🏆 OTC Desk Agent Arena — Run #131

Two new flash-tier models arrive, post the best correctness on the board twice, and still lose. GLM 5.3 Flash and Qwen 3.8 Flash across five golden workflows at each model's effort ceiling.

*2026-08-28 · run `#131` · 2 models × 5 workflows × 1 trial = 10 model-trials · both pinned to `max` · objective-only scoring (jury off) · 10 of 10 pairs `scored`, none `invalid`, zero malformed tool calls, zero truncated calls · compared against the cards of run `#129`, which was measured at 2 trials per cell**

Why this run exists

Both models were published in the two days before it ran — GLM 5.3 Flash on 2026-08-26, Qwen 3.8 Flash on 2026-08-27 — and both are aggressively cheap for the tier: `$0.075 / $0.25` and `$0.16 / $0.47` per MTok, against `$1.40 / $4.40` for GLM 5.3 and `$0.50 / $3.00` for Qwen 3.8 27B. Both carry 1M context and accept image and video input.

The question was whether the flash tier had just gained two contestants at a fraction of the price. On correctness, very nearly. On the ranking, no — and the reason is a single axis.

Headline

Consolidated cards, averaged across all five workflows. Rows 1–7 are run #129 at each model's own effort ceiling; the two newcomers are run #131 at `max`.

Model	eff	OVR	GRD	ADH	SYN	PRC	EFF	CON	Objective
GPT–5.6 Luna	max	**84**	91	94	95	90	47	89	92.6
Qwen 3.8 27B	max	**84**	91	86	99	89	58	87	89.6
Gemini 3.7 Flash	xhigh	**83**	95	92	99	92	31	89	**94.3**
DeepSeek V4 Flash *(pinned)*	max	**81**	91	91	97	88	54	69	90.6
Hunyuan 3	xhigh	**80**	88	93	77	82	78	50	86.1
GLM 5.3 Flash	max	**79**	85	92	95	88	**25**	—	88.7
Qwen 3.8 Flash	max	**78**	88	85	99	88	**20**	—	88.2
MiniMax M3	max	**72**	79	86	79	82	55	58	83.0
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. CON reads `—` for the newcomers because run #131 ran one trial per cell and consistency needs trials to disperse — that is *not measured*, not *perfect*.

Read the two halves of those rows separately. On objective score the newcomers land sixth and seventh of nine — but only 1.9 and 2.4 points behind Hunyuan 3, which outranks them, and comfortably ahead of MiniMax M3 and MiMo V2.5, which they lose to on nothing. On EFF they are **first and second worst on the board**, below Gemini 3.7 Flash's 31, which was the previous floor.

Every other axis is competitive. GLM 5.3 Flash's ADH of 92 is third best in the field. Qwen 3.8 Flash's SYN of 99 ties the best score anyone has posted. Neither model has a correctness weakness that explains its position.

The finding: correct, and expensive for it

EFF is golf-scored. A run at or under a workflow's calibrated `par_tool_calls` earns full marks; above par it decays linearly to **zero at 2× par**. So the axis does not ask whether a model got the right answer — the other four axes do that — it asks what the answer cost.

Here is every cell of run #131 against par:

Model	Workflow	par	calls	× par	EFF	OVR	Objective
GLM 5.3 Flash	high–board–portfolio–review	24	26	1.08	64	71	71.4
GLM 5.3 Flash	risk–limit–breach	25	34	1.36	63	93	**100.0**
GLM 5.3 Flash	trader–rfq–booking	35	78	2.23	**0**	81	**95.2**
GLM 5.3 Flash	ops–settlement	30	72	2.40	**0**	69	81.8
GLM 5.3 Flash	risk–manager–control	24	58	2.42	**0**	82	94.9
Qwen 3.8 Flash	ops–settlement	30	42	1.40	45	75	79.5
Qwen 3.8 Flash	risk–limit–breach	25	39	1.56	44	90	**100.0**
Qwen 3.8 Flash	high–board–portfolio–review	24	44	1.83	12	63	71.4
Qwen 3.8 Flash	trader–rfq–booking	35	108	3.09	**0**	81	**95.2**
Qwen 3.8 Flash	risk–manager–control	24	79	3.29	**0**	82	94.9

Five of ten cells score EFF **0**. And the pattern inside that is the uncomfortable part: **the cells where these models are most correct are the cells where they are most wasteful.** GLM 5.3 Flash's three zero-EFF workflows carry objective scores of 95.2, 81.8 and 94.9 — its two best results among them. Qwen 3.8 Flash's two zeroes are its 95.2 and 94.9. Neither model is thrashing because it is lost. It is arriving at the right answer by a long road.

Neither model reached par on any workflow. GLM's leanest run was 1.08× par, Qwen's 1.40×. Five of the seven established contestants come in *under* par on at least one workflow — Hunyuan 3 does it three times, at 0.65×, 0.76× and 0.83×.

Where they actually lost

Two workflows decide it, and neither is a story about getting things wrong.

`trader-rfq-booking-day` — the best correctness on the board, fourth place

Model	OVR	EFF	calls	× par	Objective
DeepSeek V4 Flash	**89**	66	46	1.30	93.7
Hunyuan 3	**88**	82	39	1.11	89.7
MiniMax M3	**81**	50	51	1.46	87.3
GLM 5.3 Flash	**81**	**0**	78	2.23	**95.2**
Qwen 3.8 Flash	**81**	**0**	108	3.09	**95.2**
MiMo V2.5	79	32	58	1.66	91.3
Gemini 3.7 Flash	78	8	70	1.99	93.7
Qwen 3.8 27B	75	10	66	1.89	87.3
GPT-5.6 Luna	73	**0**	118	3.37	89.7

Both newcomers post 95.2 — the highest objective score any model has recorded on this workflow — and finish fourth equal. DeepSeek V4 Flash wins it with 93.7, a point and a half *less* correct, on 46 calls instead of 78 and 108.

`risk-manager-control-day` — identical correctness, 5 OVR apart

Model	OVR	EFF	calls	× par	Objective
Hunyuan 3	**87**	66	31	1.29	**94.9**
Qwen 3.8 27B	**84**	56	34	1.40	91.0
DeepSeek V4 Flash	**82**	35	42	1.75	94.8
GLM 5.3 Flash	**82**	**0**	58	2.42	**94.9**
Qwen 3.8 Flash	**82**	**0**	79	3.29	**94.9**
GPT-5.6 Luna	81	16	44	1.83	96.2

Hunyuan 3, GLM 5.3 Flash and Qwen 3.8 Flash **all score exactly 94.9** on the flagship — the same 37 of 39 checks. Hunyuan does it in 31 tool calls and takes the workflow. GLM needs 58, Qwen needs 79, and both drop five OVR points for the difference. Qwen 3.8 Flash spends **2.5× Hunyuan's calls to produce a byte-identical result.**

Against their own larger siblings

The board comparison asks how these models rank. The more practical question is whether the flash variant is worth taking over the bigger model it is named after — and both pairs answer it the same way.

Neither comparison could be read off the existing data. Qwen 3.8 27B had a clean counterpart in run #129 (`max`, repaired harness, one day earlier). GLM 5.3 did not: its only full board is run #115 from 2026-08-19, at **unpinned vendor-default effort** and from before the upstream pinning and the trap fix. Differencing that against run

131 would have charged the model for a harness repair and an effort change at once.

GLM 5.3 was therefore re-run — **run #132**, `max`, same five workflows, same single trial, same unpinned budget — so the pair is like-for-like.

That re-run was worth doing for its own sake. At `max` on today's harness GLM 5.3 completes `high-board-portfolio-review-day` in **17** tool calls; the same model at vendor-default effort in run #115 took **41**. Published as a sibling delta, that artifact would have been attributed to the flash model.

Qwen 3.8 Flash vs Qwen 3.8 27B

Both at `max`. 27B is run #129 at two trials per cell, Flash is run #131 at one.

Workflow	27B OVR	Flash OVR	27B obj	Flash obj	27B calls	Flash calls
high-board-portfolio-review	84	63	85.7	71.4	24	44
ops-settlement	82	75	84.1	79.5	34	42
risk-limit-breach	94	90	100.0	100.0	28	39
risk-manager-control	84	82	91.0	**94.9**	34	79
trader-rfq-booking	75	**81**	87.3	**95.2**	66	108
mean	**83.8**	**78.2**	**89.6**	**88.2**	**37.2**	**62.4**

1.4 objective points separate them. Qwen 3.8 Flash costs roughly a sixth of the 27B on completion (`$0.47` against `$3.00` per MTok) and is, on average, very nearly as correct — and on two of the five workflows it is *more* correct than its larger sibling, beating it by 3.9 points on the flagship and 7.9 on trader-rfq.

The 5.6-point OVR gap is **EFF: 58.4 against 20.2**, on 68% more tool calls. The one real capability gap is `high-board-portfolio-review-day`, where the flash model loses 14.3 objective points and nearly doubles the calls.

GLM 5.3 Flash vs GLM 5.3

Both at `max`, both one trial per cell, run #132 against run #131. Run #132's flagship arm died twice on transport errors and was recovered with `--resume`; it is scored from the third, clean attempt.

Workflow	5.3 OVR	Flash OVR	5.3 obj	Flash obj	5.3 calls	Flash calls
high-board-portfolio-review	72	71	71.4	71.4	17	26
ops-settlement	88	69	88.6	81.8	32	72
risk-limit-breach	99	93	100.0	100.0	24	34
risk-manager-control	83	82	97.4	94.9	54	58
trader-rfq-booking	93	81	96.8	95.2	44	78
mean	**87.0**	**79.2**	**90.8**	**88.7**	**34.2**	**53.6**

2.2 objective points separate them, against a 7.8-point OVR gap — and again the gap is EFF, 64.4 against 25.4, on 57% more tool calls. The two models post *identical* objective scores on two of the five workflows (71.4 and 100.0). GLM 5.3 Flash is priced at about a fifth of GLM 5.3 (`$0.075 / $0.25` against `$1.40 / $4.40` per MTok).

The flagship is the exception, and it is informative. There GLM 5.3 spends **54** tool calls against its flash sibling's 58, and *both* score EFF **0** — the pro model is just as profligate as the cheap one. So `risk-manager-control-day`'s par of 24 punishes the whole GLM 5.3 generation rather than the flash variant in particular, and the flash model's efficiency problem, real as it is elsewhere, is not a uniform trait of its own lineage.

One incidental finding: **GLM 5.3 and GLM 5.3 Flash share the same unusual effort ladder** — `low`, `high`, `max`, with `medium` and `xhigh` both rejected. That is a property of the 5.3

generation, not something the flash variant introduced, and the upstream states it only in the untranslated tail of its error message.

The same shape, twice

Two independent model families, one conclusion: **the flash variant gives up around two objective points — 1.4 for Qwen, 2.2 for GLM — and spends 57 to 68% more tool calls to do it.** Whether that is a good trade is not a question the arena can answer, because it depends on what you pay for. Priced per token, both flash models are the better buy by a wide margin. Priced per tool call, per second of wall-clock, or per unit of rate limit, both are worse than the model they are a cheaper version of — and the OVR ranking, which weights EFF at 0.16, is measuring the second thing.

What this is not

It is not a correctness problem. On `risk-limit-breach-day` both models score a clean **100.0**, and on the flagship and trader-rfq they equal or beat everyone. Their GRD and SYN are top-half. If the ranking weighted only the four correctness axes, GLM 5.3 Flash would sit fourth and Qwen 3.8 Flash fifth rather than sixth and seventh.

It is not a routing or protocol defect. Both models emitted **zero malformed tool calls** and **zero truncated calls** across all ten matches, and those are measured zeros (`5/5 matches measured`), not unmeasured nulls. This mattered more than it sounds: Qwen 3.8 Flash's `alibaba` upstream returns tool-call continuation deltas with an **empty `id`** — 263 of 263 in a direct SSE probe — which LangChain merges over the real identifier, so nothing dispatches and the match lands on the 7.7 prohibition floor. Both models are pinned to the Anthropic wire protocol for exactly that reason. Unpinned, this article would have been about two models that appear unable to use tools at all.

It is not uniform. GLM 5.3 Flash ran `high-board-portfolio-review-day` at 1.08× par for EFF 64 — lean by any standard. Qwen 3.8 Flash's best is 1.40×. Both models *can* be economical; they are not on the workflows that matter most.

It is not unique to them. GPT-5.6 Luna — the top of the whole board — posts EFF 0 on trader-rfq at 3.37× par, worse than either newcomer. The difference is that Luna is frugal elsewhere (0.93× par on ops-settlement, EFF 96), so one expensive workflow does not define its card. Neither newcomer has that offsetting cell.

Caveats

Run #131 ran one trial per cell; run #129 ran two. This is the largest limitation here. The newcomers' cards therefore have no CON, and every figure is a single sample where the comparison field is a two-trial mean. A second trial could move an individual cell meaningfully. It is unlikely to erase the finding — the EFF gap is 20–25 against a 31–78 field, and five of ten cells are at *zero*, which needs 2× par to reach — but a depth-matched re-run is the honest way to settle it, and it has not been done.

These are two runs, not one board. #129 and #131 executed on different days under different network conditions. This page compares **cards**, which are absolute measurements — `passed/total` per axis, EFF against each workflow's own par — and stay meaningful without a shared field. It does not publish a merged ranking, and the newcomers do not appear on the leaderboard.

EFF depends on par, and par is young. `risk-limit-breach-day` (25) and `ops-settlement-day` (30) were only calibrated on 2026-08-27, one day before this run. Since EFF is the entire story here, a future par revision moves these two cards more than anyone else's.

One lost point is the harness's fault, not the models'. Both failed `s1:skill: read-risk-result` on the flagship. That skill declares no `routing:` frontmatter, so it never enters the orchestrator's known-skills table and cannot be routed to — `risk-limit-breach-day` already handles this by declining to grade it. The flagship still does, so every contestant on that workflow loses the same point to a check that measures catalogue spelunking rather than ability.

What would change the verdict

A depth-matched re-run at two trials, and an arm at `low`. Run #130 showed that four of seven models scored **higher** at `low` than at their ceiling, precisely because lower effort cuts tool calls and EFF has the widest spread of any axis. Two models whose only weakness is call volume are the most likely candidates in the field to gain from being turned down — and at `$0.075` per MTok in, the experiment is close to free.

Cards for run #131 are published on the model cards page. It is not a ranked board: a two-model run across five workflows has no field to rank within, and folding five workflows into one row would publish a cross-workflow average under a single workflow's heading.

Rendered from `docs/arena/2026-08-28-run131-flash-newcomers.md` · OTC Desk Agent Arena · Run #131.