The OTC Desk Agent Arena — Methodology & Results (Run #9)

A controlled, repeated-trial evaluation of nine flash-tier LLMs operating a real OTC derivatives trading desk, fully autonomously, with no human in the loop.

Date: 2026-06-28 · **Run:** #9 · **Task:** `risk-manager-control-day` · **Trials:** 5 per model · **Gateway:** ZenMux

---

Abstract

Most LLM benchmarks score a model on a frozen prompt. A trading desk is not a frozen prompt — it is a stateful environment where the operator must read risk, price portfolios, run scenarios, back-test hedges, and produce a governance report, each step depending on the last, with no human to approve or correct intermediate actions. The **OTC Desk Agent Arena** measures exactly this: it drives the *real* desk orchestrator end-to-end and scores whether a model can run a risk-manager's control day on its own.

Where Run #8 ranked the **frontier** tier, Run #9 turns the same harness on the **flash tier** — the fast, low-cost models a desk would actually reach for to run autonomous workflows at scale. We evaluated **nine flash models** over **five independent trials each** under the identical single execution regime (headless "YOLO" — no human-in-the-loop interrupts, no deferral tool), scored by a 31-point objective manifest combined 50/50 with a GPT-5.5 judge.

The headline is **not** the winner — it is the **price of the winner**. **Gemini 3.5 Flash tops the board (59.1)**, but at **$14.28 per match it is the single most expensive desk operator the Arena has ever measured — pricier than Run #8's frontier Opus and GPT-5.5 (~$11).** Right behind it, **Step 3.7 Flash (57.9)** lands within 0.1, is the steadiest model in the field, and costs **$1.04 — one-fourteenth as much.** "Flash" turns out to be a statement about latency, not cost.

This run also closes Run #8's biggest measurement gap. There, only the two Anthropic-native models surfaced token usage; every OpenAI-gateway model logged zero. With streaming usage capture now wired through the trace (`stream_usage=True`), **Run #9 reports exact, per-match token consumption and cost for every model** — no proxies, no estimates.

As in Run #8, we separate two failure modes that a naive leaderboard conflates: **model incapability** (the model connects fine but cannot operate the desk) versus **resource/infrastructure shortage** (the model is capable but its gateway route failed). Both

materialised this run — three models genuinely cannot operate the desk, and **Doubao** was censored by wedged routes (0/5 on its primary route; 2/5 on a sibling route, where it nonetheless posted the **highest functional score in the field, 65.3** — the run's dark horse). Each is flagged and reported separately (§6) rather than allowed to depress — or inflate — the board silently.

---

1. Introduction

Agentic benchmarks increasingly try to measure "can the model do the job," not "can the model answer the question." For a derivatives desk, the job is a *sequence* of stateful operations:

> "*Pull the latest risk on the control portfolio → price it on the control profile → find the hotspot underlying → run the Greeks landscape → stress it → back-test the hedge → write the governance report.*"

What makes this hard for an LLM is not any single step — it is doing all seven **without a human**: resolving ambiguous references ("the control portfolio") to the right entity, threading IDs from one async task into the next, and never stalling to ask for confirmation. A model that is excellent in a chat box but freezes the moment no human is present will fail here, and that failure is invisible to a single-shot benchmark.

Why the flash tier? The frontier models in Run #8 run the desk well, but at ~$11 a match they are expensive to operate at the cadence real automation implies (every portfolio, every morning). The flash tier promises most of the capability at a fraction of the price. Run #9 asks the obvious follow-up: *how much desk-operation autonomy survives the move down to the fast, cheap tier — and which flash model gives you the most of it per dollar?* The answer is more nuanced than "you get what you pay for."

What the arena measures

> **The ability to operate the desk as a competent human would, end-to-end, with zero human intervention.**

Concretely: did the model route the expected skills, call the right tools on the right entities, produce the durable artifacts (risk run, valuation, scenario, back-test, report), and narrate the governance decisions — across all seven steps, on its own.

2. The Arena

2.1 The task

The flagship workflow `risk-manager-control-day` is a **seven-step risk-manager control day** with a 31-point objective manifest. The scenario is seeded so that a specific underlying (AAPL) becomes the portfolio's risk hotspot only *after* a fresh pricing run — a model that prices the wrong portfolio, or reads stale risk, surfaces the wrong hotspot and cascades errors through the remaining steps.

2.2 The environment — real, not simulated

Each trial:

1. **Seeds fixtures** into the live desk database under fresh, arena-tagged IDs (portfolios, positions, pricing-parameter profiles), then purges them after.
2. **Creates a real agent thread** and drives every workflow step through the production `AgentService.stream_and_persist` path bound to the candidate model.
3. **Reconstructs the transcript from the trace log.** Skills routed are *ground truth* — the deep-agent loads each skill it follows by reading the actual `SKILL.md` file, so the trace records what the model genuinely did, not what it claimed.

Because it is the real orchestrator, every production safeguard, tool, and sub-agent is in play — including the failure surface (gateway latency, tool errors, sub-agent delegation), which becomes part of what we measure and, where relevant, control for.

2.3 Execution regime — headless "YOLO"

The arena drives every turn in a **single regime**: fully headless.

- **No human-in-the-loop interrupts.** HITL gates auto-clear.
- **No deferral tool.** The `propose_reply_options` card tool — the mechanism a model uses to bounce a decision back to a human — is *withheld entirely*, and the system prompt instructs the model that no user is present.

This eliminates a confound: a model cannot "pass" by deferring to a (non-existent) human. Removing the deferral path means **the model must commit to its own judgment**, which is the entire point of an autonomy benchmark. As §5 shows, three of the nine flash models fail precisely here — they stall asking for a confirmation that will never come.

> **Single-regime fairness.** Every model in Run #9 was driven through the identical headless path. A model cannot score by punting to a human, and none is advantaged by a different harness.

2.4 Scoring

Each trial's total is a 50/50 blend:

- **Objective (0–100):** fraction of the 31-point manifest satisfied — deterministic checks against the trace (correct entities priced, hotspot identified, artifacts produced, async task IDs surfaced, etc.).
- **Judge (0–100):** GPT-5.5 grades the transcript against the workflow rubric (governance narration, correctness, completeness), with retries to ride out transient gateway errors.

`total = 0.5 · objective + 0.5 · judge`.

2.5 Repetition and averaging

Every model runs **five independent trials**; its score is the **mean** of the trial totals, reported with the **standard deviation** as a first-class reliability signal. The `ArenaMatch` table enforces one row per (run, workflow, model), so the five trials are averaged into a single row whose breakdown carries the per-trial detail. A trial that fails for **resource/infrastructure** reasons (a wedged or SIGKILLed gateway route) is recorded as such and excluded from the model's mean, which is then reported with its reduced *n* and flagged — distinct from a trial where the model connected but could not operate the desk.

---

3. Models evaluated

Nine flash-tier models, all routed through the ZenMux gateway over its OpenAI-compatible protocol:

Model	Vendor	ZenMux route
Doubao Seed Evolving	ByteDance	`bytedance/doubao-seed-evolving`
Qwen 3.7 Plus	Alibaba	`qwen/qwen3.7-plus`
Step 3.7 Flash	StepFun	`stepfun/step-3.7-flash`
Agnes 2.0 Flash	Sapiens AI	`sapiens-ai/agnes-2.0-flash`
Gemini 3.5 Flash	Google	`google/gemini-3.5-flash`
GPT-5.5 Instant	OpenAI	`openai/chat-latest`
DeepSeek V4 Flash	DeepSeek	`deepseek/deepseek-v4-flash`
MiMo V2.5	Xiaomi	`xiaomi/mimo-v2.5`
Hunyuan 3 Preview	Tencent	`tencent/hy3-preview`

When the ByteDance candidate's primary route (`doubao-seed-evolving`) proved infrastructure-censored, Doubao was additionally re-run on the sibling route `bytedance/doubao-seed-2.1-turbo` to recover a functional result (§6).

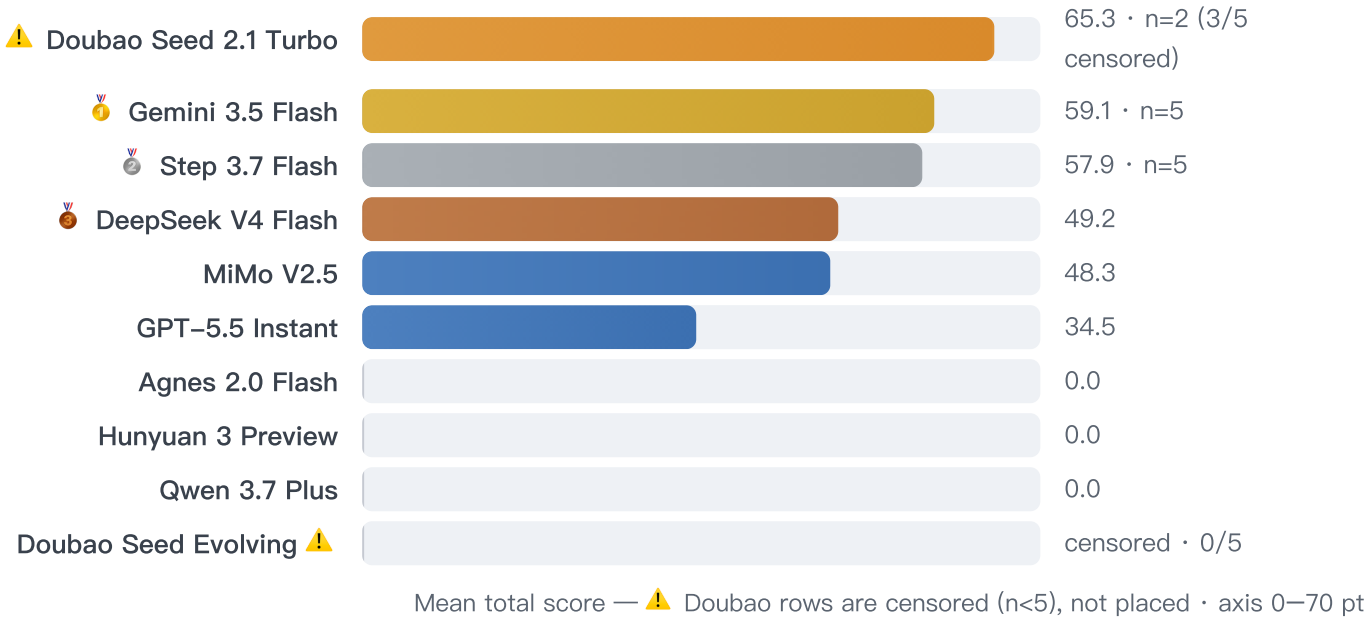
The judge throughout is **GPT-5.5** (the full model, not Instant), the same judge family as Run #8, so the two runs' judge axes are comparable.

---

4. Results

4.1 The leaderboard

Mean total score — two flash models clear the bar cleanly; a censored dark horse posts the highest functional number; a long tail collapses (1 block ≈ 2 points):



Rank	Model	Total	σ	Objective	Judge	n
⚠	Doubao Seed 2.1 Turbo ⚠	*65.3*	9.1	53.2	77.5	**2**
1	**Gemini 3.5 Flash**	**59.1**	5.5	57.4	60.8	5
2	**Step 3.7 Flash**	**57.9**	**4.3**	52.3	63.5	5
3	DeepSeek V4 Flash	49.2	15.5	40.0	58.5	5
4	MiMo V2.5	48.3	12.4	47.1	49.5	5
5	GPT-5.5 Instant	34.5	19.8	34.2	34.8	5
6	Agnes 2.0 Flash	0.0	0.0	0.0	0.0	5
7	Hunyuan 3 Preview	0.0	0.0	0.0	0.0	5
8	Qwen 3.7 Plus	0.0	0.0	0.0	0.0	5
—	Doubao Seed Evolving ⚠	*censored*	—	—	—	**0**

⚠ **The two Doubao rows are not comparable to the clean n=5 rows.** The ByteDance candidate's primary route (`doubao-seed-evolving`) was **infrastructure-censored** — it wedged on all five attempts (0 functional trials). Re-run on the sibling route `doubao-seed-2.1-turbo`, Doubao posted the **highest functional score in the field (65.3)** but completed only **2 of 5** attempts. Neither is a placed result; both are reported separately in §6. **Among models that completed all five trials, Gemini 3.5 Flash (59.1) and Step 3.7 Flash (57.9) lead.**

4.2 Headline: the winner is the worst buy

The asterisk first. Doubao Seed 2.1 Turbo posts the single highest functional score (65.3) — and on its best trial earned the highest judge score of the entire run (88.75) — but it completed only 2 of 5 attempts, too few to crown and not comparable to the clean rows (§6). The placed contest is between the two models that completed all five trials.

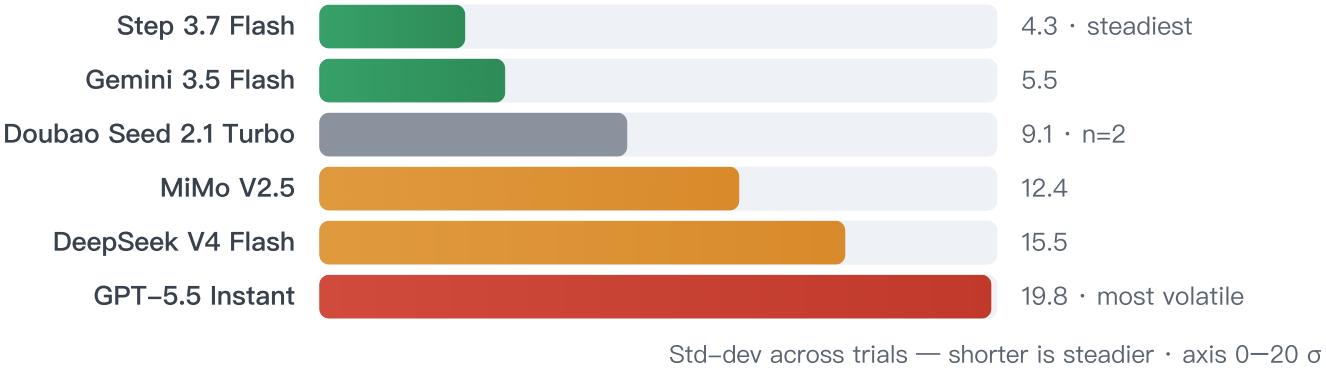
Gemini 3.5 Flash (59.1) and Step 3.7 Flash (57.9) are separated by **1.2**, well inside their combined variance — effectively a tie for the flash crown. They reach it differently: **Step wins the judge axis** (63.5 vs 60.8 — cleaner governance narration) while **Gemini wins objective coverage** (57.4 vs 52.3 — it satisfies more manifest checks, routing all 7 skills and firing 80 tool calls in its best trial).

But the leaderboard hides the real story, which only the (now exact) cost data tells: **Gemini's win costs $14.28 a match — more than any frontier model in Run #8** — because it is both pricier per token ($1.5 / $9 per M) *and* token-hungry (8.5 M prompt tokens a match). **Step delivers 98% of Gemini's score for $1.04** — a **14× cost gap for a 2% quality gap.** If you are choosing a flash model to run the desk unattended, the rank-1 model is the one you should *not* pick. See §7.

4.3 Consistency is a signal, not noise

Rank ≠ reliability. The standard deviation tells you whether you can trust a model to run the desk unattended on *any given day*:

Reliability — standard deviation across the five trials (**shorter is steadier**; 1 block ≈ 1 point of σ):



- **Step 3.7 Flash — σ 4.3:** the steadiest operator in the field, never dropping below 50.7 across five trials. Combined with its cost, this is the model you would actually deploy.
- **GPT-5.5 Instant — σ 19.8:** wildly inconsistent, swinging 18.8 → 57.0 → 18.8. Its mean (34.5) is almost meaningless: on a given day it either runs the desk competently or stalls in

read-only mode (see §5). The *instant* variant trades away exactly the consistency an autonomy use-case needs.

- **DeepSeek V4 Flash — σ 15.5:** the open-weight swinger, 22.9 → 61.7, echoing the full DeepSeek's volatility in Run #8 (σ 13.7).

4.4 The floor

Agnes 2.0 Flash, **Hunyuan 3 Preview**, and **Qwen 3.7 Plus** sit at the floor across all five trials (σ 0.0) — these zeros are reproducible and real (see §5). Three of nine flash models simply cannot operate the desk without a human.

---

5. Failure-mode taxonomy

A score near zero can mean two completely different things. The arena's trace log lets us separate them by reading what the model actually *emitted*.

Failure mode	Signature	Cause
Infrastructure wedge	run never completes; connection frozen, 0 output	gateway/route, **not the model** (recorded as a resource failure, SIGKILLed at the hard cap)
Infrastructure blank	run completes but emits **0 characters** every step	gateway returned empty streams
Model stall	run completes, emits **coherent text**, but 0 useful tool calls	the **model** can't/won't operate

The discriminator is simple: **a model that fails *behaviorally* still talks.** A model whose *pipe* failed produces nothing.

Evidence from the transcripts:

Run	tool calls	completion tok	what it emitted
Agnes 2.0 Flash — 0.0	**0**	~1,000	pseudo tool-call JSON written as prose; stalled at step 1 asking for clarification despite being given the control portfolio
Qwen 3.7 Plus — 0.0	**0**	~13,000	coherent prose, then a claimed "subagent delegation / tool-call ID error" that aborted all execution
Hunyuan 3 Preview — 0.0	**0**	~5,000	stalled before tool execution every step — delegation failure plus unnecessary clarification
Doubao Seed Evolving	—	—	*nothing* — route wedged, 0 output, SIGKILLed at 1800 s × 5 (see §6)

- **Agnes / Qwen / Hunyuan** are *genuine model results*: the pipe worked (0 wedges in 5 trials each), the model emitted thousands of characters, but it could not operate the desk. All three share the same failure: they **stall in the headless regime**, either asking for a confirmation that no human will give or emitting tool calls as natural-language pseudo-code that never parse into real calls. **Their zeros are trustworthy** — this is a capability ceiling, not a pipe failure.
- **Doubao's absence** is *not a model result*: its route never produced a single completion across five 30-minute attempts. That is the signature of an infrastructure wedge, not a model that tried and failed.

---

6. Doubao — an infrastructure-censored result, on two routes

Doubao deserves separate treatment because its raw leaderboard number badly misrepresents what we know about the model — and because we measured it on **two routes**, which together tell the real story.

6.1 The primary route: `doubao-seed-evolving` — fully censored

A trivial one-shot connectivity probe to `bytedance/doubao-seed-evolving` returned cleanly in ~4 seconds with token usage reported — the route is *reachable*. But under the sustained, multi-turn agentic load of a full control day, every one of its five attempts **wedged**: the connection froze mid-run, produced zero output, and was SIGKILLed at the 1800-second hard cap.

Outcome	Count	Interpretation
Wedged / timed out	**5**	route froze under load; SIGKILLed — *infrastructure*
Completed, functional	0	—

This is the same proxy-wedge pathology Run #8 documented for GLM 5.2: a blocking-sync freeze the in-process async timeout cannot cancel, where a fresh probe to the same model still answers in seconds.

6.2 The sibling route: `doubao-seed-2.1-turbo` — the best operator, when it runs

Because the primary route was censored, we re-ran Doubao on the sibling `doubao-seed-2.1-turbo` route (still Run #9). The route is healthier — but only just:

Outcome	Count	Interpretation
Completed, functional	**2**	**71.8 and 58.9** → mean **65.3** — *the real model*
Wedged (SIGKILLed)	2	route froze under load — *infrastructure*
Turn-timeout (slow)	1	a single turn ran past the 25-min per-turn cap — *performance*

When it completes, Doubao Seed 2.1 Turbo is **the strongest desk operator in the flash field**: a functional mean of **65.3** (above Gemini's 59.1), and on its best trial it earned a **judge score of 88.75 — the highest in the entire run** — routing all seven skills with clean governance narration. But it completes only **2 of 5** attempts, and even its successes are slow (11–24 min). For an *unattended* desk that is a disqualifying reliability profile, however high the ceiling.

6.3 How we report it

We therefore **do not place either Doubao row on the leaderboard.** Both are *resource/infrastructure* (and, for Turbo, partly *performance*) limitations of this evaluation window — explicitly distinct from the *model* limitations of Agnes, Qwen, and Hunyuan, which completed every trial and simply could not operate the desk. We report Doubao two ways, GLM-style:

- **Functional estimate (Turbo): 65.3** (mean of 2 completed runs; **n=2, low-confidence**) — its ceiling on a route that holds.
- **Primary route (Evolving): uncharacterizable** (0 of 5 completed) — a dead route this window.

A clean re-run on a healthy route is required to place Doubao with confidence.

> **This is the "mark it specially" case.** The harness tags every errored trial with a `fail_class` (`infra_timeout` for a wedge, `subprocess_error` for the turn-timeout) so a dead or slow route can never be confused with — or silently averaged into — a model's real score.

---

7. Cost & token usage

> **Note.** Unlike Run #8, these figures are **measured, not estimated** — every model in Run #9 routes through ZenMux's OpenAI-compatible gateway, and with streaming usage capture wired into the trace, each match's exact prompt and completion token counts are summed from the trace spans of that match's thread.

7.1 What we measured

Mean tokens per match (5 trials each; the workload is ~98% prompt-dominated — the desk loads large context each turn, while the model's own output is small):

Model	Prompt / match	Completion / match	Total / match
Gemini 3.5 Flash	8,530,694	164,777	**8.70 M**
Step 3.7 Flash	7,016,607	121,795	7.14 M
DeepSeek V4 Flash	4,605,724	130,497	4.74 M
Doubao Seed 2.1 Turbo ‡	4,232,275	76,871	4.31 M
MiMo V2.5	3,913,534	66,308	3.98 M
GPT-5.5 Instant	1,720,606	15,538	1.74 M
Qwen 3.7 Plus †	676,469	15,372	0.69 M
Hunyuan 3 Preview †	406,736	3,442	0.41 M
Agnes 2.0 Flash †	63,998	967	0.06 M

† zero-scoring models — their low token counts reflect early give-up, not efficiency. ‡ Doubao Seed 2.1 Turbo: mean over its **2 completed** trials (§6).

Note the **spread of 130×** in token appetite between Gemini (8.70 M) and Agnes (0.06 M). A model that stalls early is cheap precisely *because* it does no work; token volume only means value when the work gets done.

7.2 Price and measured cost

ZenMux list price (USD per million tokens) and **measured** cost per match. The no-cache column is the upper bound; the cache column assumes ~90% of the prompt-heavy context is served from cache-read (which these gateways price at ~1/10–1/50 of fresh prompt):

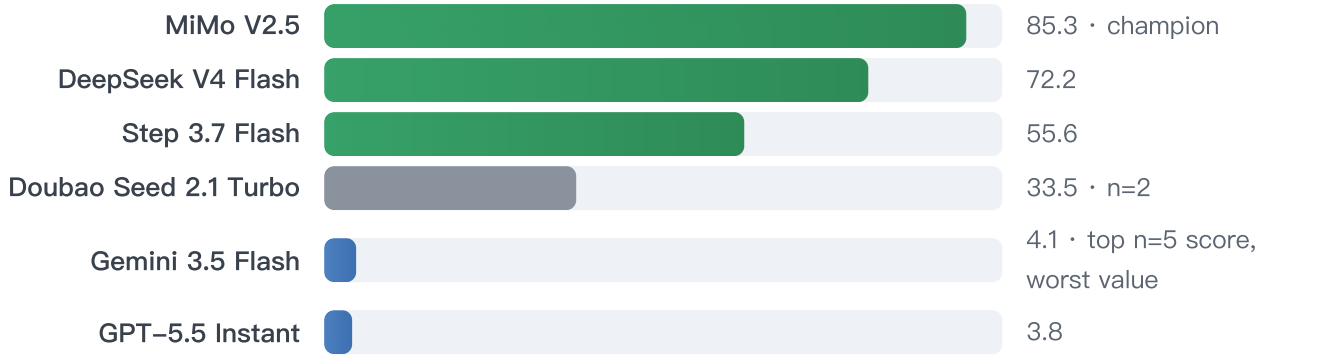
Model	in $/M	out $/M	$/match (no cache)	$/match (cached)
Gemini 3.5 Flash	1.50	9.00	**14.28**	3.91
GPT-5.5 Instant	5.00	30.00	9.07	2.10
Doubao Seed 2.1 Turbo ‡	0.423	2.113	1.95	0.50
Step 3.7 Flash	0.135	0.776	1.04	0.36
DeepSeek V4 Flash	0.140	0.280	0.68	0.11
MiMo V2.5	0.140	0.271	0.57	0.08
Qwen 3.7 Plus	0.137	0.549	0.10	0.03
Hunyuan 3 Preview	0.138	0.460	0.06	0.02
Agnes 2.0 Flash	0.100	0.200	0.01	0.01

‡ Doubao Seed 2.1 Turbo over its 2 completed trials — mid-tier on cost ($1.95), but see §6 for the reliability caveat that no price can offset.

The two costliest models in the field are the two with *frontier* pricing: Gemini 3.5 Flash ($1.5 / $9) and GPT-5.5 Instant ($5 / $30, identical to full GPT-5.5). Everything else is genuine flash pricing (~$0.1–0.14 in). "**Flash" is a latency claim, not a price claim** — and on this prompt-heavy desk workload, the per-token price compounds with token appetite to make Gemini the single most expensive operator the Arena has measured across both runs.

7.3 Cost–performance

Score per dollar (mean total ÷ measured no-cache $/match) — cost-efficiency *inverts* the quality ranking (1 block ≈ 3 points/$):



Score per dollar (measured) — cost-efficiency inverts the quality ranking · axis 0–90 pts/$

- **MiMo V2.5 — 85 points per dollar:** the cost-performance champion (48.3 at $0.57). If you want the most desk-operation per dollar, this is it.
- **DeepSeek V4 Flash (72) and Step 3.7 Flash (56)** round out the Pareto-efficient cluster — all three deliver real desk work at well under a dollar a match.
- **Doubao Seed 2.1 Turbo (33.5, n=2)** sits mid-table on value — its high score offsets its mid-tier price — but its 2-of-5 completion rate (§6) makes the number aspirational, not bankable.
- **Gemini 3.5 Flash (4.1) and GPT-5.5 Instant (3.8)** are the value floor: a frontier price tag for a flash-tier (or worse) result.

> **The practitioner's read.** The model to *deploy* is **Step 3.7 Flash** — near the top on score (57.9), the steadiest in the field (σ 4.3), and ~$1 a match. The model to *avoid* is the rank-1 winner: Gemini 3.5 Flash buys you 1.2 extra points of mean score for **14× the cost** and more variance.

8. Limitations & threats to validity

- **One workflow.** Run #9 covers a single flagship task. Rankings may shift on other desk workflows; this is a depth-over-breadth design. We're also working on other long-workflow match designs, and will make them public soon.
- **Judge model in the loop.** Half the score comes from a GPT-5.5 judge, which may carry stylistic bias. (Notably it did *not* favour its own family — GPT-5.5 Instant placed fifth, and full GPT-5.5 is not a candidate this run.)
- **Gateway variance.** All traffic shares one gateway (ZenMux). Route instability is real and asymmetric — it censored both Doubao routes this run. We detect and flag it rather than letting it silently depress a score, but it remains a confound.

- **Doubao under-sampled.** Its primary route (`doubao-seed-evolving`) returned 0/5; the sibling `doubao-seed-2.1-turbo` returned only 2/5 (mean 65.3, low-confidence). Neither is a placed result; a clean re-run on a healthy route is required. Every other model reached the full five trials.
- **Cache assumptions.** The "cached" cost column is a model of a real deployment with prompt caching; the measured, reproducible figure is the no-cache column.
- **Flash-tier framing.** These are fast/low-cost variants; a vendor's full-size model may score very differently (compare Run #8's frontier board).

---

9. Reproducibility

- **Run:** `ArenaRun #9`, status `completed`, viewable on the `/arena` page; each model's per-trial detail is in its `score_breakdown.aggregate`.
- **Regime:** headless YOLO (`mode="yolo"`) — HITL auto-cleared, deferral tool withheld.
- **Token capture:** `build_agent_model` sets `stream_usage=True`, so `usage_metadata` flows into the trace token columns for every model; per-match totals are summed over the match's trace spans (this is what makes §7 exact).
- **Engineering note:** an intermittent gateway wedge (any model, a blocking-sync freeze the in-process async timeout could not cancel) forces each match to run in an isolated subprocess under a hard `SIGKILL` wall-clock guard, with checkpoint-to-disk so the 45-match sweep is resumable across restarts. This is why the failure surface is measured rather than hidden — and why Doubao's five wedges cost wall-clock time but not data integrity.

Appendix A — per-trial totals

Model	trials	n
Gemini 3.5 Flash	56.4, 58.8, 53.3, 58.8, 68.1	5
Step 3.7 Flash	58.7, 61.3, 50.7, 60.8, 57.9	5
DeepSeek V4 Flash	22.9, 52.3, 59.2, 50.1, 61.7	5
MiMo V2.5	30.1, 49.2, 61.3, 43.2, 57.7	5
GPT-5.5 Instant	18.8, 22.6, 57.0, 55.2, 18.8	5
Agnes 2.0 Flash	0.0, 0.0, 0.0, 0.0, 0.0	5
Hunyuan 3 Preview	0.0, 0.0, 0.0, 0.0, 0.0	5
Qwen 3.7 Plus	0.0, 0.0, 0.0, 0.0, 0.0	5
Doubao Seed 2.1 Turbo ‡	71.8, 58.9 *(+1 turn-timeout, +2 wedged)*	2
Doubao Seed Evolving	*5 wedged / SIGKILLed — infrastructure-censored*	0

‡ Doubao Seed 2.1 Turbo — sibling route, functional runs only; 3 of 5 attempts failed (infrastructure/performance), so n=2 and not placed on the board (§6).

Run #9 · OTC Desk Agent Arena · generated 2026-06-28.