Run #133 — confirmation desk day: the first vision board

Dates: started 2026-08-29, completed 2026-08-30 — the leaderboard dates a board by when its run was created, so it files this one under the 29th **Workflow:** `confirmation-desk-day`, 9 steps / 33 checks, **uncalibrated par Contestants:** the four `vision`-tagged models, **2 trials each**, effort and output budget unpinned (32768 tokens, both wire protocols) **Run:** #133 on the live desk database · 8 clean trials · 0 invalid · 0 errors **Re-scored 2026-08-31:** the `gemini-3-7-flash` arm was re-run clean after two cross-contestant read channels were found and closed — see the correction and its resolution below. All other arms are as first measured.

Board

#	Contestant	Objective	OVR	GRD	ADH	SYN	PRC	EFF	CON	calls
1	**deepseek-v4-flash-vision**	86.3	**81**	**94**	99	99	60	**24**	**96**	46, 30
2	glm-5-3-flash	**87.9**	78	89	99	99	70	18	86	63, 37
3	gpt-5-6-luna	84.8	68	59	99	86	**94**	10	89	59, 74
4	gemini-3-7-flash	77.3	67	59	99	99	64	8	92	67, 115

Zero truncation, zero malformed tool calls, zero step errors across all eight trials.

> **The highest objective score does not win the board.** `glm-5-3-flash` posts 87.9 to deepseek's 86.3 and still places second: grounding carries 0.32 of OVR and deepseek reads better (94 vs 89), consistently (CON 96 vs 86), in fewer calls. And at the bottom, **two contestants share a GRD of 59 for the same deterministic reason**: `gpt-5-6-luna` and the re-run `gemini-3-7-flash` both fail the same conf-08 extraction-pipeline wall, miss the same four checks, and land on the identical grounding stat. Read that 59 with the analysis below before reading it as weak sight — and note that gemini's earlier published 84 was measured while it could still read other contestants' answers.

What this board was built to answer

Can GLM 5.3 Flash — z.ai's first multimodal model — work a confirmations desk against Gemini, GPT and DeepSeek? Yes, on correctness: it records the highest objective score in the field and ties the leaders on every designed vision trap it was given. It places second because it is neither the most consistent nor the leanest.

The more useful answer is that **the question is nearly obsolete at this tier.**

The designed vision traps are solved

Three of the four traps this workflow was built around are **saturated at 8/8** — every contestant, every trial:

Trap	Document	Truth	Rate
Image-only scan, no text layer	conf-04 GOOGL	strike 205.0	8/8
Struck-through and amended strike	conf-09 NVDA	1045.0, not the struck value	8/8
Barrier direction carried only by a ticked box	conf-10 AMZN	`DOWN_OUT`, 171.2	8/8
A term the document never states	conf-07 META	`initial_price` null, status `invalid`	8/8

That last row is the one worth pausing on. Step 7 asks for a value that is not in the document, with the strike sitting nearby as an obvious substitute. **No model substituted, in any trial.** The pilot board saw the incumbent substitute in 2 of 6 runs; on this field the behaviour is gone.

What still discriminates — 6 of 28 checks

Per-check pass-rate tally across the whole field. Keyed by label, so a check appearing in several steps is counted once per occurrence.

Check	Rate	What it is
Step 6 — ORCL notional 636,000	**1/8**	**The hardest check. A substitution failure, not a reading one — see below.**
`get_confirmation_batch` called	26/48	Procedure, not sight
Step 8 — `skipped_count` = 1	4/8	Booking restraint and counting
Step 3 — AMD strike 185.0	4/8	conf-08 page 2 — an **extraction-pipeline** failure, see below
Step 3 — AMD initial_price 178.9	4/8	Same document, same page, same cause
Step 9 — artifact names ARD-EQO-2026-04901	7/8	Carrying a reference into the summary

22 of 28 checks are dead (0/N or N/N). This tally is only valid because every arm on this board is healthy — a contaminated arm fails checks the healthy field passes and manufactures apparent discrimination. On the pilot board, with one arm poisoned by an infra defect, the same instrument read 8 checks dead where the repaired board read 25.

Neither hard check is an OCR failure

Both were re-examined against the transcripts and the source documents after the board was first published. Neither is what it looked like.

The ORCL notional is a SUBSTITUTION failure. The document states a Notional Amount of 636,000.00 in a low-contrast column; the generator also gives it `num_options: "4,000"` and a strike of 163.50. The three wrong answers are not misread digits:

Answer	Who	What it actually is
4,000	luna x2, glm x2, gemini x2	the **contract count**, reported as the notional
654,000	deepseek t1	**computed** 4,000 x 163.50
636,000	deepseek t2	correct — the stated value

(An earlier revision credited a contaminated gemini trial with the correct 636,000; its clean re-run answers 4,000 in both trials, like most of the field.)

Every contestant read the neighbouring strike 163.50 correctly, **8/8**. So the field can see that region of the page. What a faint value produces is not a wrong digit but a **substituted legible field** — the same instinct step 7 traps, minus step 7's explicit permission to answer null.

The AMD page-2 checks measure the EXTRACTION PIPELINE, not the agent. Because this workflow routes the extraction sub-call to the contestant, a contestant can fail here without ever misreading anything itself. conf-08's outcome is bimodal and perfectly split by model, identical on every one of their attempts:

- `gpt-5-6-luna` — `terms: {}`, always
- `gemini-3-7-flash` — `terms: {}`, always
- `glm-5-3-flash`, `deepseek-v4-flash-vision` — `185.0 / 178.9`, always

Stage 1 succeeds for all four (family, counterparty and reference `ARD-EQO-2026-04781` at confidence 0.99). **Stage 2 — filling the term schema from the rendered page — returns an empty dict for two of them.**

CORRECTION (2026-08-31): a cross-contestant leak, not a fabrication

An earlier revision of this report said `gemini-3-7-flash` "asserted the correct values without accessible evidence". **That was wrong, and the error is instructive.** The arena transcript records only the PARENT agent's tool calls; work done inside a `task()` subagent never appears in it. "No tool result contains the value" was therefore never valid evidence, and the trace database — which does record subagent spans — tells a different story.

What actually happened, in both gemini trials:

```
tool  task         ->  chain  general-purpose        (subagent, invisible to the transcript)
tool  read_file        /large_tool_results/call_3c5fd773d1f2455993d7552e
tool  record_answer    {"strike": 185, "initial_price": 178.9}
```

That file is not gemini's. It belongs to `glm-5-3-flash` — written by glm's own `parse_trade_confirmation` at 16:03, and read by gemini at 17:27 and again in its second trial. It contains conf-08's `strike: 185.0, initial_price: 178.9`.

`/large_tool_results/` is a **content-addressed store that is written per session but read globally.** `ContentAddressedFilesystemBackend._latest_artifact()` and `ls()` filter only on `kind == "tool_result"` and the rendered path — there is no `workflow_id`, `session_id` or thread predicate — while `capture_tool_result()` writes both ids. The workflow-scoped `list_artifacts` / `read_artifact` tools are the documented recovery route; the filesystem backend is an unscoped second door to the same store, and `glob` over it lists every other session's ids.

So a later contestant can read an earlier contestant's answers. That biases a board by POSITION IN THE FIELD — the same failure class as leftover fixture rows making each successive match's name resolution harder, and just as silent.

Measured on this board (all six gemini trials — the original pair and both remediation re-runs — plus the rest of the field):

Contestant	own conf-08 extraction	foreign reads
`gpt-5-6-luna`	**0 of 5** attempts	none
`gemini-3-7-flash`	**0 of 17** attempts, ever	**two channels — see resolution**
`glm-5-3-flash`	4 of 4	one of luna's (carried no useful value)
`deepseek-v4-flash-vision`	2 of 2	none

`gemini-3-7-flash`'s step-3 grounding passes were contaminated. Its own extraction of conf-08 never once succeeded; the values it recorded came from glm's data. `glm-5-3-flash` and `deepseek-v4-flash-vision` are unaffected — their extractions succeeded on their own, every time. Fifteen further reads of this kind appear across runs #129-#132, so this is not unique to this board.

RESOLUTION (2026-08-31): two doors closed, the arm re-run clean

Fixing this took two rounds, because the first fix's verification re-run found a second channel.

Round 1 — the CAS store. All four read surfaces of `/large_tool_results/` (`read` / `ls` / `glob` / `grep`) now resolve the calling workflow and **fail closed** when they cannot; a path owned by another session returns an explicit "belongs to another session" error. Gemini's arm was deleted and re-run — and the re-run scored higher (86.3), not lower. The trace showed why: it had stopped using the CAS route and was now reading **`/artifacts/` directly**, where the same content was still reachable.

Round 2 — the `/artifacts/` mount. The whole artifacts root was mounted with a blanket read allow, and that root *contains* the CAS blob store (`artifact_blobs/` — the round-1 fix was bypassable at the raw path), every contestant's arena transcripts (`arena/**` — the grade book), and every other thread's report workspace (`agent/thread-N/`). In the round-1 re-run gemini read another thread's desk summary and another contestant's arena transcript — 21 transcript-tree touches and 9 foreign-report touches across its four contaminated threads; the other three contestants: zero, in every trial. The mount is now scoped desk-wide: the restricted subtrees are denied outright and `agent/thread-N/` is visible only to its own thread, enforced per operation

in both agent stacks. (A third potential channel — the long-term memory layer injecting desk facts into contestants — was closed in the same pass; it was uniform across this field and carried no fixture values, so it does not caveat this board.)

Round 3 — the clean run, verified. With both doors closed, gemini's arm ran two clean trials: **zero** attempts on foreign thread dirs, arena paths or blob paths (the enumeration filtering removes the breadcrumbs, so there was nothing foreign left to find); every `/large_tool_results/` read was of its own workflow's artifacts; its one creative move — `os.listdir` inside `run_python` — died on the sandbox's WebAssembly virtual filesystem, which has no host access. Its own conf-08 extraction still failed every attempt (0/17 lifetime). The clean result: **objective 77.3, OVR 67, GRD 59** — against the 81.8/75/84 first published and 86.3/76/79 in the half-fixed interim.

Nine objective points of gemini's published position were leaked information. And the clean row is the strongest confirmation of this report's luna analysis: denied everyone else's answers, gemini lands on the IDENTICAL GRD 59, missing the identical four checks, for the identical upstream cause.

What this does NOT change: luna's GRD 59 is real and is not a sight deficit. The value `178.9` appears in **zero spans** of either luna thread — its extraction genuinely never produced it — and luna neither fabricated a number nor took one from a neighbour. It reported null and said why. Of the four contestants it is the only one that both failed the extraction and declined to obtain the answer by any other route.

One extraction failure, three of luna's four grounding misses

The cascade is why luna's GRD is an outlier rather than a gradient:

- conf-08 stage 2 returns `terms: {}` -> step 3 `strike` and `initial_price` recorded null (**2 checks**)
- conf-08 therefore validates `invalid` -> at step 8 luna skips META **and** AMD, so `skipped_count=2` rather than 1 (**1 check**)
- the fourth miss is the ORCL notional, which most of the field also misses

Luna's identical GRD of 59 on the pilot board is consistent with the same deterministic cause. And the strongest evidence arrived with the re-scored board: `gemini-3-7-flash`, run clean, lands on the **same GRD 59 with the same four misses** — the wall is the extraction pipeline, and luna was simply the only contestant honest enough to hit it in public the first time.

The board's variance is procedural, not visual

`glm-5-3-flash` scored **97.0 and then 78.8** — an 18-point spread on identical inputs, and the reason it dropped from first to second when its second trial landed.

Every point of that spread is one check. Its two trials read the documents *identically*: same strikes, same barrier, same absent term, same faint-notional miss. What differed is that trial 1 called `get_confirmation_batch` at all six graded steps and trial 2 called it at none.

`deepseek-v4-flash-vision` shows the same thing inverted. Its lower-scoring trial (81.8, 30 calls) is the one where it read **every trap correctly including the faint notional** — the only trial on the board to do so — and it scores lower than its 90.9 trial because it skipped the batch re-reads.

So the single widest-swinging check on this board rewards re-fetching a payload the model already holds. That is the concern the pilot report raised about these checks, now with a number on it: **`get_confirmation_batch` accounts for more objective variance than every vision check combined.** It measures procedure, and a model answering correctly from context it already has is arguably being efficient rather than wrong. This is the same class as the "step-scoped check penalises reading the evidence one step early" defect already recorded in `CLAUDE.md`.

Par is still not calibrated, and now there is evidence for why

`confirmation-desk-day` declares no `par_tool_calls`, so EFF runs on the legacy hyperbolic curve against a `designed_par` of **9** — the theoretical minimum. That crushes every EFF on this board into 10–24 on a 0–99 scale, and EFF is 0.16 of OVR.

The established way to calibrate is the median of the **fully-correct** trials. **This board produced none: not one of the eight trials passed every check.** The pilot's single 33/33 run is one data point, not a distribution. So par stays unset rather than being invented from a sample that does not exist, and **EFF and OVR here are not comparable with calibrated boards.**

For the record, the counted call distribution over 8 clean trials: median **61**, range **30–115**. The 115 is gemini's second clean trial — with the shared stores closed, its hunt for the conf-08 values it could not extract ran long before it settled on null, which is also why its clean EFF is the field's lowest.

Why this board can be trusted

The pilot board found three harness defects, each of which cost a real contestant real points, and each invisible to the golden replay. All three are fixed here, and this run is the evidence:

1. **`StreamChunkTimeoutError` after 120s.** Step 1 parses six documents in one tool body — 12+ vision calls synchronously — so the outer stream emits nothing for minutes and langchain's default fires on a connection that is idle rather than dead. Launched with `LANGCHAIN_OPENAI_STREAM_CHUNK_TIMEOUT_S=900`; zero contestants died to it.
2. **Unaddressable upload paths.** A model told the files are at `/artifacts/uploads/confirmations/` and then rejected for using that spelling scored 30.3 with grounding 0/10 on the pilot. Fixed; zero path errors here.
3. **A binary `read_file` poisons the conversation permanently.** deepagents returns a binary read as a media content block; three of four gateways reject that shape from a *tool* message, each in its own dialect, and the rejected message stays in the history so every later turn redraws the same 400. On the pilot this took `deepseek-v4-flash-vision` from a real 100.0 to a recorded 36.4 and last place. `BinaryReadGuardMiddleware` now intercepts it in all three hand-built agent stacks **and** in the `general-purpose` subagent that deepagents adds for itself — a fourth stack that inherited the parent's full toolset while running unguarded and unaudited.

Zero errors across eight trials is the result of those three fixes, not of an easy workflow.

Caveats a reader must carry

- **Do not read this ordering as a vision ranking.** 22 of 28 checks are saturated, separation comes from consistency, procedure and volume — and the grounding axis itself is partly measuring the extraction sub-call rather than the agent, because this workflow routes that sub-call to the contestant.
- **The two GRD 59s are not a sight deficit.** For both `gpt-5-6-luna` and the clean `gemini-3-7-flash`, it is one deterministic conf-08 extraction failure cascading into three checks, plus the field-wide notional substitution.
- **`gemini-3-7-flash`'s row is the RE-SCORED clean measurement.** Its first two published rows (81.8, then 86.3) were contaminated through two cross-contestant read channels, both now closed; the correction and resolution above carry the full history. The clean re-run is verified at the trace level: zero foreign reads, all recovery reads own-workflow, one failed sandbox probe. The other three contestants were never exposed on this check.
- **EFF is uncalibrated** (see above). Do not compare it across boards.

- **Memory injection was uniform.** All eight trials ran with the desk's long-term memory layer injecting the same five hand-seeded desk facts (none containing fixture values) into every contestant. Arena threads are excluded from memory injection and extraction going forward, so later boards run without this constant.
- **Two arms were re-run.** `gemini-3-7-flash` for the contamination above, and `glm-5-3-flash`, which lost a trial to a transport-level `RemoteProtocolError` — the peer closed the connection mid-body — where infra trials are skipped rather than retried, which would have left the headline contestant at half the field's depth with no CON. Its `scored` row was deleted and the arm re-run to two fresh trials via `--resume 133`, at the same effort and the same re-supplied output budget. **The repair changed the podium:** on the single surviving trial GLM stood at OVR 85 and first place; at full depth it is 78 and second. Its objective mean is 87.9 either way — the coincidence is real and is why OVR, not the objective percentage, ranks the board.
- **Two trials, not five.** CON is measured but from a small sample.
- **The extraction sub-call routes to the contestant.** Without that, every model would read every document with one shared extractor and every vision check would land N/N carrying no signal.

Next

- **Reconsider the `get_confirmation_batch` checks.** They are 1 of the 6 surviving discriminators and the largest single source of objective variance, and what they reward is re-fetching held context.
- ~~Scope the CAS read path~~ **Done, twice over.** The CAS reads are workflow-scoped and fail closed, and the `/artifacts/` mount that turned out to be a second door to the same bytes is scoped desk-wide (restricted subtrees denied, thread workspaces visible only to their own thread, in both agent stacks). The contaminated arm was re-run clean and this board carries the result. Boards #129-#132 retain their measured scores with the leak documented; whether any of their 15 cross-store reads moved a score has not been audited, and a reader of those boards should carry that caveat.
- **Fix the step-3 grounding check, which currently inverts the desk's own standard.** As written it scores a model for producing a number it cannot evidence and scores zero for reporting, correctly, that the term could not be extracted. Either grade the extraction outcome explicitly, or word the step like step 7 so that a justified null is a legal answer.
- **The workflow needs harder traps.** Its designed vision difficulty is solved by the whole field, and the two checks that still separate contestants turned out to measure substitution and an extraction failure rather than sight. That is where the next iteration should aim.

- **Do not calibrate par until a trial passes every check.** The instrument is the median of fully-correct trials, and that set is currently empty.
