🏆 OTC Desk Agent Arena — Run #129

The flash tier, re-measured on a repaired harness — twice. Seven models × five golden workflows × two trials, once at each model's measured effort ceiling and once with the whole field pinned to `low`.

*2026-08-27 · ceiling board = merged run `#129` (`#127` six models + `#128` Qwen 3.8 27B); `low` board = run `#130` · 7 models × 5 workflows × 2 trials × 2 efforts = 140 model-trials · objective-only scoring (jury off) · 70 of 70 pairs `scored` across both boards, none `invalid`, zero malformed tool calls, zero truncated calls, every trial measured rather than assumed***

Why this board exists

Every previous flash-tier board was measured on a harness with defects that were invisible from inside it. Four were found and fixed between runs #114 and #126, and each one moved scores without touching a model:

1. **Anthropic-protocol contestants ran capped at 4096 output tokens** — silently, for every board through #114.
2. **Every ZenMux model id was unpinned**, so each request drew an upstream provider at random and the response named none of them.
3. **A truncated turn and a malformed tool call were both unmeasurable**, so either could reach a published board as a legitimate score.
4. **The trap step had been dead since run #12**, and a leaked scenario set had turned another step into a retrieval dead-end.

The consequence is not that old boards were slightly noisy. Take one model id on one workflow — `deepseek-v4-flash` on `risk-manager-control-day`, no code change and no model change anywhere between these dates:

run	date	objective
#112	2026-08-17	**93.6**
#121 / #122 / #125	2026-08-21 → 08-25	**7.7** ×3
#126	2026-08-25 *(upstream pinned)*	**94.9**

7.7 is not a bad score, it is the *prohibition floor* — what a transcript earns for doing nothing at all, by satisfying every `tool_not_called` check through inaction. The gateway had begun routing that id to a different upstream, which returned tool calls with empty identifiers so nothing ever dispatched. A leaderboard cannot be read at all under those conditions, so the field was re-measured from scratch on the repaired harness.

Headline

rk	Model	eff	OVR	GRD	ADH	SYN	PRC	EFF	CON	Objective
1	**GPT-5.6 Luna**	max	**84**	91	94	95	90	**47**	89	92.6
2	**Qwen 3.8 27B**	max	**84**	91	86	99	89	**58**	87	89.6
3	**Gemini 3.7 Flash**	xhigh	**83**	95	92	99	92	**31**	89	**94.3**
4	**DeepSeek V4 Flash** *(pinned)*	max	**81**	91	91	97	88	**54**	69	90.6
5	**Hunyuan 3**	xhigh	**80**	88	93	77	82	**78**	50	86.1
6	**MiniMax M3**	max	**72**	79	86	79	82	**55**	58	83.0
7	**MiMo V2.5**	high	**70**	77	80	81	82	**63**	52	80.0

`OVR = round(0.32·GRD + 0.26·ADH + 0.16·SYN + 0.16·EFF + 0.10·PRC)`. The jury is off, so no subjective score enters the ranking.

The board's best operator on raw correctness does not win, and neither does the second-best. Gemini 3.7 Flash leads the objective axis by 1.7 points and places third. Qwen 3.8 27B is **fourth on objective (89.6) and second on OVR**, on efficiency alone. Luna takes the top spot from Qwen on the ADH tie-break, both being OVR 84.

*Scored on `par_tool_calls` calibrated for `risk-limit-breach-day` (25) and `ops-settlement-day` (30) — see Calibrating EFF. Before that calibration this same board read Luna 81 / DeepSeek 81 / Gemini 80, i.e. DeepSeek 2nd and Gemini 3rd. Objective scores are untouched; only EFF was re-derived.***

What changed in the harness

Each item below is a measurement, not a refactor. The evidence for each is what justifies discarding the boards that preceded it.

1. The Anthropic protocol had a hidden 4096-token output cap

`langchain_anthropic` applies `_FALLBACK_MAX_OUTPUT_TOKENS = 4096` whenever it has no profile for a model id — and it has none for **any** id routed through ZenMux, because the vendor prefix defeats its lookup. Every Anthropic-protocol contestant in every board through #114 therefore ran capped, while every OpenAI-protocol contestant ran at its provider default. `build_agent_model` now passes an explicit 32768.

Why it was invisible: a truncated turn emits a lone `reasoning` block — no text, no tool call — so the step produces nothing, yet the span is `status=success` because the HTTP call genuinely succeeded. `_is_infra_blank` corroborates blankness with step *errors*, and a truncation raises none, so the match is recorded `scored`.

Measured cost (runs #118 vs #119, budget as the only variable): an artifact was produced in **0 of 8** trials at 4096 and **7 of 8** at 32768, a **16.4-point** mean objective difference.

In this board, **MiniMax M3 is the only Anthropic-protocol contestant** — and it is the model whose historical mean moves most.

2. ZenMux upstream providers were a per-request lottery

ZenMux serves one model id from several upstreams and picks one **per request**; the response body names none of them. Measured on the streaming wire, one `task` call each:

route	tool-call continuation deltas
`deepseek/deepseek-v4-flash` on **deepseek**	`id` / `name` = `null` — correct (30 of 31)
`deepseek/deepseek-v4-flash` on **alibaba**	`id` / `name` = `""` (9 of 10)

langchain's accumulator treats a present-but-empty string as an update and overwrites the real identifiers from the first delta, while `arguments` fragments concatenate correctly — **intact args, hollow ids**. `task()` rejects every such call, the agent re-issues, and it loops to the recursion limit.

The full damage record for one model id, with **no code change anywhere**:

run	date	route	objective
#112	08-17	unpinned → landed on deepseek	93.6
#121	08-21	unpinned → landed on alibaba	**7.7**
#122	08-24	unpinned → landed on alibaba	**7.7**
#125	08-25	unpinned → landed on alibaba	**7.7**
#126	08-25	**pinned to deepseek**	**94.9**

7.7 is the prohibition floor, which inaction earns by satisfying every `tool_not_called` check — so it does not read as a failure, it reads as a model that declined to act.

Every ZenMux model now declares its upstream. Policy: pin the model owner's own infrastructure. The pinned DeepSeek route ranks as its **own contestant** (`deepseek-v4-flash-ds`), because the old slug holds both 93.6 and 7.7 and no single row can honestly average them. Exposure was worst where nobody looked — `deepseek-v4-pro` and `glm-5.2` had six upstreams each.

Alongside the pinning, the registry schema was rewritten so the three routing axes are independent and legible: `id` (the model), `provider` (the upstream serving it), `protocol` (the wire format). Previously `provider` held ZenMux's gateway routing label while the upstream rode inside the id, so one row could read `qwen/qwen3.7-max:alibaba` / `openai` / `anthropic` — three vendor-shaped words meaning three different things. That change is legibility, not behaviour; it is listed here because it is what made the qwen drift below findable.

3. Two silent-failure detectors were added

Both mirror the same design and both are **caveats on a measurement, never automatic invalidation** — the harness cannot know whether a lost turn was scoring-critical, and sweeping such matches to `invalid` would silently shrink historical boards.

- **Truncation** — reads `stop_reason: "max_tokens"` (Anthropic) *and* `finish_reason: "length"` (OpenAI). A detector that knows only one goes blind the moment the other protocol's budget is pinned. The expensive case is `severed_tool_call`: the model had chosen its tool and the cap cut the JSON argument, so the call never ran.
- **Malformed tool calls** — counts calls returned with an empty `id` and/or `name`. This is the defect in §2, and it was invisible to every other gate: nothing raises, so no step records an error; nothing truncates, so the truncation flag reads a clean zero; and the resulting 7.7 is not zero.

Three absence rules apply to both: never appended to a step's `errors` (that list feeds the infra-blank gate, so putting it there would *be* the invalidation this design rejects); `null` means **never measured** and is not `calls: 0`; and both are carried at the breakdown's top level, because `fold_trial_breakdowns` does not lift `diagnosis`.

This board reads a measured zero on both, for all seven models and all 35 matches (`matches_measured` 5/5 per model). That is the first board that can say so — and `null` would not have been good enough, because `null` means never measured, not zero.

4. Trap and fixture repairs

- The step-8 prohibition had been **dead at 5/99 since run #12**. Honest survivors were probing the exact requested set name, taking the "not found" error and abstaining — behaviour the bare check scored identically to silently substituting a different scenario. The check now exempts a *failed* probe of the exact referent (a probe is only a probe if it failed; otherwise `generate_scenario_set` minting the set under the requested name would walk straight through), and pairs the prohibition with a neutrally-worded `record_answer` null.
- A leaked near-miss scenario set (`stagflation-shock-2011-compact`) had sat beside the reserved name for five weeks — a reserved-name guard cannot see a near-miss, and a model asked for a set that does not exist **invents one**. It turned the trap into a 141 KB retrieval problem: 273 calls and 91 errors in that step alone on run #109. `_purge_match_scenario_sets` now reclaims model-created sets on trace + baseline evidence.

5. The contestant key was widened

A contestant is now `(model_id, reasoning_effort, max_output_tokens)`. Effort moves tool-call count by ~22% between `low` and `high`, so it moves EFF and CON; budget gates performance as a cliff. Two efforts are two operating regimes, never one averaged row. Effort ladders are **measured per route** and vendored, because models.dev was wrong for every route that mattered and mostly **too narrow** — the dangerous direction.

Why the historical means cannot be compared

Per-model objective means, split at the boundaries above:

model	protocol	⩽ #114	#115—#126	#127
MiniMax M3	anthropic	**48.2** *(n=13)*	87.2 *(n=3)*	**83.0** *(n=5)*
DeepSeek V4 Flash *(unpinned)*	openai_chat	81.2 *(n=30)*	**50.2** *(n=6)*	—
DeepSeek V4 Flash *(pinned)*	openai_chat	—	—	**90.6** *(n=5)*
GPT-5.6 Luna	openai_chat	89.2 *(n=38)*	91.6 *(n=6)*	92.6 *(n=5)*
Gemini 3.7 Flash	openai_chat	97.8 *(n=5)*	92.3 *(n=2)*	94.3 *(n=5)*
Hunyuan 3	openai_chat	75.0 *(n=10)*	78.2 *(n=2)*	86.1 *(n=5)*
MiMo V2.5	openai_chat	78.4 *(n=19)*	85.3 *(n=2)*	80.0 *(n=5)*

Two rows carry the whole argument:

- **MiniMax M3, the only Anthropic-protocol contestant, moves 48.2 → 83.0.** That is the profile the output-budget fix predicts, and the controlled A/B (#118 vs #119) independently measured a 16.4-point budget effect. It is *consistent with* the cap being lifted; it is not proof, because these era means also span manifest changes and the trap repair.
- **DeepSeek V4 Flash's unpinned slug goes 81.2 → 50.2 while getting better**, because the lottery started landing on the broken upstream. Its pinned successor scores 90.6.

These era means are not clean comparisons — they pool different workflow mixes, manifest versions and field sizes. They are shown to establish that the harness moved scores by more than the models differ from each other, which is the only claim needed to justify a fresh board.

Calibrating EFF

Two of the five workflows had never been calibrated. `risk-limit-breach-day` and `ops-settlement-day` declared no `par_tool_calls`, so `designed_par` fell back to the **theoretical minimum** — the sum of each step's `expected_tools`, 11 and 10 — and both stayed on the legacy hyperbolic EFF curve.

That fallback exists so the shared scoring kernel never regresses a workflow nobody has calibrated yet. It is not a par. **No trial in the arena's history has ever come close to either number**: the leanest limits run ever recorded is 15 calls, the leanest ops run 20. Scoring against a target the design has never observed is not a hard standard, it is a broken instrument — every model on both workflows was being measured against a run that does not exist.

Method — the same one the only real precedent used. Of the three already-calibrated workflows, exactly one has an empirical anchor: `high-board-portfolio-review-day` declares 24, and 24 is the **median of its 22 perfect trials**. The flagship's 24 is *not* a usable precedent (it has only 2 perfect trials ever, 42 and 50 calls, so

its par sits at the 0th percentile — it was derived analytically as 11 expected + ~13 counted overhead), and `trader-rfq-booking-day` has no perfect trials at all.

So: **median of fully-correct trials on the current manifest.**

One correction was needed before that median meant anything. Eight runs in this database are not runs at all but **merges** — `merge_runs` folds its sources' per-trial breakdowns into a new run rather than referencing them, so a naive sweep of `arena_match` counts those trials twice. They also nest: #58 contains #44, which is itself a merge of three earlier runs. Restricted to original runs, each counted once:

workflow	old par	fully-correct trials	models	min	p25	median	p75	max	new par
`risk-limit-breach-day`	11 †	31	13	15	23.0	**25.0**	31.0	85	**25**
`ops-settlement-day`	10 †	13	**3**	21	26.0	**30.0**	35.0	52	**30**

† theoretical-minimum fallback, not a declared par.

The de-duplication did not move either median — duplication is symmetric and a median shrugs it off. That is exactly why it was worth catching: the statistic looked fine, and only the sample size gave it away (36 "trials" from a workflow with twelve real runs). **Check `n` against the number of runs that could have produced it, not just the answer.**

Counting is against `counts_detail.tool_calls`, which **excludes** `META_TOOLS` (`task`, `read_file`, `write_todos`) — a skill-file read must never enter a par, or the denominator outgrows the numerator it is compared against.

Both pars survived a hard test. They were first derived before this report's `low` arm existed, on roughly half the sample. Adding 7 trials to one and 2 to the other moved `risk-limit-breach-day` by a single call (26 → 25) and `ops-settlement-day` not at all. A doubled sample and an entirely new effort regime shifting par by one call is the evidence these are anchored rather than lucky.

But par drifts with the effort mix of whoever ran the boards, and that is a real limitation of the method rather than of these two numbers. On `risk-limit-breach-day` the `low` trials sit at median 23 and the `max` trials at 27; a low-heavy sample therefore tightens EFF for every model, permanently. 25 leans low deliberately — the desk default is `low`, and an **inflated par hands back free EFF credit**, which corrupts a ranking, whereas a par slightly too strict only compresses one axis. `ops-settlement-day` shows no such skew (low 30.0 against max 29.5).

`ops-settlement-day`'s par is still **provisional, and on breadth rather than depth**. Its 13 trials come from only **three models**. `high-board-portfolio-review-day`'s 24 rests on 22 trials and `risk-limit-breach-day`'s 25 on 31 trials from thirteen models. A par set by three models can encode their shared habits as the standard; the manifest flags it for re-derivation when a fourth and fifth model finally post a fully-correct trial there.

This re-scores history, and two boards go DOWN

Ability cards are **derived on read**, never migrated — so changing a par re-scores every stored match on that workflow, retroactively. The blast radius is **19 runs, 103 matches, 191 trials**. Note this tally deliberately *keeps* merged runs: they are themselves published boards whose numbers shift. Only the par derivation excludes them. Same table, opposite rule, because one question is "what does a competent run cost" and the other is "what on the site changes".

	EFF	OVR
mean, all 191 affected trials	37.2 → **78.0** (+40.8)	84.4 → **90.8** (+6.5)

Most of that is the broken instrument being repaired. But **runs #103 and #114 lose ground** (−13.0 and −8.0 EFF), and that is the calibration working as designed: the hyperbolic curve `min(1, par/calls)` never reaches zero, so it always granted partial credit to a runaway. Golf scoring hits **exactly 0 at 2 × par**. An 85-call limits trial scored 13 on the old curve and scores **0** now.

Four published boards are affected — **#101** (a ranked leaderboard board), and the provisional cards **#104**, **#113** and **#115**. The site derives every number from the database at build time, so these move on the next build whether or not anyone says so. They should carry a rescoring note, exactly as #104 already carries one for the 39 → 38-point manifest change.

And #101's new top is not a ranking anyone should trust

Calibrating EFF does not just move that board, it **saturates** it:

rk	model	OVR	GRD	ADH	SYN	EFF	PRC	objective
1	GPT-5.6 Terra	98	99	99	99	99	93	97.4
2	Kimi 2.7	98	99	99	99	99	87	94.9
3	**Grok 4.5**	98	99	99	99	97	99	**100.0**
4	MiMo 2.5 Pro	98	99	99	99	97	90	96.2

Four models tie at OVR 98 on **identical** grounding, adherence and synthesis. The order is settled by EFF and then PRC — which places **Grok 4.5 third with a perfect 100.0 objective score**, behind models scoring 97.4 and 94.9, for spending two more tool calls. The top slot is not even robust to the constant: at par 26 it is Grok 4.5, at par 25 it is GPT-5.6 Terra. One call of calibration flips the podium.

That is not a scoring defect. It is `risk-limit-breach-day` reporting that it can no longer separate a strong field — **34 of its 36 checks pass for every contestant** — and it is the same finding this board makes about the flash tier, one layer down. Once correctness saturates, EFF *is* the ranking. Publishing #101 as an ordered leaderboard would present a tie-break artefact as a result.

The result: efficiency is the entire ranking

The correctness axes are saturated. Across seven models: GRD 77–95 (span 18), ADH 80–94 (14), SYN 77–99 (22), PRC 82–92 (10). **EFF spans 31 to 78 — a span of 47**, twice the widest correctness axis, and it stays that way *after* a calibration that raised almost every model's EFF.

Every model is at its own measured ceiling, and every model pays for it. Tool calls per trial, against each workflow's calibrated par:

workflow	par	Luna	Qwen	Gemini	DeepSeek	Hunyuan	MiniMax	MiMo
`risk-manager-control-day`	24	45 / 43	39 / 28	60 / 57	31 / 53	40 / **22**	89 / 41	**24** / 30
`high-board-portfolio-review-day`	24	**23** / 35	**17** / 30	54 / 75	37 / 30	**14** / **17**	43 / **18**	**19** / **13**
`trader-rfq-booking-day`	35	141 / 95	67 / 65	64 / 75	40 / 51	44 / **34**	45 / 57	62 / 54
`risk-limit-breach-day`	25	32 / 35	33 / **24**	30 / 28	48 / **23**	**15** / **23**	27 / **24**	**24** / 31
`ops-settlement-day`	30	**24** / 32	31 / 38	38 / 39	35 / 38	**25** ‡	**24** / **20**	**22** / 31

Bold = at or under par. ‡ one trial only — see caveats. All five pars are now calibrated, so every EFF input on this board comes from the same golf curve.

EFF is golf-scored on a calibrated par: full marks at or under par, then linear decay to zero at **2 × par**.

- **Gemini 3.7 Flash (EFF 31)** is the **only contestant with no under-par trial at all** — over par on every one of its ten trials across all five workflows, 1.8–3.1× par on the three long-calibrated ones, and past the 2× cliff on five of ten. It is the most correct agent on the board and the most expensive one to run.
- **GPT-5.6 Luna (EFF 47)** is efficient everywhere except `trader-rfq-booking-day`, where it burned **141 and 95** calls against a par of 35 — 4.0× and 2.7× par, a guaranteed EFF 0 on both trials, and it still only scored 89.7. One workflow destroys its efficiency stat.
- **Qwen 3.8 27B (EFF 58)** is the efficiency story of the board: **fourth on objective and second on OVR**. Its worst trial anywhere is **1.91× par**, so it never reaches the cliff on any workflow — and it is the only contestant to score a **perfect 38/38** on one.
- **Hunyuan 3 (EFF 78)** comes in under par on **seven of its nine trials**, the most on the board, and never exceeds 1.67× par. But it is cheap because it did less: 61.5 on `high-board` and 86.1 overall, both near the bottom of the field.

Three models — Qwen, Hunyuan and MiMo — stay off the 2× cliff entirely. The three that do not are Luna (twice, both on trader-rfq), DeepSeek (once) and MiniMax (once), plus Gemini's five.

A ceiling board measures capability, not the configuration you should ship. Run

110 measured this on one model: effort is a step at `low`, and `high` / `max` buy about

one point for 3.3× the time and +50% the calls. Run #127 is that result across a field: at their ceilings, every model pays the tax and EFF collapses for all of them.

The same field at `low` — run #130

A ceiling board answers "how good can each model get". It cannot answer "what should the desk ship", because every model is at a different point on its own ladder — `max` for four of these seven, `xhigh` for two, `high` for one. So the field was run again, all seven pinned to `low`, everything else identical: same five workflows, same two trials, same unpinned output budget, same commit. Each model is now its own control and effort is the only variable.

The board reorders.

	ceiling		`low`	
1	GPT-5.6 Luna	84	**Gemini 3.7 Flash**	**86**
2	Qwen 3.8 27B	84	DeepSeek V4 Flash	84
3	Gemini 3.7 Flash	83	Qwen 3.8 27B	81
4	DeepSeek V4 Flash	81	Hunyuan 3	79
5	Hunyuan 3	80	GPT-5.6 Luna ‡	78
6	MiniMax M3	72	MiniMax M3	77
7	MiMo V2.5	70	MiMo V2.5	76

Gemini 3.7 Flash goes from third to first, and leads on *both* axes — 86 OVR and 92.0 objective, the best of either board on correctness-per-call. Its EFF more than doubles, 31 → 76, because it drops 20.7 tool calls per trial. It gives up 2.3 objective points to do it. The board's most expensive agent at its ceiling is its best agent at `low`.

Four of seven models score *higher* on OVR at `low`. Only Qwen (−3) and Hunyuan (−1) lose ground, plus Luna, whose result needs the caveat below.

With n = 2, one bad trial is worth 30 points — and CON is the tell

Before reading any delta, the deltas have to be filtered. A cell here is the mean of two trials, so a single blown trial moves it by more than any real effect. CON measures exactly that dispersion, and **CON near zero invalidates the cell's mean**:

cell	ceiling trials	`low` trials	raw Δobj	reality
Luna / `risk-manager-control-day`	37, 38 of 39	**13**, 38 of 39	−30.9	one blown **low** trial (CON 0)
MiMo / `risk-limit-breach-day`	38, **19** of 38	37, 37	+22.4	one blown **ceiling** trial (CON 0)
Qwen / `high-board-portfolio-review-day`	30, 30 of 35	17, 19 of 35	−34.2	**real** — consistent on both sides

The two largest deltas on the board are artefacts pointing in opposite directions, and the third-largest is genuine. Nothing but the per-trial spread distinguishes them. Requiring CON ≥ 50 on **both** arms leaves **25 of 35 cells usable**:

model	usable cells	ceiling	`low`	Δ objective
GPT-5.6 Luna	4/5	91.7	91.8	**+0.1**
MiniMax M3	4/5	82.9	83.0	**+0.2**
DeepSeek V4 Flash	4/5	88.6	88.1	**−0.5**
Gemini 3.7 Flash	5/5	94.3	92.0	**−2.3**
Hunyuan 3	1/5 †	89.7	92.1	+2.4
Qwen 3.8 27B	5/5	89.6	82.7	**−6.9**
MiMo V2.5	2/5 †	71.8	64.5	−7.3

† too few clean cells to support a claim.

On clean evidence, `low` costs the field a mean 2.4 objective points — and for the three models with four or more usable cells and no collapse, it costs **nothing measurable**: Luna +0.1, MiniMax +0.2, DeepSeek −0.5. Luna's fifth-place finish on the measured board is an artefact of that one 13/39 trial; on its four clean cells it is **unchanged** between `max` and `low`.

Two models genuinely pay. **Qwen 3.8 27B loses 6.9 points**, almost all of it one collapse: on `high-board-portfolio-review-day` it drops 85.7 → 51.5, consistently across both trials (17 and 19 of 35, against 30 and 30), while cutting calls only 9. It is not trading accuracy for speed there; it stops doing the work. **MiMo V2.5 loses 7.3**, but on two clean cells only.

The effect is per-model, not per-workflow

Run #110 established that effort's sign flips across *tasks*. This board shows it also flips across *models on the same task*. On `high-board-portfolio-review-day`, where six of seven cells are usable: Gemini −7.1, Qwen −34.2 and MiMo −11.3 against DeepSeek +2.9 and MiniMax +4.3. Three models get materially worse at `low` and two get better, on one workflow. (Hunyuan's +5.7 there is unusable — CON 20 and 46.)

And OVR moves opposite to correctness. Gemini loses 7.1 objective points on that workflow and *gains* 7 OVR; MiniMax gains 4.3 and gains 9. EFF is weighted only 0.16 against correctness's 0.74 — but the correctness axes are saturated (spans 10 to 22) while EFF spans 47, so the heavy low-variance axes contribute almost nothing to the *differences between rows* and the light high-variance axis decides them. **A weight is only as influential as the spread of what it multiplies.** Read OVR as "capability per call", never as capability.

One model moves the other way: **Hunyuan 3 is the only contestant whose EFF *falls* at `low`** (78 → 63), because its call count *rises* 5.5 per trial. Lower effort does not imply less work.

What this means for the desk

Ship `low`. On this evidence the correctness cost is under a point for most of the field, the call-count saving is real (Gemini −20.7, Luna −18.6, Qwen −11.4 per trial), and the ranking a desk should act on is the `low` one — because that is the configuration it would actually run. The exceptions are specific and worth testing per model rather than assuming: Qwen 3.8 27B on portfolio-review work, where `low` halves its score, and any workflow where a model's CON is already poor.

The honest limitation is **n = 2**. Ten of 35 cells were unusable, and the filter that caught them is the same statistic (CON) that a single-trial board cannot compute at all. A three-trial board would cost 50% more and would have made every cell here readable.

Discrimination is concentrating in one workflow

Objective score by workflow:

workflow	Luna	Qwen	Gemini	DeepSeek	Hunyuan	MiniMax	MiMo	spread
`high-board-portfolio-review-day`	78.5	85.7	**95.7**	72.8	61.5	57.1	64.2	**38.6**
`risk-limit-breach-day`	98.7	**100.0**	98.7	98.7	93.4	97.3	75.0	25.0
`ops-settlement-day`	**100.0**	84.1	89.8	93.2	90.9	89.8	79.5	20.5
`risk-manager-control-day`	**96.2**	91.0	93.6	94.8	94.9	83.3	89.8	12.9
`trader-rfq-booking-day`	89.7	87.3	**93.7**	**93.7**	89.7	87.3	91.3	**6.4**

`trader-rfq-booking-day` puts five of seven models within 4.0 points. `risk-limit-breach-day` puts three models on the *identical* 98.7 and hands Qwen a perfect 38/38 — the near-dead signature the run #113 validity audit flagged (34 of 36 checks scoring N/N carry no ability signal while still occupying the denominator).

`high-board-portfolio-review-day` is doing most of the discriminating, and inside it Gemini 3.7 Flash's 95.7 is **10.0 points clear of second place** while three of seven models sit in the 57–65 band. (Qwen 3.8 27B halved that gap: before it joined, Gemini led this workflow by 17.2.) A single workflow separating the field by that margin deserves a per-check pass-rate tally before it anchors any conclusion — the same instrument that found 15 of 50 checks non-discriminating on run #58 and reordered that board.

Caveats

- **n = 2 per cell.** CON (trial consistency) is 50–89; the low-CON models — Hunyuan 3 at 50, MiMo V2.5 at 52, MiniMax M3 at 58, DeepSeek V4 Flash at 69 — have real trial-to-trial variance, so their means are the least stable numbers here. MiniMax's flagship trials burned 89 and 41 calls; Hunyuan's, 40 and 22.
- **`hunyuan-3` / `ops-settlement-day` rests on ONE trial.** The second was dropped as an infra trial (infra trials are skipped, not retried). Its 90.9 has no trial-to-trial corroboration.
- **Both boards hit a network outage, and they failed differently.** The ceiling board's ~1h24m outage (09:39–11:03 UTC) made calls *block* rather than fail fast, so the loop sat still and recorded nothing wrong; the pair straddling it returned 89.8, in line with the field. The `low` board's outage failed fast instead — ~234 `ConnectError`s swept **four `risk-manager-control-day` arms** to `invalid` / `infra_blank`. Those four were re-run to completion with `--resume 130`, which deletes the stale row first, so all 35 are `scored` at the same effort and budget. Neither board recorded an unscored or contaminated pair.
- **`gpt-5-6-luna` on `risk-manager-control-day` at `low` is the one cell to distrust.** Its trials scored 13/39 and 38/39 (CON 0). The 38 matches its ceiling arm exactly; the 13 is an outlier of unknown cause, and it is what puts Luna fifth on the `low` board. It is one of the four arms re-run after the outage, though the re-run itself completed cleanly and the other three show nothing unusual.
- **`ops-settlement-day`'s par rests on 13 fully-correct trials from only THREE models.** Its weakness is breadth, not depth: a par set by three models can encode their shared habits as the standard. It is the least-settled number on this board and is flagged in the manifest for re-derivation once a fourth and fifth model post a fully-correct trial there.
- **The par calibration re-scored 191 trials across 19 runs**, four of them published. Their EFF and OVR move on the next site build, and run #101's podium reorders. See Calibrating EFF.
- **This is a MERGED board.** Qwen 3.8 27B ran as its own run, in its own database and process, and was folded in through `store.merge_runs` (`merged_from: [127, 128]`). That is the sanctioned path — the 17-model run #58 board was built the same way — and the conditions match on every axis that scoring depends on: same commit, byte-identical manifests, same denominators (39·63·35·38·44), ceiling effort, unpinned budget, two trials, and both flags **measured** rather than null. The two runs also overlapped in wall-clock time (06:25–12:20 vs 07:30–23:45 UTC), so they share gateway conditions and the same network outage rather than being weeks apart. What does *not* match: Qwen's transcripts were produced by a different process against a different SQLite file. Merged rows carry `transcript_path=None` by design, so the drilldown reaches Qwen's transcripts through source run **#128** and everyone else's through **#127**.
- **Qwen 3.8 27B is a 27B model in a flash-tier field.** It belongs on cost grounds, but it is not the same kind of artefact as the others — read its rank as "this is what a small dense model buys you here", not as a like-for-like tier comparison.
- **Not comparable to runs #8–#126** for any of these models, for every reason in this report. That is the point of the board, not a limitation of it.
- **The jury is off** (default since run #11), so this is an objective-only ranking.

What this board does and does not say

Does: on a repaired harness, at each model's measured effort ceiling, across five golden desk workflows, these seven flash-tier models are separated by 14 OVR points (84 → 70) and by 14.3 objective points (94.3 → 80.0). Correctness is broadly solved in this tier; cost is not — and the top **four** are within 3 OVR points of each other.

Also does: at `low` — the setting a desk would actually ship — the board reorders, **Gemini 3.7 Flash leads on both axes**, and the correctness cost of dropping from each model's ceiling is **under one point for three of the four models with enough clean cells to measure**. The exception is specific and reproducible: Qwen 3.8 27B halves its score on portfolio-review work.

Does not: it does not establish the ceiling ordering as settled — 1st and 2nd there differ by an ADH tie-break, and the `low` board disagrees with it. It does not measure any intermediate effort. And with n = 2 it cannot read 10 of its 35 paired cells at all, because one blown trial moves a cell further than any real effect does.

Follow-ups, in order: 1. A per-check pass-rate tally on `high-board-portfolio-review-day`, before its 38.6-point spread is treated as ability signal — it is now carrying most of the board's discrimination *and* Qwen's entire `low` regression. 2. Add a rescoring note to published boards #101, #104, #113 and #115 before the next site build moves their numbers silently. Run #101's podium reorders and its top four tie at OVR 98. 3. **Run three trials, not two, on the next board.** Ten of 35 paired cells here were unreadable because a single trial blew up, and the two largest deltas on the board were artefacts pointing in opposite directions. A third trial costs 50% more and would have made every cell readable. 4. Investigate Qwen 3.8 27B on `high-board-portfolio-review-day` at `low`: 17 and 19 of 35 against 30 and 30 at `max`, on 9 *fewer* calls. It is not trading accuracy for speed; it stops working.
