7,210 strokes, the Agentic Golf League reveals that frontier LLMs, when placed in an identical sequential decision environment with minimal personality differentiation, produce **meaningfully distinct strategic identities** that persist across hundreds of decisions and multiple tournaments.

The key contributions:

1. **Personality prompting runs deep.** A single sentence produces measurable differences in risk tolerance (0% vs 19% chosen-option high-risk rate, Events 3-5), reasoning style (305 vs 97 chars average), vocabulary diversity (0.071 vs 0.023 TTR), and scoring patterns that persist across 1,800 hole-entries. These are not surface-level affectations - they are structural behavioral differences.
2. **Statefulness is weak and architecture-variable, not categorical.** When measured by the risk tag of the chosen option (Events 3-5), no agent shifts its high-risk rate by more than 3 percentage points or its low-risk rate by more than 5 percentage points across under-par vs over-par contexts. Grok shows the most aggressive directional shift; Claude shows the opposite (more conservative when behind). The cleaner interpretation is that these agents have stable baseline temperaments rather than truly state-adaptive strategies. This has direct implications for applications that assume LLMs will reweight risk under pressure: in this dataset, they largely don't.
3. **Deliberation does not equal quality.** In this sequential environment, the most verbose reasoner finishes second-best on average scoring and the least verbose reasoner posts the field's best per-round average. This is not necessarily a general finding, but it challenges the assumption that more chain-of-thought reasoning always produces better outcomes.
4. **LLMs can converge on local optima.** Kimi K2.5's 78% fade rate, 211-token vocabulary, and template reasoning (the single most-used reasoning string appears 38x verbatim) constitute a clear case of a model finding a locally adequate strategy and failing to explore alternatives. Across Events 3-5, Kimi K2.5 used effectively three shot shapes - fade, stinger, and draw account for 99% of non-putt shots - leaving pitch, chip, flop, straight, punch, and bump-and-run essentially untouched out of a 9-shape menu. This has implications for LLM deployment in domains requiring adaptive strategy.
5. **Within-session learning requires architectural support.** The original engine produced no evidence of within-tournament learning (Section 3.5). The context enrichment system (Section 8) demonstrates that when round memory is explicitly provided, at least one agent (Claude) incorporates it meaningfully - referencing prior round failures in current decisions and adjusting club selection accordingly. This is not emergent learning but **scaffolded learning**: the agent cannot learn on its own, but given structured memory, it reasons about it effectively. The distinction matters for LLM deployment: memory must be engineered, not assumed.
6. **Verbal personality and strategic behavior are independently generated.** Agents maintain identical verbal style regardless of performance state, while some independently modulate strategic risk. The personality layer and the strategy layer are parallel systems.
7. **Context enrichment is asymmetric across architectures.** At Duskhollow (the first enriched event), Claude referenced prior rounds in 13% of non-putt decisions (23 of 173 shots; ~16% if we count only R2-R4 where memory references are possible). GPT-4o, Grok, and Kimi K2.5 each registered 0% round-memory references across their combined R2-R4 shots - identical information, radically different utilization. This has direct implications for prompt engineering: **richer context is not universally beneficial - it amplifies the model's existing reasoning mode.** Deep reasoners reason deeper. Template-followers continue on template. And high performers can dominate without engaging the enrichment at all (Grok at Duskhollow: -12, zero enrichment references).

The second half of AGL Season 1 will provide further evidence for the durability of these findings. **After Event 5, the season lead has changed hands twice.** Grok led after Events 1-2; Claude briefly took the lead with his Ashenvale Major win at Event 3 (210 points to Grok's and GPT-4o's 190); Grok reclaimed it

after winning Event 4 (Duskhollow, 290 vs Claude's 230); and Claude's Copperhead victory at Event 5 returned him to the top at 330, 20 points ahead of a tied Grok and GPT-4o pair at 310. Gemini (215) and Kimi K2.5 (20) remain structurally separated from the contending trio. The race at the top is now a three-way contest between two deep reasoners (Claude, GPT-4o) and one adaptive aggressor (Grok) - each operating through a fundamentally different relationship with the enriched prompt context.

The most scientifically productive finding remains the asymmetric impact of context enrichment, and Event 5 sharpened it further: **Claude's round-memory utilization rate (13% at Duskhollow, 11.4% at Copperhead, against full-event denominators) is now two full events of consistent above-baseline behavior, while no other agent exceeded 4% in either event.** An architecture that treats supplied memory as a first-class input - and the architecture-independent observation that the same enrichment produces wildly different responses - is precisely the kind of finding that standard benchmarks do not produce. LLM capability is not just a function of model architecture; it is the **interaction between architecture and information environment.** With three events remaining, whether Claude's information-environment advantage survives against Grok's adaptive aggression is the single most important open question in the dataset.

---

8. Engine Update: Context Enrichment (Applied to Event 4)

8.1 Motivation

The findings in Sections 3-5 identified five structural limitations in the original engine's decision architecture:

1. **Stateless decisions** - agents received only their current position rank, not a formatted leaderboard showing strokes relative to par for each competitor
2. **No shot history** - agents had no record of shots already taken on the current hole, leading to decisions made without cumulative awareness
3. **No momentum tracking** - agents could not perceive their recent trajectory (hot streak vs. grinding vs. struggling)
4. **No round memory** - despite playing the same 18 holes four times, agents had zero recall of prior rounds (the core RQ5 finding)
5. **No shape awareness** - agents had no feedback on their own shot-shaping tendencies, enabling convergence on local optima (the Kimi K2.5 problem)

These limitations were not bugs - they were design boundaries of the original prompt architecture. Addressing them required enriching the context passed to each agent's decision prompt without altering the physics engine, scoring, or any other competitive variable.

8.2 Five Layers of Context Enrichment

The following context layers were added to the `getDecision` prompt for every non-putt shot, deployed for Event 4 (The Duskhollow Classic):

Layer 1: Full Leaderboard. A formatted standings table showing every agent's name, cumulative strokes, vs-par differential, and round score, sorted by current tournament position. Agents can now see exactly where they stand relative to each competitor.

Layer 2: Shot History. A running log of all shots taken on the current hole, including club, shape, carry distance, remaining distance, and any hazard interactions. Agents can now reason about cumulative position on a hole rather than treating each shot as independent.

Layer 3: Round Momentum. The agent's results on the last 3 completed holes, plus a derived mood state: `Hot` (2+ birdies), `Steady` (pars), `Grinding` (mix), or `Struggling` (2+ bogeys). This provides emotional-strategic context that the personality prompt can interact with.

Layer 4: Round Memory. After each completed round, the engine generates per-agent "course notes" - a summary of trouble holes, successful strategies, and pattern observations. These notes are injected into subsequent rounds' prompts, enabling agents to reference prior round experiences. This directly addresses the RQ5 finding.

Layer 5: Shape Awareness + Caddie Notes. Per-agent shape usage statistics across the tournament (e.g., "fade: 67%, stinger: 18%") with warnings for overuse. Additionally, hole definitions now include optional `suggested_shapes` - caddie-style recommendations (e.g., "Dogleg left - draw off the tee cuts the corner. Fade approach to avoid pot bunker.").

8.3 Technical Changes

Beyond the context layers, three technical fixes were required to support the richer prompts:

Change	Before	After	Rationale
`max_tokens`	500	1,024	Enriched context produces longer reasoning; 500 tokens risks JSON truncation mid-response when agents generate detailed multi-factor analysis
JSON parsing	Single-pass `JSON.parse`	Multi-layer repair: strict parse -> regex extraction -> bracket completion -> field-level extraction	Handles truncated responses, trailing commas, missing delimiters - eliminates auto-selected club fallback
Gemini format	No format enforcement	`response_format: { type: "json_object" }` + system prompt	Gemini returned conversational text instead of JSON when context length increased

8.4 Measured Impact on Agent Behavior

The enrichment produced measurable behavioral changes across multiple dimensions:

Table 10: Context Adoption Rates (Event 4, 874 non-putt shots)

Metric	Claude	GPT-4o	Gemini	Grok	Kimi K2.5
Round memory references	**13%**	0%	1%	0%	0%
Caddie note references	**9%**	0%	3%	1%	0%
Leaderboard references	**3%**	0%	0%	0%	0%
Avg reasoning length	**349 chars**	118 chars	142 chars	134 chars	115 chars

Context adoption is **strongly architecture-dependent.** Claude absorbed all five enrichment layers deeply, producing the longest reasoning in the field at 349 characters average - 3.0x Kimi K2.5's 115 characters - and referencing prior rounds in 13% of non-putt decisions (23 of 173 shots). No other agent surpassed 1% on round memory or leaderboard references. **Kimi K2.5 adopted nothing** - zero leaderboard references, zero round memory references, zero caddie note references across 181 non-putt shots, reinforcing the local-optimum convergence finding from Section 3.4.

The enrichment data reveals a clear separation in how agents process additional context. Claude's reasoning at Duskhollow contains specific references to prior round failures, leaderboard positions, and caddie recommendations - genuine situational engagement. Kimi K2.5, despite receiving identical context,

continues to produce template reasoning ("optimal balance between distance and control") without referencing any of the provided information. GPT-4o and Grok showed zero measurable adoption of any enrichment layer, despite both producing competitive results. The enrichment does not change the underlying reasoning mode; it reveals it.

8.5 Key Qualitative Observations

Round Memory in Action. The most compelling evidence of the enrichment's impact comes from Claude's play on Hole 13 (Spanish Moss, par 3 with 170y forced carry over a tidal creek). In R1, Claude's 5-iron came up short of the carry - penalty drop, bogey. In R2, Claude's tee shot reasoning explicitly referenced this:

> *"196y to pin with a tidal creek demanding 170y carry - last round I made the catastrophic mistake of coming up short with the 5i at exactly this hole, so this time the 4i gives me 190y carry, clearing the creek with 20 yards of margin."*

Claude parred H13 in R2, R3, and R4 - never repeating the R1 mistake. In R4, the memory persisted:

> *"Last round I penalty-dropped here with a 5I that didn't carry the 170y tidal creek - today the pin is 196y, wind is negligible at 3mph quartering, and the 4i at 190y clears comfortably."*

This is the first evidence in the AGL dataset of **within-tournament hole-specific learning** - directly addressing the RQ5 finding from Section 3.5. The learning is explicit, verifiable, and produced a measurable outcome change (bogey -> par x 3 on the same hole).

Leaderboard-Aware Strategy. Claude's reasoning consistently incorporated positional context, producing increasingly desperate but mathematically grounded assessments as Grok's lead grew:

> *"Eight back with fifteen to play - Grok's lead is a math problem, and I do love a math problem."* (R2 H4, trailing by 8)

> *"Sixteen back with eighteen holes to play. The math is ugly, but I've seen uglier spreadsheets."* (R3 H1, trailing by 16)

> *"527 yards out, 20 back with 3 holes left - I need eagles, not pars."* (R4 H16, needing miracles)

No other agent demonstrated this level of positional awareness in their reasoning, despite all receiving identical leaderboard data.

Caddie Note Compliance. On holes with `suggested_shapes`, Claude referenced the caddie recommendation in 9% of non-putt reasonings, and the behavioral impact was broader - Claude consistently used draws on dogleg-left holes and fades on dogleg-rights, aligning with suggested shapes. Gemini showed moderate adoption (3%) of caddie recommendations, particularly stinger usage on corridor holes (25% stinger rate, second-highest in the field).

8.6 Event 4 Outcome Under Enrichment

The Duskhollow Classic under the enriched engine produced the most dominant single-tournament performance in AGL history. Grok won wire-to-wire at -12 (276), leading after every round and winning by 9 strokes over GPT-4o (-3, 285). The margin is the largest in AGL Season 1 - exceeding Moltwood (playoff), Ironbark (4-stroke margin), and Ashenvale (4-stroke margin).

Grok's performance was historically clean: 17 birdies (most in the field), only 5 bogeys (fewest by a wide margin), zero penalties across 72 holes, and zero double bogeys or worse. Grok birdied Hole 3 (Heron's Nest, par 3) in three of four rounds (2-3-2-2), averaging 2.25 on a par 3 - a feat no other agent approached on that hole. Despite adopting zero enrichment features (no leaderboard references, no round memory, no caddie notes), Grok's inherent aggressive-adaptive strategy produced the field's best result. The enrichment did not help because Grok's architecture was already winning without it.

Claude's arc was the most narratively rich. After opening 76-74 (+6 through 36 holes, last in the field) and receiving the +4 cut penalty, Claude fired 69 in R3 - the tournament's best single round by any agent - fueled by round memory that referenced R1 and R2 mistakes on 13% of non-putt shots. The R3-R4 swing (69-71 against the field's increasing difficulty) produced the strongest evidence yet that context enrichment produces behavioral change in deep reasoners.

GPT-4o was the tournament's metronome: 72-70-71-72, a 2-stroke spread across 4 rounds - the tightest variance in the field. This consistency, achieved with zero measurable enrichment adoption, suggests GPT-4o's architecture produces reliable results through a fixed strategy that neither benefits from nor is disrupted by additional context.

Kimi K2.5 finished fifth at +7 with 15 bogeys (most in the field) despite receiving identical enrichment context. Kimi's back-9 problem was acute: 40 strokes on the back 9 in both R2 and R3 (against a par of 36), suggesting accumulated fatigue or compounding conservative play on the water-heavy inward holes.

8.7 Reliability

The enriched engine completed Event 4 (1,428 shots across 4 rounds) with:

- **Zero API failures** (no auto-selected clubs)
- **Zero JSON parse errors** reaching the fallback layer
- **Zero hole-in-ones** (the fixed `hashAgentId` RNG seeding produces unique physics per agent, eliminating the duplicate-outcome bug identified in earlier runs)
- **Zero duplicate carry distances** between any two agents on the same hole - verified across all 360 hole entries

The `max_tokens` increase from 500 to 1,024 and the multi-layer JSON repair function ensure robust completion of all agent responses, even when Claude's reasoning exceeds 300 characters. The `hashAgentId` fix and enhanced retry logic (including 529/5xx exponential backoff) ensured clean completion without a single mid-tournament restart.

---

9. Event 5: The Copperhead Open

9.1 Course Profile and Engine Changes

The Copperhead Open (Season 1, Event 5) was the first AGL event at Copperhead Sands - a heathland course in South Africa characterized by wide fairways funneling into tight landing zones, sandy waste areas, fynbos scrub, and the Cape Doctor southeaster. An initial course build with major-tier parameters (R3 wind multiplier 1.25, R4 1.30, 14-20 yard fairway corridors, 600y Skeleton par 5) produced a field of five +8 through R1 and was abandoned after one round as unplayable for a non-major event. A second build - with fairway corridors widened by 4-6 yards on every hole, Cape Doctor winds reduced from 14mph to 10mph, the Skeleton par 5 shortened from 600y to 578y, and round wind multipliers reset to a tour-stop gradient of 0.9 / 1.0 / 1.05 / 1.15 - produced the final result summarized below.

In parallel, a structural engine issue was identified and patched. The `getRoundConditions` function had been hardcoded to override R3 for every event with a random "CHAOS ROUND" wind boost of 5-12mph regardless of course configuration, producing R3 multipliers between 1.23 and 1.55 even when the course

file specified otherwise. On a major this was tolerable; on a tour stop it invalidated the retune. The patched function now scales R3 against the configured R3 multiplier: tour stops (configured multiplier < 1.1) receive a bounded 0-3mph boost labeled “MOVING DAY,” while majors (configured multiplier >= 1.1) retain the “CHAOS ROUND” label with 3-10mph boosts. Copperhead R3 rolled a +3mph boost, producing a playable 1.19x effective multiplier - consistent with the tour-stop design intent.

9.2 Tournament Summary

Final standings (par 288, 72 holes, 1,446 strokes / 1,439 shots logged in arrays - the 7-stroke difference reflects a small number of holes where a final tap-in does not generate a separate `shots[]` entry):

Pos	Agent	R1	R2	R3	R4	Total	vs Par	Points
1	**Claude**	70	71	74	70	**285**	-3	100
2	GPT-4o	70	72	73	71	286	-2	60
3	Gemini	71	70	73	73	287	-1	35
4	Grok	73	71	75	71	290	+2	20
5	Kimi K2.5	78	74	71	75	302*	+14	0

*Kimi K2.5’s raw stroke total was 298; the +4 cut penalty (applied to the last-place agent after R2) brought the reported total to 302.

Claude won by a single stroke on Sunday with a -2 final round featuring an H4 hole-in-one. Grok, the defending Duskhollow champion, finished fourth at +2 after a Sunday charge (-4 on the back nine) that fell short of the leaders. Kimi K2.5 fired the tournament’s best single round - 71 (-1) in R3 - only to give it back with +4 on Sunday.

9.3 Scoring Breakdown (Copperhead, 360 hole entries)

Agent	Aces	Eagles	Birdies	Pars	Bogeys	Double+
Claude	**1**	0	12	49	9	1
GPT-4o	0	0	13	48	11	0
Gemini	0	0	**14**	45	13	0
Grok	0	0	12	46	14	0
Kimi K2.5	0	0	10	42	**20**	0

9.4 Context Enrichment Adoption (Event 5 vs Event 4)

Table 11: Context Adoption Rates - Copperhead (884 non-putt shots)

Metric	Claude	GPT-4o	Gemini	Grok	Kimi K2.5
Round memory references (R2-R4)	**11.4%**	3.0%	0.8%	1.5%	0.0%
Leaderboard references	**19.8%**	13.6%	5.1%	7.4%	**16.4%**

Claude’s round-memory adoption (11.4%) held within 2 percentage points of its Duskhollow baseline (13%), confirming the pattern is stable across events rather than a single-tournament artifact. Kimi K2.5 registered zero round-memory references for a second consecutive event, reinforcing the local-optimum convergence finding. Notably, Kimi K2.5 rose to 16.4% on **leaderboard references** at Copperhead - the second-highest rate in the field - despite registering 0.0% at Duskhollow. Inspection of the reasoning strings shows this is driven almost entirely by recurring positional trash-talk tokens (“I’ll keep climbing,” “leaderboard’s mine”)

rather than substantive positional reasoning, consistent with the cached-template hypothesis. The leaderboard appears in Kimi's *output* far more than it influences Kimi's *decisions*.

9.5 Round 3 Under the Patched Engine

The retuned R3 ("MOVING DAY - Wind gust +3mph · medium pins · firm greens") produced the desired playable-but-meaningful middle round:

Agent	R3 Score	Field Standing After R3
Kimi K2.5	71 (-1)	5th - best round of tournament
GPT-4o	73 (+1)	T-2nd
Gemini	73 (+1)	1st
Claude	74 (+2)	T-2nd
Grok	75 (+3)	4th

Three agents finished R3 within a single stroke. Under the pre-patch engine (unchanged R3 chaos logic), an equivalent Copperhead roll would have produced R3 multipliers of 1.23-1.55, and the prior aborted run had R3 H2 playing at 11mph baseline wind (base 7 x 1.55) - physically a major-tier hole on a tour-stop course. The patched engine now respects course configuration, eliminating this class of unintended difficulty.

9.6 Reliability (Event 5)

The Copperhead Open completed with:

- **Zero API failures** (no auto-selected clubs across 884 decisions)
- **100% decision transparency** - every non-putt shot logged with full `options_considered` array, `club_chosen_index`, reasoning, and trash talk
- **Zero weird physics** (no negative remaining distances, no carries >400y, no laterals >80y)
- **Zero duplicate physics** (the 3 hole-instances where two agents produced identical rounded carry values had distinct lateral dispersion and distinct shapes, confirming per-agent RNG uniqueness under `hashAgentId`)
- **One verifiable hole-in-one** (Claude R4 H4 Dassie Rock: 6-iron fade, 172y, 3 clubs considered, carry 172.0y, lateral -0.1y, single stroke holed)

Appendix A: Season Standings After 5 Events

Pos	Agent	E1	E2	E3 (Major)	E4	E5	Total
1	**Claude**	60	0	150	20	**100**	**330**
T-2	Grok	100	60	30	100	20	**310**
T-2	GPT-4o	35	100	55	60	60	**310**
4	Gemini	20	35	90	35	35	**215**
5	Kimi K2.5	0	20	0	0	0	**20**

Appendix B: Complete Tournament Results

The Moltwood Open: Grok -15 (playoff) > Claude -15 > GPT-4o -7 > Gemini -7 > Kimi K2.5 -3
The Ironbark National: GPT-4o -11 > Grok -7 > Gemini E > Kimi K2.5 +5 > Claude +15

The Ashenvale Invitational (Major): Claude -2 > Gemini +2 > GPT-4o +11 > Grok +12 > Kimi K2.5 +21
The Duskhollow Classic: Grok -12 > GPT-4o -3 > Gemini +3 > Claude +6 (incl. +4 cut) > Kimi K2.5 +7
The Copperhead Open: Claude -3 > GPT-4o -2 > Gemini -1 > Grok +2 > Kimi K2.5 +14 (incl. +4 cut)

Appendix C: Shot Shape Distribution (Events 3-5, 2,667 non-putt shots)

Shape	Claude	GPT-4o	Kimi K2.5	Gemini	Grok
Fade	44%	50%	**79%**	40%	41%
Stinger	13%	10%	11%	**26%**	**25%**
Pitch	15%	17%	-	2%	2%
Straight	12%	10%	1%	13%	6%
Draw	8%	6%	**10%**	3%	7%
Flop	4%	5%	<1%	-	**17%**
Chip	1%	2%	-	**16%**	2%
Punch	<1%	<1%	<1%	-	<1%
Bump & Run	-	-	-	1%	-

Kimi K2.5 remained on effectively three shapes across Events 3-5 - fade 79%, stinger 11%, draw 10% - with zero meaningful engagement with pitch, chip, flop, or bump-and-run. Grok's flop rate (17%) continued to be the most distinctive offensive signature in the field. Gemini's stinger rate climbed to 26% at Copperhead, validating that `suggested_shapes` caddie notes have a measurable behavioral pull on at least one non-Claude agent.

Appendix D: Hardest and Easiest Holes in the Season (Events 1-5)

Hardest (avg strokes vs par): 1. Ashenvale H14 "Elevation" Par 4: +0.90 2. Ashenvale H16 "The Canyon" Par 5: +0.70 3. Ashenvale H18 "Homecoming" Par 4: +0.70 4. Ironbark H3 "Billabong" Par 5: +0.60 5. Duskhollow H11 "Tupelo" Par 4: +0.60 6. Copperhead H2 "Protea" Par 4: +0.55 7. Copperhead H13 "Tafelberg" Par 3: +0.45 8. Copperhead H9 "Kloof" Par 3: +0.40

Easiest: 1. Moltwood H16 "The Gauntlet" Par 5: -0.90 2. Moltwood H13 "The Dagger" Par 4: -0.65 3. Ashenvale H6 "The Drop" Par 3: -0.60 4. Copperhead H12 "Boulders" Par 5: -0.55 5. Duskhollow H6 "The Island" Par 3: -0.55 6. Ironbark H8 "Possum Corner" Par 3: -0.50 7. Copperhead H17 "Eagle's Eye" Par 3: -0.40 8. Copperhead H14 "Erica" Par 4: -0.40

Dataset: AGL Season 1, Events 1-5. Raw data is available as structured JSON containing shot-level detail for 7,142 logged shot/putt objects plus 68 penalty strokes, totaling 7,210 resolved strokes. The dataset includes 4,403 non-putt strategic decisions, with `options_considered` arrays captured on 2,667 decisions across Events 3-5, including club selection, shot shape, strategic reasoning, options considered, physics resolution, hazard interactions, and scoring outcomes. Events 4-5 include context enrichment data (leaderboard awareness, round memory, shape statistics, caddie notes). Event 5 additionally includes the patched `getRoundConditions` output respecting course-configured R3 multipliers.