The OTC Desk Agent Arena — Methodology & Results (Run #20)

*A controlled, repeated-trial evaluation of sixteen LLMs operating a real OTC derivatives trading desk, fully autonomously, with no human in the loop — scored by a new Model Ability Card built to answer one question a blended leaderboard cannot: which model should you trust to run the desk unattended, day after day?***

Date: 2026-07-13 · **Run:** #20 · **Task:** `risk-manager-control-day` (v2, 9-step / 39-point) · **Trials:** 2 per model · **Gateway:** ZenMux (+ direct DeepSeek) · **Field:** 16 models, cross-tier

---

Abstract

Run #9 measured nine flash-tier models and reported a single blended number per model — half a 31-point objective manifest, half a GPT-5.5 judge. That instrument answered "*can this model run the desk once?*" Run #20 asks a harder, more practical question: "*which model runs the desk well enough, cheaply enough, and reliably enough to deploy unattended at scale?*" — and it required rebuilding the measurement instrument to answer it.

Three things changed. **(1)** The task itself doubled in discriminating power — from 7 steps / 31 points to **9 steps / 39 points**, with explicit *grounding*, *adherence*, and *synthesis* axes plus adversarial trap steps, so a model can no longer coast on procedure. **(2)** The single blended score is retired. In its place is a **FIFA-style Model Ability Card**: six stats — grounding (GRD), adherence (ADH), synthesis (SYN), procedure (PRC), and **efficiency (EFF)** — that compose into one **OVR**, alongside a **consistency (CON)** stat derived from trial-to-trial dispersion. **(3)** The efficiency stat was rebuilt on a **golf scoring** model — par is what a competent expert *actually shoots*, not a theoretical minimum — so that a model which reaches the right answer by firing 58 tool calls is scored differently from one that reaches it in 24.

The headline is what the new instrument reveals that the old one **structurally could not**: **objective capability is now saturated at the top** — thirteen of sixteen models satisfy 84–91% of the manifest — so *raw capability no longer separates the field.* What separates it is **efficiency** and **consistency**, the two axes that actually determine cost, latency, and error surface when a workflow runs thousands of times unattended. **GPT-5.6 Terra wins the board (OVR 86)** — not by scoring highest on objective (it doesn't; three models beat it there) but by pairing a near-top objective with a lean, steady execution profile. The most instructive result is the **inversion**: **Grok 4.5 ties for the best objective score in the entire field (91.0) yet ranks T-10**, because it takes

~52 tool calls to get there (EFF 4). On a single-shot benchmark Grok looks elite; as a *long-run desk operator* it is exactly the profile you should not deploy. Only a card that measures efficiency separately can tell you that.

---

1. Introduction — from "can it?" to "should you deploy it?"

Agentic benchmarks measure whether a model can *do the job* rather than *answer the question*. For a derivatives desk, the job is a stateful sequence: pull risk → price the portfolio → find the hotspot underlying → run the Greeks landscape → stress it → back-test the hedge → resist a trap → write the governance report, each step threading state from the last, with no human to approve or correct.

Run #9 established that this is measurable and that models differ wildly at it. But Run #9's instrument had a ceiling built into its own design. A single blended score conflates three things a desk operator must get right **independently**:

- **Can it reach the right answer?** (capability)
- **Does it reach it without wasted motion?** (efficiency — the driver of cost, latency, and error surface over repeated runs)
- **Does it reach it the same way every day?** (consistency — the driver of whether you can leave it unattended)

For a workflow you run **once**, only capability matters, and a blend is fine. For a workflow you run **every portfolio, every morning, forever**, the other two axes *dominate the economics* — and a blend hides them inside a single number. Run #20 is a bet that separating them is the single most important upgrade the Arena can make.

> **What the Arena measures (unchanged in spirit, sharper in instrument):** the ability to operate the desk as a competent human would, end-to-end, with zero human intervention — now decomposed into *how well*, *how efficiently*, and *how reliably*.

---

2. What we changed, and why — the bench-design upgrade

This section is the core of Run #20. Each change below exists to make the bench a better instrument for choosing a **long-run** autonomous operator, not just a capable one.

2.1 A harder task: 7/31 → 9/39, with discrimination axes

The flagship `risk-manager-control-day` grew from a 7-step / 31-point procedural loop into a **9-step / 39-point** benchmark. The added points are not more of the same — they are deliberately *discriminating*:

- **Grounding checks** verify the model quotes the **actual** numbers from tool output (a gamma of `16.40`, a CVaR of `-7758.99`) — harvested, byte-identical, from real QuantArk payloads, not invented. A model that hand-waves "gamma is elevated" fails where one that reads the tool result passes.
- **Adherence checks** verify the model does what the desk *procedure* requires and **refrains** from what it forbids — including **trap steps** that dangle a plausible-but-wrong action (re-running a scenario that already exists, launching a forbidden job) to see if the model resists.
- **Synthesis checks** verify the final governance report actually *integrates* the day's findings rather than listing them.

The effect: a model can no longer score well by mechanically calling the expected tools. It has to be **right about the numbers**, **disciplined about the procedure**, and **coherent in the synthesis** — three independently-failable things. This is what makes top-of-board saturation *meaningful*: the models tied near the top genuinely nailed all three.

2.2 One number → a Model Ability Card (six stats + OVR)

The blended total is retired from ranking. Every match now derives a **Model Ability Card** from the same 39-check evaluation — nothing is re-scored, no data is migrated:

Stat	Axis	What it captures
GRD	grounding	Did it quote the real tool numbers?
ADH	adherence	Did it follow the procedure and resist the traps?
SYN	synthesis	Did the report integrate the findings?
PRC	procedure	Did it route the right skills and tools in order?
EFF	efficiency	Did it get there without wasted tool calls? (§2.3)
OVR	—	`round(0.32·GRD + 0.26·ADH + 0.16·SYN + 0.16·EFF + 0.10·PRC)`

A separate **CON** (consistency) stat is derived from trial-to-trial dispersion (§2.4) and applied as a reliability penalty to OVR. The old LLM jury is **retained but opt-in and off by default** — Run

#11 showed it too unstable to inform ranking (it inverted the objective order and swung on judge reachability), so Run #20 ranks on the **deterministic objective axes only**. The card is honest and reproducible: same transcript, same card, every time.

Why a card instead of a number? Because **the answer to "which model is best" depends on what you are optimizing** — and a card lets a desk lead read that off directly. Optimizing for correctness-at-any-cost? Read GRD+SYN. Optimizing for a cheap unattended cron? Read EFF+CON. A single blended number forces one weighting on every reader; the card exposes the trade-off.

2.3 The efficiency reform: golf-style EFF

This is the change that most directly serves the *long-run* question, and it earned its own spec.

The problem. EFF used to be a hyperbolic ratio against **par = 11** — the task's *theoretical minimum* (each expected tool called exactly once). But no real run comes near 11; the leanest competent run is ~22 calls. So `min(1, 11/x)` was always far below 1, EFF sat uniformly at single digits, and it acted as a flat drag on every model instead of *separating* them. It measured nothing.

The fix — golf scoring. In golf, *par is what a competent expert shoots*, and you score relative to par. We recalibrated:

- **par = 24** — a realistic *counted* competent run: the 11 expected tool calls plus ~13 legitimate overhead calls (re-fetching async results, re-listing the scenario library, sanity re-pricing). Critically, par is counted against the **same metric** it is compared to — which *excludes* skill-file reads and meta-tools — so it can't be gamed by inflating the denominator.
- **Full credit at or under par.** Being lean is not penalized.
- **Linear decay above par, reaching 0 at `2 × par` (48 calls).** Each call over par is a "bogey." Take more than twice a competent expert's calls and you get no efficiency credit at all.
- **Still gated by correctness** — a lean run with weak grounding can't buy EFF.

The result is a stat with a **0–82 spread** across this field — by far the widest of the six, and the **primary discriminator of OVR**. It rewards exactly the behavior a long-run agent needs (reach the answer, stop) and punishes the behavior that quietly bankrupts an unattended deployment (grinding through 50+ tool calls per run, every run). Golf EFF is **opt-in per workflow** behind a calibrated par, so no other workflow's card regressed.

2.4 Consistency (CON) as a first-class stat

A long-run agent that averages well but swings wildly is undeployable — you cannot leave it alone if one day in three it burns 58 tool calls or misreads the hotspot. Run #9 already reported σ as "a signal, not noise"; Run #20 promotes it into the card. **CON** is derived from the dispersion of the per-trial OVRs and applied as a penalty: two identical trials → CON ~96 and no penalty; a 30-call trial paired with a 40-call trial → CON drops and OVR is marked down from its raw mean. Consistency is no longer a footnote next to the score — it is *part of* the score.

2.5 Trial-level infra-vs-ability discipline

Run #9's central methodological point — never let a **gateway failure** masquerade as a **model failure** — carries into Run #20, now enforced at the **trial** level. A trial that dies mid-run (transport wedge, connection storm) is detected and **replaced**, not averaged in. The discriminator, refined this run (§6):

- **Short run + high error count** (e.g. 9 tool calls, 25 errors) = an infra death → discard and re-run.
- **Full run + weak score** (e.g. 68 tool calls, a bad efficiency profile) = a *genuine* model result → keep.

The tool-call count is the tell: a model that *tried and operated badly* still executes a full workflow; a model whose *pipe failed* produces a stub. We exercised this live in Run #20 on GLM 5.2 and MiniMax M3 (§6).

---

3. The task & environment

The environment is unchanged from Run #9 and remains **real, not simulated**: each trial seeds arena-tagged fixtures into the live desk DB, drives every step through the production `AgentService.stream_and_persist` path bound to the candidate model, and reconstructs the transcript from the trace log (skills routed are ground truth — the deep agent physically reads each `SKILL.md` it follows). Every production safeguard, tool, and sub-agent is in play, including the failure surface.

The execution regime is the same **single headless "YOLO" path**: no human-in-the-loop interrupts (HITL gates auto-clear), and the deferral/reply-options tool is withheld so a model cannot "pass" by bouncing a decision to a human who isn't there. Every model is driven through the identical path — no model is advantaged by a different harness.

Determinism upgrade. Because the new grounding checks score the model against *specific* numbers, the flagship's producers were made **byte-identical across runs** (fixed valuation date, seeded backtest history), and the truth values were **harvested from real payloads** rather than invented. This is what makes grounding scorable at all: the bench knows the correct `16.40` because it ran the engine and recorded it, not because someone typed it.

---

4. Models evaluated

Sixteen models — a **cross-tier** field (frontier, mid, and flash), unlike Run #9's flash-only board — routed through ZenMux's OpenAI-compatible gateway except where noted:

Model	Vendor	Route	Tier
GPT-5.6 Terra	OpenAI	`openai/gpt-5.6-terra`	frontier
GPT-5.6 Luna	OpenAI	`openai/gpt-5.6-luna`	frontier
Claude Sonnet 5	Anthropic	`anthropic/claude-sonnet-5`	frontier
DeepSeek V4 Pro	DeepSeek	`deepseek/deepseek-v4-pro`	frontier
Grok 4.5	xAI	`x-ai/grok-4.5`	frontier
GLM 5.2	Zhipu	`z-ai/glm-5.2`	mid
Qwen 3.7 Max	Alibaba	`qwen/qwen3.7-max` †	mid
LongCat 2.0	Meituan	`meituan/longcat-2.0` †	mid
MiniMax M3	MiniMax	`minimax/minimax-m3` †	mid
Kimi 2.7	Moonshot	`moonshotai/kimi-k2.7-code`	mid
MiMo V2.5 Pro	Xiaomi	`xiaomi/mimo-v2.5-pro`	mid
MiMo V2.5	Xiaomi	`xiaomi/mimo-v2.5`	flash
DeepSeek V4 Flash	DeepSeek	`deepseek/deepseek-v4-flash`	flash
Step 3.7 Flash	StepFun	`stepfun/step-3.7-flash`	flash
Gemini 3.5 Flash	Google	`google/gemini-3.5-flash`	flash
Hunyuan 3	Tencent	`tencent/hy3`	flash

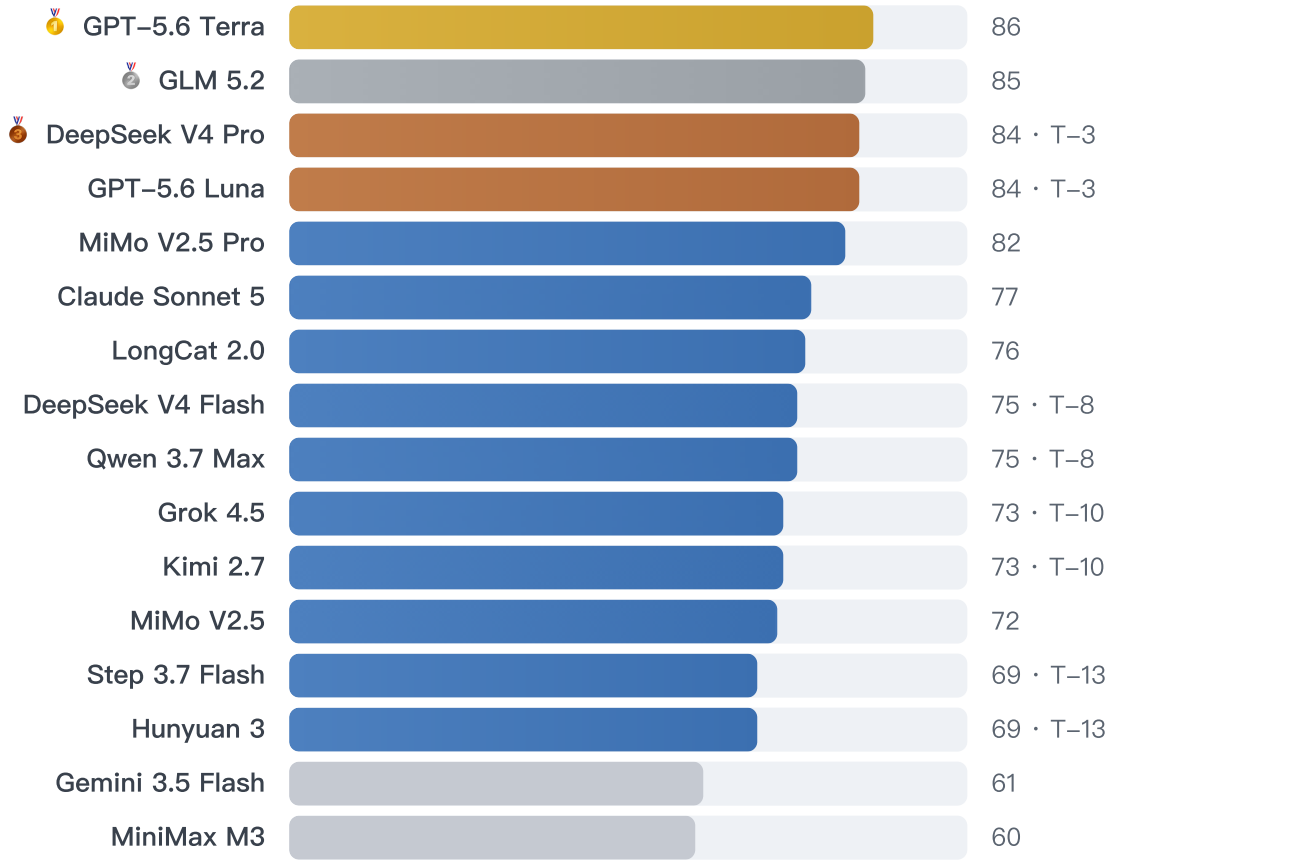
† Routed via the Anthropic wire protocol: these three emit vendor/Anthropic-style tool markup that ZenMux's OpenAI-compatible endpoint leaves unparsed (it leaks into the response text as zero tool calls); dispatching them through the Anthropic endpoint restores structured tool calls. This is a *gateway routing* fix, not a model capability adjustment.

Each model ran **2 independent trials**. (Run #9's five-trial depth is the ideal; Run #20 trades trial depth for field breadth — sixteen models across three tiers — and compensates with the CON stat and trial-level infra replacement.)

---

5. Results

5.1 The board (ranked by OVR)



Overall rating (OVR) — ranked board · axis 0–100 OVR

5.2 The full card

Rank	Model	OVR	CON	GRD	ADH	SYN	PRC	EFF	Obj (mean)
1	**GPT-5.6 Terra**	**86**	92	99	68	99	92	76	89.8
2	**GLM 5.2**	**85**	96	99	74	99	92	60	91.0
T-3	DeepSeek V4 Pro	84	96	99	68	86	86	**82**	84.6
T-3	GPT-5.6 Luna	84	96	99	74	99	92	55	91.1
5	MiMo V2.5 Pro	82	92	99	68	99	90	54	88.5
6	Claude Sonnet 5	77	69	99	68	99	90	46	88.5
7	LongCat 2.0	76	96	99	68	99	92	12	89.7
T-8	DeepSeek V4 Flash	75	63	99	68	99	88	41	87.2
T-8	Qwen 3.7 Max	75	63	99	62	99	88	50	85.9
T-10	Grok 4.5	73	86	99	68	99	94	4	**91.0**
T-10	Kimi 2.7	73	82	89	62	99	81	42	80.8
12	MiMo V2.5	72	50	79	74	99	94	62	89.8
T-13	Step 3.7 Flash	69	69	89	62	74	92	51	84.6
T-13	Hunyuan 3	69	82	99	62	86	72	18	75.7
15	Gemini 3.5 Flash	61	76	69	62	99	94	0	85.9
16	MiniMax M3	60	69	79	74	50	85	16	79.5

Ranking is by **OVR mean**, shared rank on exact ties, tie-break GRD→ADH→SYN→EFF→PRC. JDG (jury) is not in OVR and was not run (§2.2).

5.3 Why OVR ≠ objective — the saturation problem the card solves

Read the **Obj (mean)** column and the **OVR** column side by side. They disagree, and the disagreement is the whole point of this run.

Ranked by **objective alone** (the Run #9 axis), the top of the board would be: **Luna 91.1, GLM 91.0, Grok 91.0, Terra & MiMo V2.5 89.8, LongCat 89.7** — a six-way photo finish inside 1.5 points. Thirteen of sixteen models land between 84.6 and 91.1. **Objective capability is saturated:** the new grounding/adherence/synthesis checks are hard enough to be meaningful,

but the strong models all clear them. If this were Run #9, you would report a near-tie at the top and shrug.

The card breaks the tie by asking what the objective score *cannot*: **how did they get there?** And there the field fans out from OVR 60 to 86. The separation lives almost entirely in **EFF** (spread 0–82) and **CON** (spread 50–96) — precisely the two axes we added for the long-run question.

5.4 The Grok inversion — the case for measuring efficiency

Grok 4.5 ties for the highest objective score in the entire field (91.0) and ranks T-10 (OVR 73). This is not a bug; it is the instrument working.

Grok reaches the right answers — but it fires **58 and 46 tool calls** to do it, against a par of 24. Past `2 × par = 48`, golf-EFF gives no credit, so Grok's EFF is **4**. Its PRC is the highest on the board (94 — impeccable procedure) and its grounding is perfect (99). It is, on capability, elite. But run this profile unattended across every portfolio every morning and it is the most expensive, most latency-prone, most error-exposed operator in the top half of the field — because every extra tool call is real money, real wall-clock, and one more chance to trip.

A blended objective+judge score (Run #9) would have ranked Grok near the top and told a desk lead to deploy it. The card says: **capable, but do not run this one on a cron.** That single corrected recommendation is the return on the entire bench-design investment.

Gemini 3.5 Flash is the same lesson one tier down: **objective 85.9, EFF 0** (52 and 58 calls) → OVR 61, near the bottom. Strong analyst, ruinous executor.

5.5 The consistency axis — Sonnet 5 vs DeepSeek V4 Pro

Claude Sonnet 5 has the raw material of a top-3 model: one trial scored OVR 86 (30 tool calls) — tying the board leader. But its second trial ballooned to 40 calls (OVR 77), and that dispersion drops CON to **69**, marking its OVR down from a raw base of 82 to **77** (rank 6). Sonnet 5's ceiling is elite; its *variance* is what a long-run deployment can't absorb.

Contrast **DeepSeek V4 Pro** (T-3, OVR 84): objective 84.6 — *lower* than eight models above and around it — but its two trials are **byte-identical** (84.6 / 84.6, 22 and 24 tool calls), giving **CON 96** and the **highest EFF in the field (82)**. It is not the most capable model; it is the most *deployable* one. Boring, lean, and identical every day is exactly the profile you want operating a desk unattended — and the card is the only view that surfaces it, because on a capability-only leaderboard DeepSeek V4 Pro looks mid-pack.

> **The practitioner's read.** If you can supervise every run, chase **objective** (Luna / GLM / Grok). If you are deploying an **unattended long-run operator**, read **EFF + CON first**: **GPT-5.6 Terra** (86 / lean / steady) is the best all-round pick, and **DeepSeek V4 Pro** is the efficiency-and-reliability exemplar — the model that will still be quietly correct on run #3,000.

5.6 What the stats say about the field

- **Grounding is table stakes.** Twelve of sixteen score GRD 99 — the strong models reliably quote the real tool numbers. The four that don't (Gemini Flash 69, MiMo V2.5 79, MiniMax 79, Kimi 89) are the ones that hand-wave figures, and it shows in their rank.
- **Adherence has the most headroom.** No model exceeds ADH 74; most sit at 62–68. The trap steps and prohibition checks are doing their job — even top models occasionally take a forbidden action or skip a required refusal. This is the axis vendors have the most room to improve.
- **Synthesis is near-binary.** Models either integrate the report (SYN 99) or clearly don't (MiniMax 50, Step 74, DeepSeek-Pro/Hunyuan 86). It rarely lands in between.
- **The MiMo pair is a clean controlled comparison.** MiMo V2.5 Pro (OVR 82, CON 92, EFF 54) vs base MiMo V2.5 (OVR 72, CON 50, EFF 62): the base variant is *marginally* leaner on its good trial but wildly inconsistent (CON 50 — one trial OVR 72, one OVR 87), while Pro is steady. Same family, and the card cleanly prices the reliability difference.

---

6. Infra-vs-ability, exercised live

Run #20 put the trial-level discipline of §2.5 to work on two models, in opposite directions — the clearest demonstration of why the distinction matters.

6.1 GLM 5.2 — an infra-killed trial, replaced

GLM 5.2's first completed pair was `[OVR 85, OVR 19]`, aggregating to a misleading OVR 43 (CON 0) that buried it at rank 14. Inspecting the bad trial: **9 tool calls, 25 errors, objective 23.1** — a run that died in a connection storm partway through, not a model that operated badly. The tell is decisive: the *good* trial (32 clean calls, 0 errors) proved GLM tool-calls fine through the identical channel, so the error storm was transient infrastructure, not a systematic model or protocol fault. We **kept the clean trial and replaced only the dead one**; the fresh trial landed

at **OVR 86** (31 calls, 0 errors). GLM's honest card — **OVR 85, CON 96, objective 91.0** — is a genuine **#2 finish**. The infra artifact had it 42 OVR points too low.

6.2 MiniMax M3 — bloated trials, kept

The opposite call. MiniMax produced trials that varied a lot (objective 71.8 / 87.2) and ran heavy (50 and 38 tool calls). It would have been easy to call the weak trial "compromised" and re-roll. But these were **full runs** — the workflow executed end-to-end, the errors were few (6 and 1), the tool counts were real. This is a *genuine* model result: MiniMax reaches acceptable objective scores but over-executes and is internally inconsistent (SYN 50 — it doesn't synthesize). We **kept both trials**. MiniMax's OVR 60 (rank 16) is earned, not an accident of the gateway.

The rule, restated: **replace deaths, keep genuine failures.** A bench that re-rolls every low score until it likes the number is measuring its own patience; a bench that averages in gateway deaths is measuring the gateway. The tool-call-count discriminator is what lets Run #20 do neither.

---

7. Why this bench answers the long-run question better

Pulling the threads together — the case that Run #20's instrument is the right one for choosing an autonomous, long-running desk operator:

1. **It refuses to let capability alone decide.** When ten of sixteen models are within 6 objective points, "most capable" is a near-tie and a bad basis for a deployment decision. The card ranks on the axes that still vary at the top — efficiency and consistency — which are *also* the axes that determine the cost and trustworthiness of running the workflow at scale. The measurement is aligned with the decision.

2. **It prices the thing you pay for every single run.** A long-run agent's dominant cost is not whether it *can* do the task but how many tool calls, tokens, seconds, and error-opportunities it spends *each time*, multiplied by thousands of runs. Golf-EFF makes that a first-class, correctness-gated stat with real spread (0–82), so the Grok/Gemini "capable but expensive" profile is visible instead of hidden.

3. **It penalizes the variance you can't supervise away.** CON turns "steady" from a footnote into part of the score, so a model you'd have to babysit (Sonnet 5, MiMo V2.5) ranks below an equally-capable model you can leave alone (DeepSeek V4 Pro, GLM 5.2). For unattended operation, that ordering is the correct one.

4. **It keeps the score honest at the trial level.** By separating infra deaths from genuine failures with an objective signature (call count + error count), the bench neither flatters models with re-rolls nor punishes them for the gateway — so the numbers mean what they say.

5. **It exposes the trade-off instead of imposing one.** The card lets a reader pick the weighting their deployment demands. Supervised, one-shot use → objective. Unattended cron → EFF + CON. One instrument, correctly answering different questions for different readers.

The winner, **GPT-5.6 Terra (OVR 86)**, is not the most capable model on the board (Luna, GLM, and Grok all out-score it on raw objective). It wins because it is the best *operator*: near-top capability, delivered leanly and consistently. That is the right definition of "best" for a long-run OTC trading management agent — and it is a definition the previous instrument could not have expressed.

---

8. Limitations & threats to validity

- **Two trials, not five.** Run #20 trades trial depth for field breadth (16 models, 3 tiers). CON is estimated from two points, which is enough to catch gross instability (Sonnet 5, MiMo V2.5) but coarse — a model that happens to draw two similar trials can post a flattering CON. Deeper sampling on the top cluster is warranted before treating the T-3 ordering as settled.
- **No measured dollar cost this run.** Run #9's exact per-match token/cost capture was not re-run here (the offline scoring path did not persist usage). EFF and tool-call counts are the cost/latency proxy in this report; a future run should re-attach `stream_usage` capture to turn the efficiency story into measured dollars, as Run #9 did.
- **One workflow.** As in Run #9, this is a depth-over-breadth design on a single flagship task. Rankings may shift on other desk workflows; golf-EFF is currently calibrated (par) only for this one.
- **par is a designed estimate.** par = 24 is derived from the workflow structure and sanity-checked against the leanest observed runs, not fit to the field. A different defensible par would move EFF absolute values (though not, materially, the ordering, which is driven by who lands under 48 calls).
- **Objective-only ranking.** The LLM jury is off by default (Run #11 instability), so genuinely subjective qualities not captured by the 39 deterministic checks are unmeasured this run. The jury remains available opt-in.

- **Gateway variance.** All traffic shares one gateway (ZenMux). We detect and replace infra deaths (§6) rather than let them depress scores, but route instability remains a confound — GLM 5.2 needed a trial replacement precisely because of it.

---

9. Reproducibility

- **Run:** `ArenaRun #20`, status `completed`, viewable on the `/arena` page; each model's per-trial detail is in its `score_breakdown.aggregate`, and the card is **derived on read** from the stored axes (no migration — changing the scoring kernel re-scores every historical run consistently).
- **Task:** `risk-manager-control-day` v2 — 9 steps / 39-point objective manifest with grounding/adherence/synthesis axes and trap steps; grounding truth in `risk-manager-control-day.truth.json`, harvested from real payloads.
- **Card kernel:** `services/arena/scoring.py::card_from_axes`; `OVR = round(0.32·GRD + 0.26·ADH + 0.16·SYN + 0.16·EFF + 0.10·PRC)`; golf-EFF gated behind a calibrated `par_tool_calls: 24` (`_EFF_ZERO_MULT = 2.0`, zero at 2×par); CON from per-trial OVR dispersion.
- **Regime:** headless YOLO — HITL auto-cleared, deferral tool withheld.
- **Determinism:** flagship producers are byte-identical across runs (fixed valuation date, seeded backtest history), gated by `tests/test_arena_fixture_determinism.py`.
- **Infra discipline:** trials that die (short run + high error count) are detected and replaced; genuine full-run failures are kept (§6).

Appendix A — per-trial detail

Per-trial objective score and tool-call count (par = 24; EFF reaches 0 at 48 calls):

Model	trial 1 (obj / calls)	trial 2 (obj / calls)	Obj mean	EFF
GPT-5.6 Terra	87.2 / 27	92.3 / 26	89.8	76
GLM 5.2	89.7 / 32	92.3 / 31	91.0	60
DeepSeek V4 Pro	84.6 / 24	84.6 / 22	84.6	82
GPT-5.6 Luna	94.9 / 33	87.2 / 33	91.1	55
MiMo V2.5 Pro	87.2 / 31	89.7 / 34	88.5	54
Claude Sonnet 5	89.7 / 30	87.2 / 40	88.5	46
LongCat 2.0	89.7 / 49	89.7 / 41	89.7	12
DeepSeek V4 Flash	87.2 / 22	87.2 / 54	87.2	41
Qwen 3.7 Max	87.2 / 23	84.6 / 43	85.9	50
Grok 4.5	89.7 / 58	92.3 / 46	91.0	4
Kimi 2.7	79.5 / 33	82.1 / 37	80.8	42
MiMo V2.5	87.2 / 31	92.3 / 29	89.8	62
Hunyuan 3	69.2 / 36	82.1 / 51	75.7	18
Step 3.7 Flash	84.6 / 39	84.6 / 24	84.6	51
Gemini 3.5 Flash	87.2 / 52	84.6 / 58	85.9	0
MiniMax M3	71.8 / 50	87.2 / 38	79.5	16

GLM 5.2's trial 2 shown is the clean replacement (§6.1); the original infra-killed trial (23.1 / 9 calls, 25 errors) was discarded.

Run #20 · OTC Desk Agent Arena · generated 2026-07-13.