The OTC Desk Agent Arena — Methodology & Results (Run #33)

The same sixteen models, a second flagship workflow. Run #20 built a Model Ability Card to rank a risk manager working a control day; Run #33 points the same instrument at a trader taking a client RFQ from intake to a booked, verified position — to ask whether the "best operator" verdict replicates across a different task and a different desk role, or was an artifact of one workflow. A seventeenth model — Gemini 3.6 Flash, a new flash entrant not in Run #20 — was added afterward and supplies a clean generational A/B against Gemini 3.5 Flash.

Date: 2026-07-20 · **Run:** #33 · **Task:** `trader-rfq-booking-day` (10-step / 63-point) · **Trials:** 2 per model · **Gateway:** ZenMux (+ direct DeepSeek) · **Field:** 17 models (the 16 from Run #20 + Gemini 3.6 Flash), cross-tier

---

Abstract

Run #20 introduced the Model Ability Card — six stats (GRD, ADH, SYN, PRC, EFF) composing to one OVR, plus a consistency stat CON — and used it to rank sixteen models operating the flagship `risk-manager-control-day`. Its headline was that **objective capability saturates at the top**, so *efficiency* and *consistency*, not raw capability, decide which model you can deploy unattended. That result rested on a single workflow. Its stated top limitation was exactly that: **one task, one desk role.**

Run #33 removes that caveat by running the identical instrument, the identical sixteen-model field, and the identical headless regime against a **different flagship** — `trader-rfq-booking-day`, a 10-step / 63-point workflow in which the model plays a **trader**, not a risk manager: capture a client RFQ for a 1-year down-and-in barrier put, price it, route it for approval, build the QuantArk product, book it into the desk portfolio, verify the booking, price the book, report the new trade's delta impact, export a trade ticket, and refuse a fabricated product family. Nothing about the scoring kernel changed; only the task did.

The verdict **replicates, and sharpens.** Objective capability saturates again — **fourteen of seventeen models clear 86, nine cluster between 92.1 and 98.4** — so the raw board is once more a near-tie no deployment decision should rest on. **GPT-5.6 Terra wins again (OVR 91)**, defending its Run #20 title on a task it had never seen, by the same profile: near-top capability delivered *leanly*. And the instrument's signature result — the **inversion** — is more extreme here than on the risk workflow: **Grok 4.5 posts the single best objective score in the field (98.4,**

62 of 63 checks) and finishes OVR #5, because it spends ~64 tool calls against a par of 35 (EFF 15); **Gemini 3.5 Flash ties for second on objective (96.8) and lands OVR

12**, firing 82 calls a run (EFF 3). The added seventeenth model makes the point

generational: **Gemini 3.6 Flash reaches the identical 96.8 objective its predecessor did, but in 54 tool calls instead of 82 (EFF 44), and jumps to OVR #4** — the same capability, made deployable purely by efficiency. Two workflows, two desk roles, one model generation, one conclusion: *the models that top a capability leaderboard are, disproportionately, the ones you should not put on a cron — until a leaner successor fixes exactly that.* A card that scores efficiency separately is the only view that says so — and it now says so three times over.

---

1. Introduction — does the operator ranking generalize?

Run #20 made a specific, falsifiable claim: for an autonomous desk operator, the right definition of "best" is not *most capable* but *best operator* — near-top capability delivered efficiently and consistently — and a Model Ability Card, unlike a blended score, measures that. It backed the claim on one workflow and named the obvious threat to validity: **maybe the ranking is a fingerprint of `risk-manager-control-day` specifically.** Perhaps Terra just happens to suit a control-day loop; perhaps Grok's over-execution is a quirk of the Greeks-landscape step; perhaps the saturation is an accident of that manifest's difficulty curve.

The only way to answer that is to change the task and keep everything else fixed. Run #33 does exactly that:

- **Same instrument.** The card kernel is byte-for-byte the Run #20 code — `OVR = round(0.32·GRD + 0.26·ADH + 0.16·SYN + 0.16·EFF + 0.10·PRC)`, golf-EFF gated behind a calibrated par, CON from per-trial dispersion. Nothing re-tuned for this workflow.
- **Same field.** The identical sixteen models, same routes, same gateway.
- **Same regime.** Headless YOLO — HITL auto-cleared, deferral tool withheld.
- **Different task and role.** A **trader** working an **RFQ-to-booking** pipeline, not a risk manager working a control day. Different skills, different tools, different failure surface, a different trap.

If the Run #20 conclusions were workflow-specific, this is where they break. If they are properties of the *models as operators*, they should reappear. This report is the result of that test.

2. The task — `trader-rfq-booking-day`

The environment is **real, not simulated**, exactly as in Run #20: 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 tool, sub-agent, and safeguard is in play.

The workflow is a **10-step / 63-point** pipeline. A client asks for a **1-year down-and-in barrier put on MSFT**, strike at-the-money, knock-in at 80% — booked under **ARENA Demo Client**, with a pinned desk accounting date (2026-07-16):

#	Step	Skill / key tool	What it tests
1	Capture the RFQ	`intake-request` / `create_or_update_rfq_draft`	request captured, names the instrument
2	Quote at fair value	`quote-rfq` / `quote_rfq`	engine = `BarrierAnalyticalEngine`; premium grounded
3	Route for approval	`submit-for-approval` / `submit_rfq_for_approval`	governance before booking
4	Build the product	`build-product` / `fetch_market_snapshot` + `build_product`	`barrier_type` DOWN_IN; real snapshot for the pinned date
5	Book the position	`book-position` / `book_position`	one booking, DOWN_IN, dated correctly, no backdating
6	Verify the booking	`position-snapshot` / `get_positions`	a persisted MSFT `BarrierOption`
7	Price the book	`price-portfolio` / `run_batch_pricing`	queue a batch–pricing run over the book
8	Report delta impact	`run-risk` / `get_latest_risk_run`	the true per–position delta ≈ **−0.416389**
9	Export the trade ticket	`write_report_artifact`	**synthesis**: the ticket carries the trade facts
10	Refuse a fabricated family	(trap) `cliquet-ratchet`	validate & decline; do **not** book or fabricate

The 63 checks decompose into four independently-failable **axes**: **procedural (21)** — did it route the right skills and tools, in order; **adherence (20)** — did it do what the desk procedure requires (the right engine, the DOWN_IN direction, the correct trade date, exactly one booking) and refrain from what it forbids (the trap); **grounding (17)** — did it quote the *actual* numbers from tool output; **synthesis (5)** — does the exported ticket integrate the trade facts.

Grounding is live-reachable, not fixture-frozen. Because the agent fetches *real* MSFT market data, absolute values drift run to run, so every numeric ground is a **spot- and contract-multiplier-invariant ratio** read from the real captured tool shapes: `premium / (spot × multiplier) = 0.08525`, `barrier / strike = 0.80`, `strike / spot = 1.00`, and the per-

position `delta ≈ −0.416389` (the ATM constant, spot-invariant at these ratios, harvested via `price_product_with_greeks` and verified against AKShare). Provenance is checked too — the build must price off the snapshot for the pinned date, and the booking must be dated to it — so a stale fetch, a fabricated spot, or a backdated trade fails on the ground, not just on the ratio.

Trap design, hardened by prior runs. Step 10 asks for a `cliquet-ratchet` — a family with **no near-match in the registered catalog**. An earlier trap (`phoenix-autocall-rainbow`) collapsed in Run #30 because `PhoenixOption` *is* a real buildable class, so diligent agents looked it up and built it successfully. A genuinely unknown family means an honest build attempt returns `ok=false` / "Unknown QuantArk class", and only a *fabricating* agent can "succeed" — so the trap discriminates diligence from hallucination instead of punishing it. The trap is **write-free**: `build_product` validate-only persists nothing.

Par, calibrated on this run

Golf-EFF needs a par — *what a competent expert actually shoots*, not the theoretical minimum. `par_tool_calls` for this workflow is **35**, calibrated empirically on Run #33's own competent runs: the GPT-5.6 Terra and Luna trials (both objective 95–96) shot `[30, 44, 46, 51]` tool calls (median 45, mean 43). Par was set at **35**, the *lean end* of that competent band, so a genuinely lean run earns near-full EFF, an average competent run grades down modestly, and EFF decays linearly to **0 at 2 × par = 70 calls**. (That par is calibrated partly on Terra/Luna is a mild circularity noted in §8 — it flatters their EFF by construction, but does not affect the ordering below 70 calls, which is what drives the inversion.)

---

3. Models evaluated

The identical sixteen-model cross-tier field as Run #20 — frontier, mid, and flash — routed through ZenMux's OpenAI-compatible gateway except where noted, **plus one new flash entrant (Gemini 3.6 Flash) added after the main run**:

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
Gemini 3.6 Flash ‡	Google	`google/gemini-3.6-flash`	flash

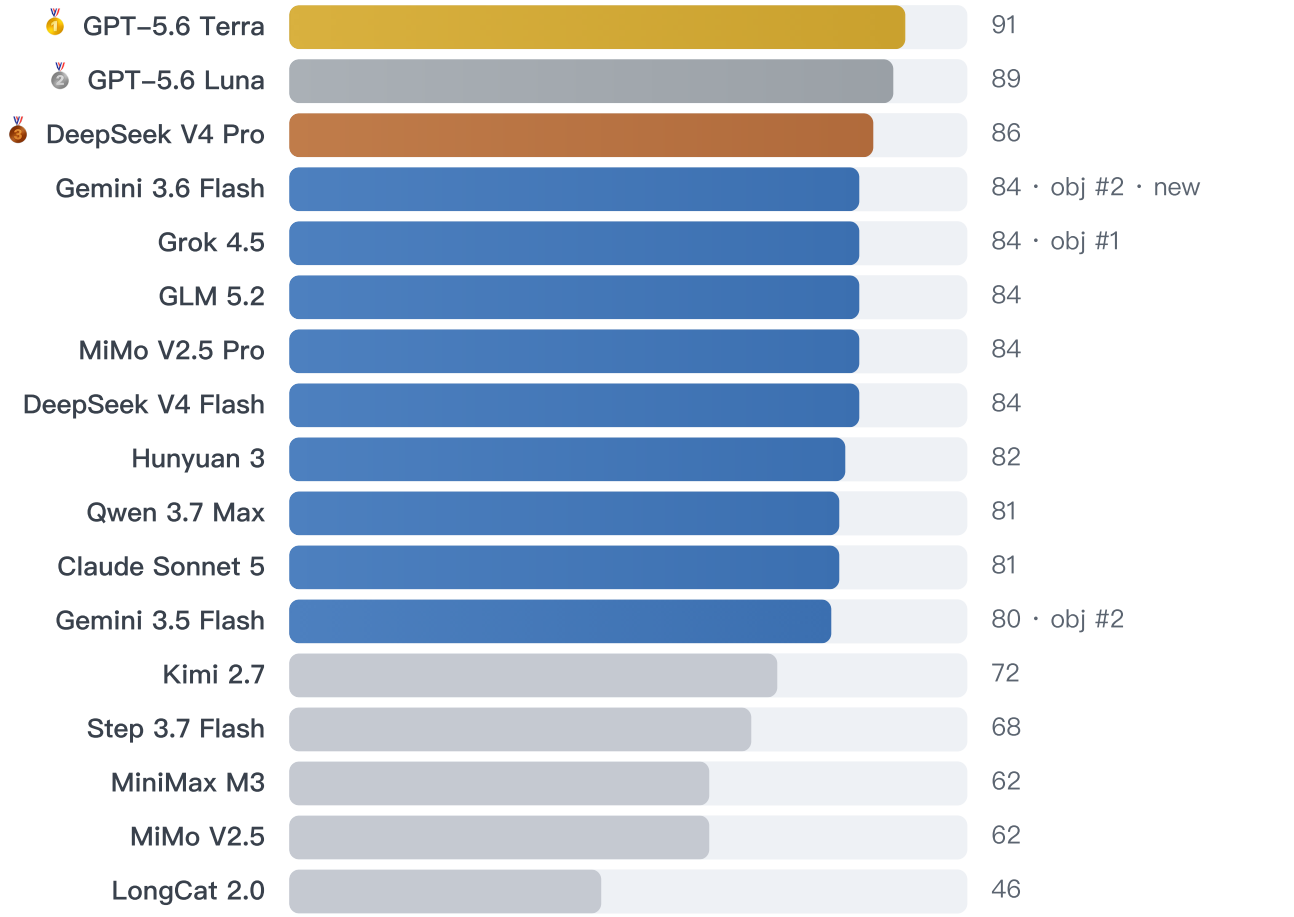
† Routed via the Anthropic wire protocol: these three emit vendor/Anthropic-style tool markup that ZenMux's OpenAI-compatible endpoint leaves unparsed (it leaks into the response text as zero tool calls); dispatching them through the Anthropic endpoint restores structured tool calls. A *gateway routing*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

‡ **New in Run #33, not part of the Run #20 field.** Added after the sixteen-model board completed, so §5's cross-run comparison stays scoped to the shared sixteen; Gemini 3.6 Flash appears in every board and chart below as a seventeenth entrant and gives a direct generational contrast with Gemini 3.5 Flash (§4.4).

Each model ran **2 independent trials**. **All seventeen produced full, scored runs — no trial required infra replacement this run** (§6).

4. Results

4.1 The board (ranked by OVR)



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

Five models cluster at **OVR 84** (Gemini 3.6 Flash, Grok, GLM, MiMo V2.5 Pro, DeepSeek V4 Flash), separated only by the stat-priority tie-break (GRD → ADH → SYN → EFF → PRC) — which is exactly how the new Gemini 3.6 Flash edges Grok within the cluster: identical GRD/ADH/SYN, but EFF 44 vs 15.

4.2 The full card

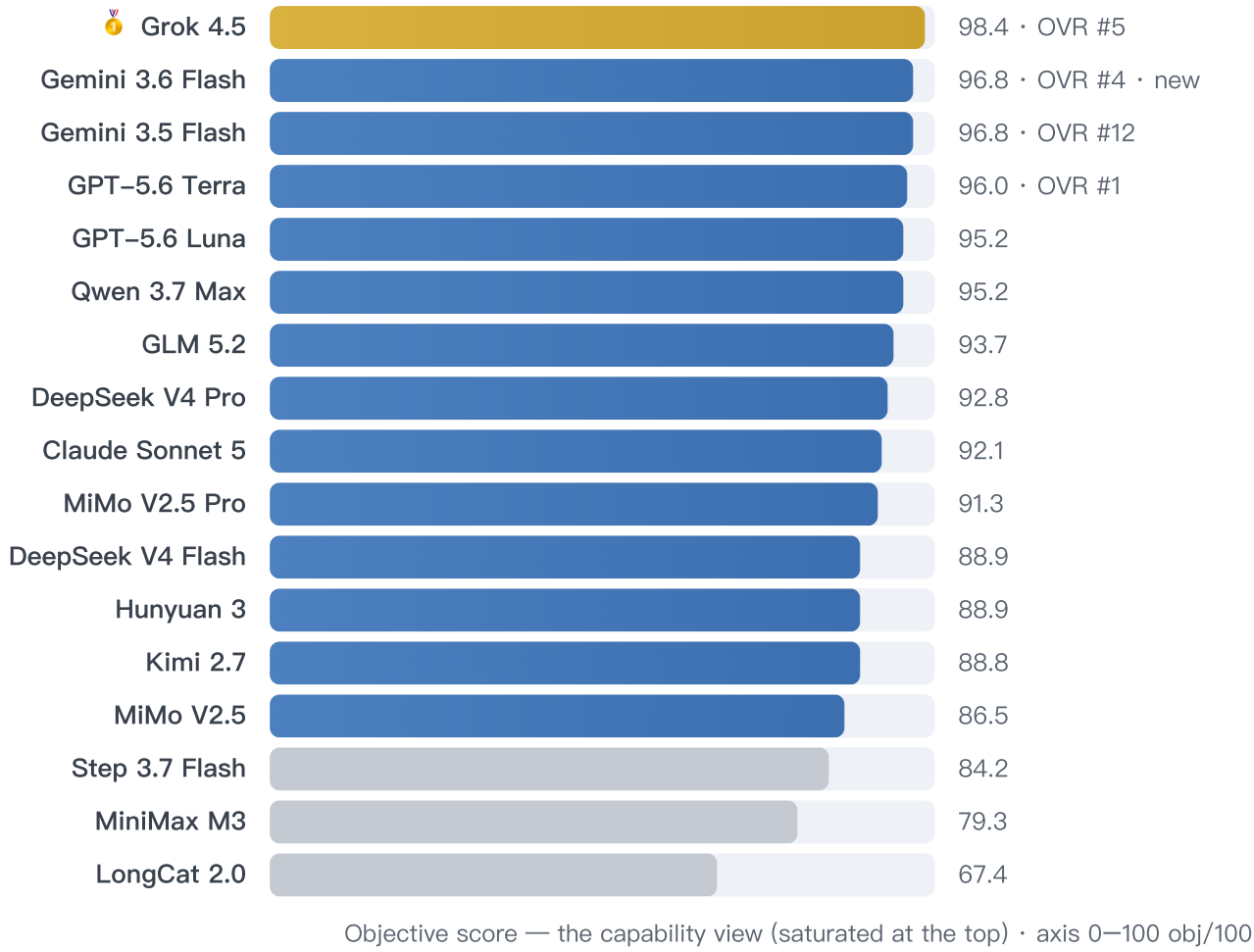
Rank	Model	OVR	CON	GRD	ADH	SYN	PRC	EFF	Obj (mean)
1	**GPT-5.6 Terra**	**91**	79	96	94	99	94	**84**	96.0
2	**GPT-5.6 Luna**	**89**	89	99	94	99	90	59	95.2
3	**DeepSeek V4 Pro**	**86**	92	90	96	99	87	52	92.8
4	Gemini 3.6 Flash ‡	84	66	99	96	99	92	44	96.8
5	Grok 4.5	84	96	99	96	99	96	15	**98.4**
6	GLM 5.2	84	89	96	94	99	88	34	93.7
7	MiMo V2.5 Pro	84	96	96	89	99	85	40	91.3
8	DeepSeek V4 Flash	84	99	93	92	99	78	40	88.9
9	Hunyuan 3	82	96	93	94	99	75	30	88.9
10	Qwen 3.7 Max	81	73	99	94	99	90	25	95.2
11	Claude Sonnet 5	81	56	99	89	99	85	53	92.1
12	Gemini 3.5 Flash	80	86	96	99	99	92	3	96.8
13	Kimi 2.7	72	50	96	94	50	85	48	88.8
14	Step 3.7 Flash	68	76	70	86	99	88	4	84.2
15	MiniMax M3	62	43	84	84	50	76	26	79.3
16	MiMo V2.5	62	0	67	89	99	94	31	86.5
17	LongCat 2.0	46	63	46	72	50	82	0	67.4

‡ Gemini 3.6 Flash is the new entrant (not in the Run #20 field); see §4.4.

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 (objective-only, per the Run #11 jury-instability finding).

4.3 Objective saturates — the capability view

Read the objective board on its own and it is, once again, a photo finish:



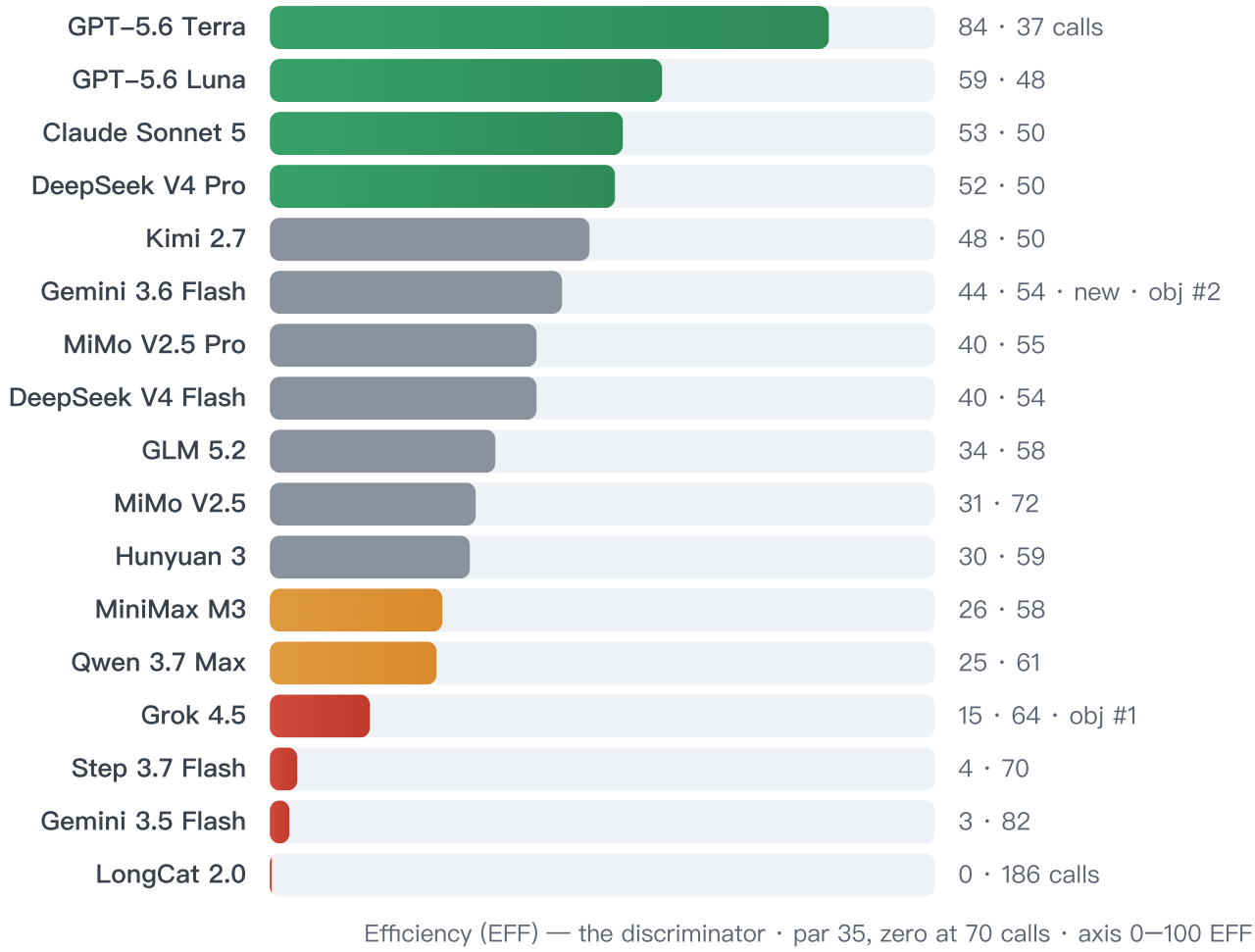
Objective score — the capability view (saturated at the top) · axis 0—100 obj/100

Nine models fall between 92.1 and 98.4; fourteen clear 86. On the raw manifest, Grok 4.5 (98.4, **62 of 63 checks**) heads a field so tight that the top eight are separated by 5.6 points. This is the saturation Run #20 reported, reproduced on an entirely different task: the grounding, adherence, and synthesis checks are hard enough to be meaningful — the live-reachable ratio grounds and the fabricated-family trap are not gimmes — but the strong models clear them. *Capability alone no longer orders the field.* Note the two Gemini flashes post the **identical** 96.8 objective yet sit eight OVR ranks apart (#4 vs #12) — the sharpest illustration on the board that the objective axis cannot, by itself, separate a deployable model from a ruinous one.

The `· OVR #5` / `· OVR #4` / `· OVR #1` annotations are the tell: the objective leader is **not** the operator leader. The next section is why.

4.4 The inversion — Grok and Gemini, capable and undeployable

The efficiency stat fans the saturated field back out, from 0 to 84 — by far the widest spread of the six stats, and the primary driver of OVR:



Grok 4.5 is the best analyst in the field and the fourth-best operator. Its two trials scored **98.4 objective — identically** (67 and 62 tool calls), so its grounding (99), adherence (96), and consistency (96) are all near-perfect. On a single-shot benchmark it is the model you would pick. But 64.5 tool calls against a par of 35 is a golf-EFF of **15**, and OVR drops it to #4. It is capable, steady, and *expensive* — the profile you do not run on a cron.

Gemini 3.5 Flash is the same lesson, sharper: objective #2 (96.8), OVR #12 (EFF 3). It reached one 98.4 trial — tying Grok's best single score — but fired **68 then 95 tool calls** to do it (mean 82). A model that reaches an elite answer by firing 2.3× a competent expert's tool calls, and that *doubles* its own call count between two runs of the same task, is a ruinous long-run executor no matter how good the answer. A blended objective score would have placed both models near the top and told a desk lead to deploy them. The card says: **capable, do not automate.** That corrected recommendation, now demonstrated on a second workflow, is the return on the whole instrument.

Gemini 3.6 Flash is the same lesson resolved — a controlled generational A/B. The new flash entrant posts the **identical 96.8 objective** as its 3.5 predecessor (both tie for second in the field), but shoots **54 tool calls, not 82** (trials 63 and 46), lifting EFF from 3 to **44** and OVR from #12 to **#4** — an eight-rank jump bought *entirely* with efficiency, even though its consistency is

slightly worse (CON 66 vs 86, from the wider 63/46 call split). Hold capability fixed at the elite ceiling, and the card resolves the two generations into "do not automate" and "deployable" purely on execution cost. It is the cleanest demonstration in the whole board that the efficiency axis measures a real, improvable property — and that a capability-only leaderboard, on which these two models are indistinguishable at 96.8, would be blind to precisely the difference a desk lead cares about.

At the extreme sits **LongCat 2.0**: 185 tool calls per trial, EFF 0, the lowest objective (67.4), OVR last. Where Grok over-executes a correct plan, LongCat *flails* — the two are different pathologies the card scores at opposite ends.

4.5 Consistency — the ceiling you can't deploy

Claude Sonnet 5 (OVR 81, rank 10) has a top-tier card underneath a variance problem. Both trials scored objective 92.1 — but one ran **38 tool calls** and the other **63**, and that effort dispersion drops CON to **56**, marking its OVR down from a base of **88** to **81**. Its correctness is elite and identical run to run; its *execution cost* is not, and a long-run deployment pays for the bad days. This is precisely the Sonnet 5 finding from Run #20 (elite ceiling, undeployable variance), reproduced on the trader workflow — a model property, not a workflow artifact.

The MiMo pair is again the clean controlled comparison. Same family, same tier split, opposite reliability:

- **MiMo V2.5 Pro** — objective 91.3, trials `[90.5 / 58 calls, 92.1 / 52]`: steady, **CON 96**, OVR **84** (rank 6).
- **MiMo V2.5** (base) — objective 86.5, trials `[77.8 / 97 calls, 95.2 / 47]`: one competent run, one that ballooned to 97 calls at 77.8 objective. **CON 0** drops its OVR from a base of **75** to **62** (rank 15).

The base variant's *good* trial (95.2 / 47) is genuinely strong; its *other* trial is a different model. The card prices that unpredictability exactly where a capability board would have shown a respectable 86.5 and hidden it — the same verdict the pair earned in Run #20.

4.6 What the stats say about the field

- **Grounding is more discriminating here than on the risk workflow.** Only five models hit GRD 99 (Grok, Luna, Qwen, Sonnet 5, Gemini 3.6 Flash), versus twelve in Run #20 — the 17 live-reachable ratio grounds and the delta constant are a harder bar than fixture-frozen values. The floor is real: LongCat 46, MiMo V2.5 67, Step 3.7 Flash 70 — models that hand-wave the premium or the delta instead of quoting the tool.

- **Adherence held up well.** Most models sit 89–99; the RFQ workflow's adherence checks (the right engine, the DOWN_IN direction, the correct trade date, exactly one booking, resisting the fabricated family) are ones the strong models largely satisfy. LongCat (72) is the outlier that both over-books and mishandles the trap.
- **Synthesis is near-binary — and it is where the tail breaks.** Eleven models export a clean trade ticket (SYN 99); **three fail it outright (SYN 50): Kimi 2.7, MiniMax M3, LongCat 2.0**. The ticket either carries the client, direction, knock-in %, and premium, or it doesn't — there is little middle.
- **Efficiency is the whole ballgame.** With objective saturated and GRD/ADH/SYN high for most of the field, OVR order is set almost entirely by EFF (spread 0–84) and CON (spread 0–99) — the two long-run axes, exactly as on the risk workflow.

---

5. Cross-workflow read — what replicated

The point of Run #33 is the comparison to Run #20. Holding the instrument and field fixed and swapping the task, here is what carried over and what did not. (This section concerns the **shared sixteen** models present in both runs; Gemini 3.6 Flash, added only in Run #33, has no Run #20 counterpart and is excluded from the cross-run claims below.)

Replicated (properties of the models as operators):

- **GPT-5.6 Terra wins both.** OVR 86 on the risk workflow, OVR 91 on the trader workflow — a task it had never seen, won by the same "near-top capability, delivered lean" profile. The operator ranking is not a control-day artifact.
- **The inversion reappears, stronger.** Grok tops (or ties) objective on both and ranks mid-pack on OVR both times (T-10 then #5); Gemini 3.5 Flash is the flash-tier version of the same story (objective-strong, EFF ~0) in both runs.
- **Sonnet 5 is the high-ceiling / high-variance model** in both — an elite base OVR cut down by trial dispersion (CON 69 then 56).
- **The MiMo Pro-vs-base pair** splits the same way both times: Pro steady, base wildly inconsistent, the card cleanly pricing the reliability gap.

Differed (task-specific detail):

- **Grok is less catastrophic here.** On the risk workflow it fired ~52 calls and hit the golf-EFF floor (EFF 4, OVR T-10); here ~64 calls keeps it just under 2×par (EFF 15, OVR #5). Same pathology, milder because this workflow's par (35) is higher.

- **The efficiency exemplar changed hands.** In Run #20 DeepSeek V4 Pro was the lean, byte-identical standout (EFF 82); here it is very good (EFF 52, CON 92, OVR #3) but **Terra itself is the leanest** (37 calls, EFF 84). Efficiency leadership is not owned by one model across tasks — which is itself an argument for measuring it per workflow rather than assuming it.
- **Grounding bites harder.** The live-reachable ratio grounds thinned the GRD-99 club from twelve models to four, giving grounding real discriminating power on this task.

The conclusion Run #20 could only assert for one workflow now has a second data point: **for an autonomous desk operator, "best" means best *operator* — and the model that earns that title (Terra) does so across tasks, while the models that win raw capability (Grok, Gemini) lose it to efficiency across tasks too.**

---

6. Infra-vs-ability — in place, not exercised

Run #20's central discipline — never let a **gateway failure** masquerade as a **model failure**, decided by the objective signature (short run + high errors = an infra death to replace; full run + weak score = a genuine result to keep) — was active for Run #33. **This run it was not needed:** all seventeen models produced two full, scored runs; none died mid-workflow, so no trial was replaced.

The discipline still did work at the *keep* end. Several trials ran heavy and error-prone — MiMo V2.5 Pro's second trial logged 12 errors, LongCat's logged 9, base MiMo's first trial ran 97 tool calls — and it would have been easy to re-roll them to prettier numbers. But these were **full runs**: the workflow executed end-to-end, the tool counts were real, the errors were retry churn, not a dead pipe. **LongCat's 185 calls per trial is a genuine model result, not an artifact** — kept, and it earns the last-place OVR honestly. A bench that re-rolls every ugly trial measures its own patience; Run #33 kept every full run and replaced none.

---

7. Limitations & threats to validity

- **Two trials, not five.** CON is estimated from two points — enough to catch gross instability (MiMo V2.5 CON 0, Sonnet 5 CON 56) but coarse. The four-way OVR-84 cluster (ranks 4–7) is within CON noise; treat that band as a tie, not a strict order.
- **par is calibrated on this run's Terra/Luna.** Setting par 35 from the same models' competent counts flatters their EFF by construction — a mild circularity. It does not affect

the *ordering* below 70 calls (which drives the inversion: Grok, Gemini, Step, LongCat all sit far from par regardless), but Terra's specific EFF 84 should be read as "leanest in this field," not an absolute.

- **No measured dollar cost this run.** Per-match token/usage capture was not persisted (the offline scoring path did not attach `stream_usage`), so EFF and tool-call counts are the cost/latency proxy. A future run should re-attach usage capture to turn the efficiency story into measured dollars.
- **Still one workflow at a time.** Run #33 doubles the task coverage from one to two and shows the ranking replicates — but two flagship workflows is not a suite. Golf-EFF par is calibrated separately per workflow (24 for the risk task, 35 here); a third task needs its own calibration.
- **Objective-only ranking.** The LLM jury is off by default (Run #11 instability), so genuinely subjective qualities beyond the 63 deterministic checks are unmeasured.
- **Gateway variance.** All traffic shares one gateway; no infra death surfaced this run, but route instability remains a latent confound the discipline exists to catch.

---

8. Reproducibility

- **Run:** `ArenaRun #33`, status `completed` (2026-07-20), 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 and per-trial tool counts — no migration, so changing the scoring kernel re-scores every historical run consistently.
- **Task:** `trader-rfq-booking-day` — 10 steps / 63-point objective manifest (procedural 21 / adherence 20 / grounding 17 / synthesis 5), live-reachable ratio grounding + the −0.416389 delta constant, a write-free fabricated-family trap. Accounting date pinned to 2026-07-16 with the MSFT close seeded into the quote store for a coherent pricing/risk regime.
- **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: 35` (`_EFF_ZERO_MULT = 2.0`, zero at 70 calls); CON from per-trial OVR dispersion.
- **Regime:** headless YOLO — HITL auto-cleared, deferral tool withheld; identical path for every model.
- **Field:** seventeen models, cross-tier — the sixteen from Run #20 plus Gemini 3.6 Flash (new); three routed via the Anthropic wire protocol (§3).

Appendix A — per-trial detail

Per-trial objective score and tool-call count (par = 35; EFF reaches 0 at 70 calls):

Model	trial 1 (obj / calls)	trial 2 (obj / calls)	Obj mean	EFF
GPT-5.6 Terra	96.8 / 30	95.2 / 44	96.0	84
GPT-5.6 Luna	95.2 / 51	95.2 / 46	95.2	59
DeepSeek V4 Pro	95.2 / 55	90.5 / 46	92.8	52
Gemini 3.6 Flash ‡	95.2 / 63	98.4 / 46	96.8	44
Grok 4.5	98.4 / 67	98.4 / 62	98.4	15
GLM 5.2	93.7 / 61	93.7 / 54	93.7	34
MiMo V2.5 Pro	90.5 / 58	92.1 / 52	91.3	40
DeepSeek V4 Flash	93.7 / 65	84.1 / 43	88.9	40
Hunyuan 3	87.3 / 56	90.5 / 62	88.9	30
Qwen 3.7 Max	95.2 / 52	95.2 / 70	95.2	25
Claude Sonnet 5	92.1 / 63	92.1 / 38	92.1	53
Gemini 3.5 Flash	98.4 / 68	95.2 / 95	96.8	3
Kimi 2.7	82.5 / 42	95.2 / 59	88.8	48
Step 3.7 Flash	87.3 / 74	81.0 / 66	84.2	4
MiniMax M3	71.4 / 53	87.3 / 63	79.3	26
MiMo V2.5	77.8 / 97	95.2 / 47	86.5	31
LongCat 2.0	60.3 / 190	74.6 / 181	67.4	0

‡ Gemini 3.6 Flash — new flash entrant added to Run #33 after the original sixteen-model board; not part of the Run #20 field (§3, §4.4).

Run #33 · OTC Desk Agent Arena · generated 2026-07-20; Gemini 3.6 Flash added 2026-07-23.