🏆 OTC Desk Agent Arena — Run #134

Gemini 3.8 Flash is the more accurate model and finishes behind the one it replaces. The extra spend is not retries and not subagent fan-out: it is five times as much filesystem searching, at the same hit rate, in the same places.

> **Correction, 2026-09-04.** This report originally read the over-execution as a property of the model. It is mostly a property of the **effort setting**. Run `#138` re-ran the same model at `low` across all six workflows: calls fall 53%, objective gives up 1.8 points, and **OVR rises from 79 to 86** — a better consolidated card than its predecessor's 83. The findings below stand as a description of `xhigh`; they do not support the claim that the newer model is worse. The evidence is in **The effort arm**, below.

2026-09-03 · run `#134` · 1 model × 6 workflows × 2 trials (5 workflows) and 1 trial (confirmation-desk-day) · pinned to `xhigh` · output budget unpinned · objective-only scoring (jury off) · 6 of 6 arms `scored`, zero truncated calls, zero malformed tool calls · compared against the `gemini-3-7-flash` arm of run `#127`, the same route at the same effort · control arm: run `#135`

Why this run exists

Gemini 3.8 Flash was published on 2026-09-01. The desk's previous Gemini contestant, 3.7 Flash, holds an unusual position in the standings: it has posted the **highest objective score in the flash tier** (94.3 on run #129) while ranking third, because its efficiency is the worst of the leading group. The obvious question for a successor was whether it kept the accuracy and fixed the cost.

It did the opposite of the second half.

Headline

Consolidated card, six workflows, published to `/arena/models.html` as cards only — one model has no field, so a rank of #1 of 1 would measure nothing.

Model	eff	OVR	GRD	ADH	SYN	PRC	EFF	CON	Objective
Gemini 3.8 Flash	xhigh	**79**	91	97	99	95	**4**	89	**95.6**
Gemini 3.7 Flash *(run #129)*	xhigh	**83**	95	92	99	92	**31**	89	94.3

On the five workflows the two arms share, at the same effort on the same route:

	Gemini 3.8 Flash	Gemini 3.7 Flash
Objective score	**97.1**	94.3
Workflows won on objective	**4 of 5**	1 of 5
Perfect 100.0 scores	**2**	0
Tool calls per trial	95.9	**52.0**
EFF	4	**31**
Mean OVR	80.6	**83.0**

The newer model is better at the work and ranks lower for it. It wins four of the five workflows on correctness, posts two perfect scores where its predecessor posts none, and spends **84% more tool calls** to do it. EFF is scored as golf against each workflow's calibrated par, so that difference is not a rounding adjustment — it is the whole ranking.

Workflow	3.8 obj	3.8 calls	3.7 obj	3.7 calls
ops-settlement-day	**100.0**	84.5	89.8	**38.5**
risk-limit-breach-day	**100.0**	45.0	98.7	**29.0**
risk-manager-control-day	**97.4**	92.0	93.6	**58.5**
trader-rfq-booking-day	**96.8**	136.5	93.7	**69.5**
high-board-portfolio-review-day	91.4	121.5	**95.7**	**64.5**

Every row goes the same way. There is no workflow where 3.8 is both more accurate and cheaper, and none where it is cheaper at all.

Where the calls actually go

The scores say *how much* more it spends. The transcripts say *on what*. Counting every tool call in all ten trials across the five shared workflows, then averaging per trial:

Tool	3.8 Flash	3.7 Flash	Delta
`grep`	64.5	14.5	**+50.0**
`ls`	22.5	**0**	+22.5
`glob`	23.0	6.5	+16.5
`read_artifact`	17.5	2.5	+15.0
`inspect_artifact`	11.5	2.5	+9.0
`list_artifacts`	10.0	3.0	+7.0

Grouping those together against everything else:

	3.8 Flash	3.7 Flash
Filesystem / search calls	**149**	29
Domain and quant tool calls	352	253
Search as a share of all calls	**29.7%**	10.3%

Filesystem search is 55% of the entire increase. The remaining 45% is spread thinly over two dozen domain tools, none of which moves by more than about ten calls. The single most striking cell in the table is `ls`: across ten trials and five workflows, Gemini 3.7 Flash never calls it once, and 3.8 calls it 22 times.

Where it searches is consistent. Ranked by target, 3.8's search calls go to `/skills` (53), the filesystem root (25), `/large_tool_results` (24) and `/artifacts` (8). Its predecessor touches `/skills` 24 times and almost nothing else.

What it is not

Three cheaper explanations are available, and the transcripts rule out all three.

It is not retrying after failures. Gemini 3.8 Flash records *fewer* errors than its predecessor, 0.3 per workflow against 0.8. Whatever the extra calls are, they are not recovery.

It is not subagent fan-out. Both models dispatch essentially the same number of `task()` subagents, 1.5 against 1.4 per workflow. The extra work is happening in the orchestrator's own loop.

It is not worse searching. This is the one that matters, and it is the one that survives contact with the data least comfortably for the intuitive story. Classifying every search call by what came back:

	3.8 Flash	3.7 Flash
Returned content	**78%**	79%
Returned nothing	17%	14%
Refused (foreign store)	5%	3%
Exact repeat of an earlier call	23%	21%

The hit rates are the same. The repeat rates are the same. Both models also concentrate their searching **late** in an episode rather than front-loading it as orientation — 55% of 3.8's search calls fall in the last third, against 60% for 3.7.

So this is not a model that searches badly, or searches differently, or searches in the wrong places. It searches *exactly the way its predecessor does*, with the same productivity per call, and does five times as much of it before it stops. The deficit is not in the searching. It is in the stopping.

The confound, and the control that sizes it

There is a problem with everything above, and it is ours rather than the model's.

The 3.7 baseline ran on 2026-08-26. Run #134 ran on 2026-09-02, and in between we changed the harness. Gemini 3 routes return an encrypted **thought signature** beside every tool call and require it echoed back on the assistant turn carrying that call; `ChatOpenAI` does not preserve that field, by design, so the second model turn of every tool loop was being refused with a 400. Run #134 initially failed on all six workflows before we fixed it. (The control that established this was a harness defect rather than a new-model limitation: 3.7 Flash, which has four published boards on this route, fails identically once the field is stripped.)

That fix hands the model back its own prior reasoning. It is entirely plausible that doing so makes a model more willing to keep pulling on a thread — which would mean we had measured our own change and attributed it to Google.

So we re-ran **Gemini 3.7 Flash today, with the fix**, on `ops-settlement-day`. That workflow is the right instrument: 3.7's own trial-to-trial spread there is a single call, so one trial separates the hypotheses cleanly.

Counting the transcripts, so that total calls and search calls come from one instrument:

Arm	Tool calls	Search calls	Objective
3.7 Flash, run #127, **without** the fix	45, 46	1, 4	89.8
3.7 Flash, run #135, **with** the fix	**55**	**10**	90.9
3.8 Flash, run #134, with the fix	85, 98	29, 43	100.0

The fix is not inert. It moves 3.7 from 45.5 calls to 55, and its search calls from 2.5 to 10. Had we published the naive comparison, part of what we called a model regression would have been our own patch.

It does not, however, come close to explaining the gap:

Component	Calls	Share of the gap
Total gap, 3.8 with fix vs 3.7 without	+46.0	100%
Attributable to the harness change	+9.5	**21%**
Attributable to the model	+36.5	**79%**

The same decomposition holds for search calls alone (22% harness, 78% model), and the two independent instruments — the trace harvest that feeds EFF, and a direct count of the transcripts — agree on the split to the percentage point.

The conclusion is robust to the correction. EFF is scored on the trace harvest, which puts 3.8 at 84.5 calls on this workflow. Strip the entire harness component out and it still runs `ops-settlement-day` in 75 calls against a calibrated par of **30** — 2.5× par, which floors EFF just as thoroughly as 2.8× does. No plausible adjustment for the harness rescues the ranking.

The effort arm: what actually drives the searching

Added 2026-09-04, runs `#136` / `#137`.

Everything above holds `xhigh` fixed and varies the model. The obvious complementary test is to hold the model fixed and vary the effort — and the arena already contained half the answer for free. Run #127 and run #130 are **Gemini 3.7 Flash on the same five workflows with effort as the only variable**:

	3.7 at `xhigh`	3.7 at `low`
Tool calls per workflow	56.4	**35.5**
Search calls per workflow	5.8	**0.5**
Search as a share of calls	10.3%	**1.4%**

Search falls **91%**. On three of the five workflows it goes to exactly zero. Whatever the file rummaging is, it is switched on by reasoning effort rather than baked into the model.

So we ran Gemini 3.8 Flash at `low` across **all six workflows** (run `#138`), one trial each — same model, same route, same harness, same commit, with effort as the only variable. It is the paired arm for run #134, and it reverses this report's framing.

	`low` (run #138)	`xhigh` (run #134)
OVR	**86**	79
GRD	91	91
ADH	92	97
SYN	99	99
PRC	94	95
EFF	**47**	4
Objective	93.8	**95.6**
Tool calls per workflow	**75.3**	161.6
Search calls per workflow	**26.7**	51.3

Halving the effort halves the work and raises the score. Calls fall 53%, search calls fall 48%, objective gives up 1.8 points — and OVR rises seven, from 79 to 86, because EFF goes from 4 to

47. Grounding is *identical* at both efforts (91 and 91); the entire correctness cost is 5 points of adherence.

Per workflow, `low` wins four, ties one and loses one by a single point:

Workflow	`low` OVR	`xhigh` OVR	Δ	`low` obj	`xhigh` obj
ops-settlement-day	**98**	83	**+15**	100.0	100.0
risk-limit-breach-day	**97**	84	**+13**	100.0	100.0
high-board-portfolio-review-day	**80**	72	+8	85.7	91.4
trader-rfq-booking-day	**87**	81	+6	96.8	96.8
confirmation-desk-day	71	71	0	87.9	87.9
risk-manager-control-day	82	**83**	−1	92.3	97.4

Four of the six score identically on objective at both efforts, including both perfect 100.0s. The accuracy `xhigh` buys is confined to two workflows — `risk-manager-control-day` and `high-board-portfolio-review-day` — and on the golf curve it costs far more in calls than it returns.

The effort effect on searching is not uniform, and that is the most useful detail here. On five workflows filesystem search all but disappears at `low`. On `confirmation-desk-day` it barely moves: 151 search calls at `low` against 159 at `xhigh`, even though total calls halve from 490 to 256. That workflow's rummaging is **inherent to the task**, not bought by effort — which is exactly why it is the one workflow where dropping the effort changes nothing at all, OVR 71 either way.

So the corrected reading of run #134 is simple. `xhigh` on this model is a misconfiguration. At the desk default it is not merely competitive but the strongest consolidated card the flash tier has produced — **OVR 86 against the 83 its predecessor posts at its own ceiling.**

Scope. Run #138 is **one trial per cell**, so CON is *not measured* rather than perfect, and single-sample noise is real: `ops-settlement-day` was measured twice at `low`, in run #137 and run #138, and scored 93.2 and then 100.0. Read the per-workflow rows as indicative and the direction — five of six workflows improving, none worse by more than a point — as the finding.

Caveats

- **The harness control is one workflow and one trial.** ZenMux subscription quota ran out mid-investigation, so the paired 3.7 arm on `trader-rfq-booking-day` was deliberately abandoned to preserve what remained. The 21% figure is measured on `ops-settlement-day` and is **not** established as a constant across workflows. It is sized, not solved. (The *effort* arm, by contrast, now covers all six workflows.)
- **A comparability break now exists for Gemini contestants**, in the same way run #115 created one for Anthropic-protocol models when the output budget was raised. Every Gemini board from #134 onward runs with the thought signature preserved; #113, #127, #129 and #130 did not, and the control says that is worth roughly a fifth of the call count on at least one workflow. Read Gemini EFF across that boundary with care.
- **`confirmation-desk-day` carries one trial, not two.** Its second was lost to the same quota exhaustion and recorded `invalid` / `infra_blank` rather than scored, so its CON is *not measured*, not perfect. That workflow is also excluded from every 3.7 comparison in this report, because 3.7 has no arm on it at pinned effort.
- **Objective scores are not adjusted for the harness change at all.** The correction above applies to call counts. Whether the restored signature also helped 3.8's accuracy is untested, and the 3.7 control moved only 1.1 points on that axis.
- Zero truncation and zero malformed tool calls, **measured** rather than assumed: `trials_measured` equals `trials_total` on all six arms.

What would change the verdict

- ~~**A paired `low` arm.**~~ **Run across all six workflows (`#138`), and it changed the verdict** — see the effort arm above. A `low` card is now published beside the `xhigh` one. What remains is **depth**: run #138 is one trial per cell, so it carries no CON and the two `low` measurements of `ops-settlement-day` differ by 6.8 objective points. A two-trial re-run would turn the direction into a number.
- **The rest of the harness control.** Four more paired 3.7 arms would turn the 21% from a single-workflow measurement into a real coefficient, and would settle whether the effect is uniform or concentrated in workflows with large artifact trees.
- **A prompt-level test of the stopping rule.** The finding is that the model does not know when to stop searching, not that it searches badly. If that is right, it should be correctable from the orchestrator prompt rather than requiring a different model — and a cheap A/B on `ops-settlement-day` would show it.

Cards for this run are published at `/arena/models.html`. Nothing here reaches the leaderboard: a rank is relative and this run had no field.
