The OTC Desk Agent Arena — Methodology & Results (Run #94)

*The third workflow of the triptych. Run #20 built the Model Ability Card and ranked a risk manager's control day; Run #33 replicated the verdict on a trader's RFQ-to-booking day and crowned the same winner. Run #94 points the instrument at a board-governance review day** — and the board reorders. GPT-5.6 Terra, #1 on both prior workflows, lands

7; GPT-5.6 Luna falls from the podium to #10; Gemini 3.6 Flash posts the first perfect

card in arena history. This report explains why — and why the explanation is neither "the workflow is broken" nor "OpenAI turned the models down," but something more useful: **the ranking signal itself migrated, and the card lets you watch it move.***

Date: 2026-07-29 · **Run:** #94 · **Task:** `high-board-portfolio-review-day` (8-step / 35-check, post-audit manifest) · **Trials:** 2 per model · **Gateway:** ZenMux · **Field:** 18 models, cross-tier (the Run #33 seventeen + Doubao Seed 2.1 Pro)

Abstract

Two workflows and two desk roles had produced one conclusion: objective capability saturates at the top, so *efficiency* (EFF) and *consistency* (CON) decide deployment — and GPT-5.6 Terra, the leanest near-top operator, won both boards. Run #94 runs the same instrument, field, and headless regime against `high-board-portfolio-review-day`, a governance workflow rebuilt after the Run #58 scoring-validity audit: 35 checks, zero non-discriminating, answer-level grounding keyed to values that exist **only inside a persisted report artifact**, and a calibrated par of 24 that the field — for the first time — actually meets (median calls/par **0.96**, vs 1.50 on Run #20 and 1.74 on Run #33).

That last number is the whole story. On the first two boards, 85–97% of trials over-executed par, EFF spread across a 0–84 range, and EFF was the strongest single rank-correlate of OVR (ρ≈0.63–0.66): the boards rewarded exactly the trait Terra has in extremis. On high-board the field runs *at* par, EFF compresses (ρ 0.44), and the ranking signal migrates to **synthesis** (σ 28.3, **ρ 0.87** — the widest, most rank-decisive axis on any board to date) and **grounding** (ρ 0.81). Terra's profile is the mirror image of that signal: both trials, identically (CON 99, objective stdev

0.0), it issued **one** portfolio-scoped `list_reports` call, received an honest empty result, declared last quarter's governance report nonexistent, recorded `null`, and shipped a terse final artifact missing the required evidence — losing 6 checks across GRD, SYN, and PRC while running 9–12 tool calls against a par of 24. The efficiency that won two titles and the evidence-abandonment that lost this one are **the same behavioral policy** measured under different task demands. Luna's drop is a different mechanism: one near-perfect trial (97.1) and one that whiffed the opening skill routing (77.1) — CON 30, the consistency discount doing its job.

To separate "OpenAI family trait" from "Terra-specific policy," we added **GPT-5.6 Sol** to the registry and ran it through the identical board (run #95). Sol hit the identical trap — the very same filtered-first `list_reports` call, the same honest empty result — and did the opposite: a nineteen-call evidence hunt across the artifact store, grep, and the report registry that found the report, quoted 211.34, and closed out **two perfect trials (35/35, objective 100.0, stdev 0.0)** at 45 and 30 calls against par 24 (EFF 43, OVR 85). The abandonment is not a family trait; it is Terra's policy. The three GPT-5.6 variants share the same opening move and then span the lean↔thorough spectrum nearly end to end — which reads as deliberate productization of behavioral policies, not a capability change. The harness could be strengthened to *remove* this discriminator — seeding the fixture report's `request_payload.portfolio_id` would make Terra's filtered query succeed — but the audit's own standard cuts the other way: filtered-first was the *majority* opening move (9 of 13 transcripts), the recovery path is real and cheap (six models retried unfiltered and found the report), and "persistence when a presupposed document fails to surface on the first query" is a trait a desk genuinely selects on. Three boards, one instrument, and the triptych's refined conclusion: **there is no workflow-independent "best operator" — there are behavioral policies, and a card that decomposes them is what lets you match a model to the job.**

---

1. Introduction — the question the first two boards begged

Run #33's abstract closed on a confident replication: same winner, same profile, two desk roles. It also quietly assumed the profile itself — *near-top capability delivered leanly* — was workflow-independent. High-board was built to stress exactly the part of desk work the first two flagships under-sampled: **evidence retrieval against persisted state, and synthesis of that evidence into a governance deliverable.** A control day and a booking day are procedural gauntlets — do the steps, in order, without inventing numbers. A board-review day is an *evidential* gauntlet: the numbers you must quote live in a prior quarter's report artifact, a persisted risk run, and a view you yourself constructed three steps ago; the deliverable is judged on whether that evidence actually arrives inside it.

If the "best operator" verdict were workflow-independent, high-board should re-elect Terra. It did not — and the decomposed card turns that upset from an anecdote into a mechanism.

2. The task — `high-board-portfolio-review-day`

The model plays the desk's **board-review operator** on a governance day: build a board-scoped review view over the Desk Control Book, count its Snowball concentration, read the **persisted governed risk run** (consume-only: the workflow seeds a completed `RiskRun`; dispatching a fresh one is off-script), quote the governed NVDA delta, resist an over-claim trap, retrieve **last quarter's board governance report** and quote the prior governed valuation it states, and draft the new quarter's governance report as Markdown, grounded in the day's evidence.

Scoring runs on the post-audit manifest — the Run #58 validity audit found 15 of the original 50 checks carried no signal (dead, free, or double-scored); the rebuilt manifest has **35 checks** (12 grounding / 11 procedural / 7 adherence / 5 synthesis), a live-verified **zero** of which are non-discriminating (field pass rates span 7/18–17/18, with a handful of deliberate floors). Two properties matter for this report's headline:

- **Answer-level grounding, keyed to artifact-only values.** The prior governed valuation (**211.34**) is stated *only* in the seeded report's artifact body — never in `result_payload` — so it cannot be computed, guessed, or brute-forced by enumeration; the model must find and open the right report. The value sits >12% from every other graded number, far outside `rel_tol`, so a swapped answer fails.
- **A calibrated par of 24** (measured from competent runs, not the theoretical minimum), which the field met: this is the first board where EFF stopped being the scarcest stat.

The manifest was **de-biased twice on 2026-07-27 — both times in Terra's favor** — before this board ran: a double-jeopardy fix (its missed `get_report` had been charged three times: the expected-tool check, a bare duplicate `tool_called`, and the success sequence) and an explicit "as Markdown" in the final prompt (Terra had reasonably chosen DOCX, got `kind="binary"`, and lost all five synthesis greps to an unstated format preference). What Terra loses on this board, it loses to checks that survived an audit *and* two rounds of fairness repair.

3. Field & conditions

The Run #33 seventeen plus **Doubao Seed 2.1 Pro** (added to the field 2026-07-29, backfilled onto both prior boards). Two trials per model, jury **off** (objective axis only; `subjective_mode="disabled"`), banked as per-model runs #85–#93 and folded with the shipped `scoring.fold_trial_breakdowns` kernel into board #94. One infrastructure note:

GLM 5.2's initial rows were scored across a broken integration — through ZenMux's OpenAI-compatible gateway it emits tool calls with empty-string ids, so every persona delegation failed before starting; after pinning it to `protocol: anthropic` (the fourth model needing the pin) it was re-run clean and the board re-merged. Its row below reflects the pinned re-run. Gemini 3.5 Flash lost one trial to infra (skipped, not retried, per the trials design) and is scored on n=1.

4. Results

4.1 The board

#	Model	OVR	CON	GRD	ADH	SYN	EFF	PRC	Objective
1	Gemini 3.6 Flash	**99**	99	99	99	99	99	99	100.0
2	Grok 4.5	96	99	99	99	99	80	99	100.0
3	DeepSeek V4 Flash	91	96	91	99	89	81	94	94.3
4	DeepSeek V4 Pro	88	76	95	99	99	68	86	94.3
5	Qwen 3.7 Max	88	92	82	99	99	76	90	91.4
6	LongCat 2.0	83	50	90	99	99	72	94	95.7
7	GPT-5.6 Terra	83	99	82	99	**59**	82	81	82.9
8	Step 3.7 Flash	81	86	74	99	79	74	90	85.8
9	Claude Sonnet 5	79	59	74	92	99	84	81	84.2
10	GPT-5.6 Luna	74	**30**	78	99	69	82	94	87.1
11	Gemini 3.5 Flash	74	n/1	74	99	59	33	99	85.7
12	Hunyuan 3	71	40	82	99	40	78	81	80.0
13	Kimi 2.7	68	33	78	99	40	76	76	77.2
14	MiMo 2.5	68	40	74	99	50	76	68	74.3
15	Doubao Seed 2.1 Pro	66	0	78	99	50	78	90	82.8
16	MiMo 2.5 Pro	65	10	78	99	50	76	68	75.7
17	GLM 5.2	64	99	70	99	0	64	58	62.9
18	MiniMax M3	42	0	50	78	40	35	32	48.6

Headlines the prior two boards make legible:

- **Gemini 3.6 Flash posts the arena's first perfect card** — 35/35 checks in both trials, 21–24 calls against par 24, CON 99. On Run #33 the same model was the "capability made deployable" story (identical objective to its predecessor at 2/3 the calls); here it simply has no weakness for any axis to find.
- **Grok 4.5's rehabilitation is the EFF story inverted.** Its profile barely changed — it remains a maximally correct, somewhat verbose executor (100.0 objective, 28–29 calls). What changed is the denominator: 28 calls against par 35 on trader-rfq earned EFF 15 and OVR #5; 28 against par 24 here is only 1.2×par on the golf curve — EFF 80, OVR #2. Two boards agree on what Grok *is*; they price it differently.
- **The OpenAI pair inverts.** Terra #1 → #7 and Luna #2/#3 → #10 — examined next.

4.2 Why Terra fell — one policy, six checks

Terra's card reads GRD 82 · ADH 99 · **SYN 59** · EFF 82 · PRC 81, **CON 99, objective stdev 0.0** — both trials failed the *same* six checks. That consistency is the diagnosis: this is not noise, not infra, not a bad day; it is a policy.

The transcript shows the policy in one exchange. Step 7 asks: "*Pull last quarter's board governance report for context*" and requires `record_answer` of the valuation it states. Terra called `list_reports(portfolio_id=9101, limit=100)` — a reasonable, *documented* filter (the tool's `portfolio_id` matches against `request_payload`) — and received `{"reports": [], "total": 0}`, because the seeded report is a desk-level row whose `request_payload` carries no portfolio id. A scoped filter **cannot** surface it. Terra's response, verbatim:

> "*No prior-quarter board governance report is persisted for Desk Control Book: the report registry returned zero reports for portfolio 9101. Therefore, there is no report-stated prior governed valuation to quote; I recorded it as* `null`."

One call, one empty result, one confident declaration of nonexistence — against a user turn that *presupposes the report exists*. The miss cascades through six checks: the `get_report` expected-tool check, the report-type grounding check, the 211.34 answer quote, two synthesis greps on the final artifact (which, in Terra's characteristically terse rendition, omits "Snowball" and the 17.5 NVDA delta), and the success-sequence check. Twelve and nine tool calls in its two trials — half of par — is the same leanness that won Runs #20 and #33, now operating as evidence-abandonment.

The field context makes it a fair loss. Opening with a *filtered* query was the majority move — 9 of the 13 banked transcripts did — so the trap is not "you filtered"; it is "what did you do when the filter came back empty against a presupposition." Six models retried unfiltered, found report #5, and quoted 211.34. Three (Terra, Kimi, MiMo 2.5 Pro) declared defeat. Qwen never opened the report and recorded the *current* valuation (238.05) — precisely the swapped-answer failure the 211.34 keying was designed to catch. MiniMax, remarkably, quoted 211.34 *without* ever calling `get_report` — it read the materialized artifact file directly, and the answer-level check credited the answer while the trace checks scored the skipped procedure: the two-layer grading working exactly as designed.

4.3 Why Luna fell — variance, not weakness

Luna's two trials: **97.1** (one missed check — the same artifact-grep on "17.5") and **77.1** (missed the `portfolio-maintenance` skill routing at the *opening* step, muddled the container-vs-view distinction, and dragged wrong counts through every downstream answer). At step 7 — Terra's stumbling block — Luna is actually the field's most dogged retrier: four filtered variants, then unfiltered, then `get_report`, then the correct 211.34. Luna at full strength is a top-3 operator on this board; Luna's problem is that you get that operator one trial in two. CON 30 applies the card's inconsistency discount (up to −18% of OVR at CON 0) and lands it #10. On a governance task run unattended, that discount is the instrument saying something true.

4.4 GPT-5.6 Sol — family trait, or Terra's policy?

Terra's collapse invites a vendor-level reading: did OpenAI quietly turn the family down? The falsifiable version of that question is an A/B inside the family. We registered **GPT-5.6 Sol** (the third 5.6 variant on ZenMux, absent from all prior boards), probed its tool-call ids clean through the production factory path, and ran it through the identical manifest as run #95 — same fixtures, same par, same jury-off regime, same day.

Sol: objective 100.0 in both trials (stdev 0.0), 35/35 checks — the board's third perfect objective — at 45 and 30 tool calls (par 24). Folded card: GRD 99 · ADH 99 · SYN 99 · PRC 99 · **EFF 43**, CON 66 (the 45→30 call swing), **OVR 85** — which would slot ~#4 on the board, between the DeepSeek pair and LongCat.

The step-7 transcript is the experiment's payoff. Sol opened with **the exact same call Terra did** — `list_reports(portfolio_id=9101, limit=100)` — and received the same `{"reports": [], "total": 0}`. Then, where Terra declared nonexistence, Sol launched a nineteen-call hunt: two `list_artifacts` sweeps, five `inspect_artifact` / six `read_artifact` probes into its own session evidence, a `grep` for "governance" over large tool results,

`get_report(report_id=5)`, and a correct `record_answer(prior_governed_valuation=211.34)` with the source cited.

So the family picture across three siblings on one manifest:

Variant	Opening move	Recovery policy	Calls (par 24)	Objective	OVR
Terra	filtered `list_reports`	**none** — declares absence, records null	9–12	82.9	83
Luna	filtered `list_reports`	4 filtered retries → bare → found	21–23	87.1 (97.1/77.1)	74
Sol	filtered `list_reports`	exhaustive multi-modal hunt → found	30–45	**100.0**	85

The shared opening (all three filter first, as did most of the field) looks like common heritage. The recovery policies diverge completely — and they are each variant's *signature across the whole day*, not a step-7 quirk: Terra runs every step at half par, Sol runs every step heavy, Luna sits at par with an unstable opening. Three conclusions follow. **First, the "intelligence turned down" hypothesis is refuted**: a same-family, same-gateway, same-day variant just posted a perfect board. **Second, Terra's loss is a policy, not a defect** — its two prior titles and this defeat are one coherent behavior priced by different tasks. **Third, the variants read as deliberate productization** — OpenAI appears to ship the lean/balanced/ thorough trade-off as a product line, and the card measures exactly the axis they differentiate on: Sol buys +17 objective points over Terra for ~3× the tool calls; EFF 82 vs 43 is that invoice, itemized.

5. The signal migration — how this workflow differs from the other two

The cross-workflow comparison is the report's core finding. Same instrument, same field, three tasks:

	Run #20 (risk-mgr)	Run #33 (trader-rfq)	Run #94 (high-board)
Median calls / par	1.50	1.74	**0.96**
Trials over par	85%	97%	31%
Trials over 2×par	24%	19%	3%
EFF: field σ / ρ vs OVR	24.1 / **0.63**	22.2 / **0.66**	15.5 / 0.44
SYN: field σ / ρ vs OVR	12.8 / 0.43	18.3 / 0.58	**28.3 / 0.87**
GRD: ρ vs OVR	0.73	0.56	**0.81**
CON: field mean	80	75	**56**
Check mix (dominant axis)	22/39 procedural	21+20/63 proc+adh	**12/35 grounding**

Read down the columns and the mechanism is explicit:

- **On the first two boards, EFF was the scarcest stat and the strongest single rank signal.** Nearly the whole field over-executed par; the axis with the widest spread and highest OVR correlation was efficiency. A lean-by-policy model banks a large, reliable edge — Terra's two titles.
- **High-board pays out on different axes.** The field runs at par (so EFF compresses and decorrelates), adherence saturates (mean 97.4 — the prohibitions are easy here), and what varies is whether models *retrieve and carry evidence*: synthesis spreads across the full 0–99 range with ρ 0.87, grounding ρ 0.81. The task changed which behaviors are scarce; the ranking followed the scarcity.
- **High-board also destabilizes.** Field-mean CON drops from ~80/75 to 56 — seven models show CON ≤ 40 here. The opening step (route to `portfolio-maintenance`, respect the container/view distinction) is a variance generator the procedural workflows didn't have: miss it and every downstream count is wrong (Luna's bad trial; Doubao's CON 0).

The refined triptych conclusion, then, is not "Run #33's verdict was wrong" — on procedural workflows it stands — but that it was **scoped**: "*best operator*" *is a match between a model's behavioral policy and a workflow's scarce axis, not a total order over models.* Terra remains the model you want on a well-mapped procedural pipeline where every extra call is waste. It is demonstrably not the model you want walking into a governance review where the evidence must be hunted down. The card — because it decomposes — is what lets a desk make that assignment; a single-number leaderboard would have hidden the migration entirely.

6. Harness vs. behavior — what we deliberately did not "fix"

Every upset invites the question: strengthen the harness until the upset disappears? Three decisions in this run drew that line, and the audit doctrine drew it for us:

1. **The step-7 discriminator stays.** Seeding the fixture report's `request_payload.portfolio_id` (a one-line `$seed` change) would make Terra's filtered query succeed — and flip a check that currently splits the field 15/3 into a free pass. The audit standard condemns spreads caused by *broken fixtures*; this fixture is not broken: the report is discoverable by the default call, the recovery costs one tool call, and most of the field executes it. Persistence under a failed first query is desk-relevant behavior, measured fairly.
2. **The unfairness that *was* found was already repaired — in Terra's favor.** The 2026-07-27 double-jeopardy and "as Markdown" fixes (§2) are the model of the right harness response: charge one mistake once, never grade an unstated preference. Post-repair, what remains of Terra's deficit is behavior, not instrumentation.
3. **Behavioral diversity is the finding, not the bug.** The per-check field tally — the same instrument that condemned 15 checks on Run #58 — certifies every check on this board discriminates or floors deliberately. When a healthy instrument reorders a board across tasks, the correct inference is that the *tasks* differ, and model selection should too.

7. Limitations & threats to validity

- **n=2 trials.** CON from two samples is coarse (one bad trial → CON 30); the stat is honest about disagreement but cannot distinguish a 50%-variance model from an unlucky one. The banked-runs design makes topping up cheap.
- **Step-7 taxonomy is last-trial-only** for models whose earlier trial transcript was superseded; aggregate check pass-rates (15/18 on `get_report`) carry the both-trials signal.
- **Cross-board comparisons ride on par calibration.** Par 24 here is measured; Run #33's par 35 was recalibrated once already. EFF's cross-workflow meaning is only as stable as those measurements.
- **One model scored on n=1** (Gemini 3.5 Flash, infra-skipped trial) and is flagged in place.
- **Scores are manifest-relative.** Terra's 62.0 on the pre-audit high-board (Runs #34/#58) vs 82.9 here is instrument repair, not model change; no cross-manifest comparison appears in this report.

8. Reproducibility

Ground truth is the live DB: `store.leaderboard(run_id=94)` (aggregates), per-trial breakdowns in `arena_match.score_breakdown` (runs #85–#93), transcripts under `artifacts/arena/<run>/high-board-portfolio-review-day/<model>/transcript.json`. The Sol A/B is run #95 via the standard `queue_arena_run` + `execute_arena_run_task` path. Axis-spread and Spearman numbers derive from the stored cards exactly as `_derive_card` serves them; the discrimination table's script is checked into the session record. QuantArk is pinned (`quantark==0.3.0`) as of `ea07986`, so every graded constant in this run reproduces from a clean environment.

Appendix A — step-7 behavioral taxonomy (banked transcripts)

Model	First call filtered?	`list_reports` calls	Retried bare?	Opened report?	Recorded answer
Gemini 3.6 Flash	yes	2	yes	yes	211.34 ✓
GLM 5.2	yes	2	yes	yes	211.34 ✓
GPT-5.6 Luna	yes	5	yes	yes	211.34 ✓
LongCat 2.0	no (bare)	3	—	yes	211.34 ✓
Hunyuan 3	no (bare)	1	—	yes	211.34 ✓
Claude Sonnet 5	yes	2	yes	yes	null ✗ (opened, then declined to quote)
Step 3.7 Flash	yes	4	yes	no	null ✗
GPT-5.6 Terra	**yes**	**1**	**no**	**no**	**null ✗ (declared nonexistent)**
Kimi 2.7	yes	2	no	no	null ✗
MiMo 2.5 Pro	yes	2	no	no	null ✗
Qwen 3.7 Max	yes	1	no	no	238.05 ✗ (current valuation — the swap trap)
MiMo 2.5	(0 calls)	0	—	no	null ✗
MiniMax M3	(0 calls)	0	—	no	211.34 ✓ (read the artifact file directly)
GPT-5.6 Sol (run #95)	yes	1	no — hunted via artifacts + grep instead	yes	211.34 ✓

Rendered from `docs/arena/2026-07-29-run94-otc-desk-agent-arena.md` · OTC Desk Agent Arena · Run #94.