Migration assessment

Decision brief

Measured research: 29-30 September 2026. Frontier illustration added 2 October 2026.

Hold the migration decision

No measured candidate completed final confirmation. Use the findings below to narrow the next assessment; none of these results authorizes a production switch.

What Pisama has established	Why it matters
Quality alone did not settle the choice	GLM / Sail met the original quality and error screen, but was later set aside for latency.
A faster route had a reliability problem	GLM / Crusoe met its quality-consistency, latency and cost checks, but 5 of 40 attempts hit rate limits.
Aggregate improvement can hide a harmful change	The historical Terra pair improved task success, yet one attempt made an impermissible simulated cancellation.

Recommended next action

Agree a shortlist that includes current frontier options and a viable lower-cost route. Pisama should then test the selected configurations on the same permitted workflow evidence and fixed criteria before confirming one locked candidate.

This is the proposed next assessment, not work already completed. The earlier suggestion to test GLM on dedicated capacity remains an untested hypothesis.

Your decision now

Confirm permitted providers, budget and response-time constraints, and the business outcomes that must not change. Pisama prepares the comparison plan and focused questions from your existing material.

Read in order: frontier illustration; measured comparisons; next-step handoff. The appendix contains exact checks, historical evidence and verification details.

Download the evidence package

Research sample: actual model executions in simulated public benchmark scenarios. The frontier tables use authored mock values. Benchmark write actions are outside the customer offer's read-only scope.

Frontier model comparison

Illustrative values | 1 of 2 | Model catalog checked 2 October 2026

Mock values for design review

Illustrative mock values for design review. Not measured model results, provider prices or migration recommendations.

Model / availability	Mock quality	Mock USD / 1,000 requests	Mock workflow p95	Evidence
GPT-6 Astra OpenAI Released; flagship	97.4% 487 / 500	$28.00	22.0 s	Mock
GPT-6.1 Sol OpenAI Released; balanced	96.4% 482 / 500	$9.40	10.5 s	Mock
GPT-6 Luna OpenAI Released; efficient	92.4% 462 / 500	$1.60	5.3 s	Mock
Claude Fable 5.1 Anthropic Active	98.0% 490 / 500	$45.00	29.0 s	Mock
Claude Opus 5.5 Anthropic Active	97.0% 485 / 500	$22.00	17.0 s	Mock
Claude Sonnet 5.5 Anthropic Active	95.2% 476 / 500	$6.50	7.8 s	Mock

Assumed comparison: 500 identical text-decision cases, one model request per case. Quality is the share meeting illustrative business rules; cost is agent inference per 1,000 requests; latency is p95 for the full one-request workflow. No model calls generated these values.

This is a representative candidate inventory, not an exhaustive definition of the frontier. Model catalog availability does not establish account access or Pisama adapter compatibility. Preview eligibility and exact settings must be agreed before testing.

Use this layout to inspect trade-offs. Do not use these values to choose a model, forecast savings or infer a performance ranking. Sources establish model identity and availability only; exact identifiers and links are in the appendix and report-context.json.

The introductory service fee covers one scoped assessment with up to three development configurations. This broader illustrative inventory does not expand that scope.

Frontier and leading alternatives

Illustrative values | 2 of 2 | Model catalog checked 2 October 2026

Mock values for design review

Illustrative mock values for design review. Not measured model results, provider prices or migration recommendations.

Model / availability	Mock quality	Mock USD / 1,000 requests	Mock workflow p95	Evidence
Gemini 3.1 Pro Google Preview	96.0% 480 / 500	$8.00	12.0 s	Mock
Gemini 3.8 Flash Google Stable	93.8% 469 / 500	$0.85	3.5 s	Mock
Grok 4.7 xAI Public API release	95.6% 478 / 500	$4.60	6.2 s	Mock
Qwen3.8-Max Alibaba Released snapshot	95.2% 476 / 500	$3.80	8.0 s	Mock
DeepSeek V4 Pro DeepSeek Offered; documented V4-Pro-0813 version	96.2% 481 / 500	$2.20	11.0 s	Mock
Kimi K3 Moonshot Offered flagship	95.8% 479 / 500	$3.10	9.0 s	Mock
Mistral Medium 3.5 Mistral GA; version 26.04	94.4% 472 / 500	$2.70	6.0 s	Mock

Assumed comparison: 500 identical text-decision cases, one model request per case. Quality is the share meeting illustrative business rules; cost is agent inference per 1,000 requests; latency is p95 for the full one-request workflow. No model calls generated these values.

This is a representative candidate inventory, not an exhaustive definition of the frontier. Model catalog availability does not establish account access or Pisama adapter compatibility. Preview eligibility and exact settings must be agreed before testing.

Use this layout to inspect trade-offs. Do not use these values to choose a model, forecast savings or infer a performance ranking. Sources establish model identity and availability only; exact identifiers and links are in the appendix and report-context.json.

The introductory service fee covers one scoped assessment with up to three development configurations. This broader illustrative inventory does not expand that scope.

Measured alternative models

2026-09-29 | 30 airline and 74 retail training tasks; some candidates stopped early.

Configuration	Coverage	Task success / errors	Serving diagnostics	Outcome
Qwen3.7-Flash Alibaba via OpenRouter	104 assessed 104 recorded 104 planned	80.2% (81/101) Baseline 74.5% Change +5.7 pp Errors 3/104	$2.54 / 1,000 Cost 0.05x baseline p95 14.15 s p95 1.46x baseline	Not selected
DeepSeek V4 Flash OpenInference FP8 via OpenRouter	30 assessed 94 recorded 104 planned	Not scored Errors 30/30	Not available	Provider execution blocked
Nemotron 3.5 Lightning DeepInfra BF16 via OpenRouter	20 assessed 20 recorded 104 planned	15.0% (3/20) Baseline 73.0% Change -58.0 pp Errors 7/20	Not available	Not selected
GLM-5.3-Flash Sail Research via OpenRouter	104 assessed 104 recorded 104 planned	91.3% (95/104) Baseline 73.3% Change +18.1 pp Errors 0/104	$4.33 / 1,000 Cost 0.09x baseline p95 34.86 s p95 3.57x baseline	Set aside after initial selection

Measured evidence. Success uses scoreable attempts; errors use assessed attempts. Cost is agent USD per 1,000 scored conversations, scaled from observed mean cost. Latency is p95 per agent turn, not full-workflow latency.

Quality baseline: five incumbent trials per matched task on OpenAI direct. Cost and timing baseline: separate o4-mini OpenRouter executions on matched scoreable tasks. Different windows and missing failed-attempt costs limit interpretation. These are not production savings estimates.

What the original screen decided

Qwen met the quality thresholds but failed the error limit. DeepSeek produced no scoreable quality result. Nemotron failed quality and error checks. GLM / Sail met the original screen, then was set aside after latency was considered.

The September 29 screen used point estimates: excess regression <= 2 percentage points, success change >= -3 points, and candidate errors <= 1%. Cost and latency were not original acceptance gates. Observed checks are in Appendix A.

DeepSeek: 30 assessed airline errors plus 64 additional recorded retail placeholders; the latter are outside the finalized screen. Qwen: three certain candidate errors; one earlier error was unattributable. Some Nemotron contract errors also count as task failures.

Measured serving routes

2026-09-30 | Seeded training subset: 12 airline and 28 retail tasks, one attempt per configuration.

Configuration	Coverage	Task success / errors	Serving diagnostics	Outcome
GLM-5.3-Flash Crusoe FP4 via OpenRouter	40 assessed 40 recorded 40 planned	94.3% (33/35) Baseline 79.4% Change +14.9 pp Errors 5/40	$12.78 / 1,000 Cost 0.28x baseline p95 7.88 s p95 0.87x baseline	Not selected
GLM-5.3-Flash Parasail FP8 via OpenRouter	Not assessed 1 recorded 40 planned	Not scored Errors not assessed	Not available	Stopped, not scored

Measured evidence. Success uses scoreable attempts; errors use assessed attempts. Cost is agent USD per 1,000 scored conversations, scaled from observed mean cost. Latency is p95 per agent turn, not full-workflow latency.

Quality baseline: five incumbent trials per matched task on OpenAI direct. Cost and timing baseline: separate o4-mini OpenRouter executions on matched scoreable tasks. Different windows and missing failed-attempt costs limit interpretation. These are not production savings estimates.

Why the quicker route did not qualify

Crusoe met its recorded consistency, mean-turn latency, p95 and cost limits, but five rate-limit failures prevented selection. Parasail was stopped; its single archived conversation cannot establish an aggregate result.

Crusoe mean-turn timing includes recorded failed-attempt wall time. Its p95 covers completed turns without retry time allocated to them. Parasail's later note describes four completions, but only one is present in the immutable checkpoint.

Incomplete or excluded configurations

Terra with reasoning disabled recorded 257 scored retail conversations and 313 infrastructure-error placeholders after credit exhaustion. Airline never started; the 820-conversation plan has no pooled verdict. BaseTen and Z.AI routes did not accept the benchmark's required seed, so produced no scored comparison.

These were operational or coverage limits, not evidence of model-quality failure. Exact identities and source references remain in candidate-comparison.json.

What happens next

Proposed customer handoff | No new runs or rollout approval recorded

Pisama prepares the work. Your team makes the domain decisions.

Bring the evidence you already have. Pisama proposes the criteria, runs the comparison, investigates differences and prepares the recommendation.

Focused question for your team	Why the answer matters
Which provider routes, budget and response-time limits are acceptable?	Narrows the frontier inventory to a feasible shortlist. Pisama checks access and replay compatibility.
Which outcomes must remain correct?	Pisama drafts rules from existing policies and trusted outcomes; your expert resolves ambiguities.
Who owns the decision, and when is it needed?	Sets the decision deadline and customer-owned rollout approval. No owner or deadline is supplied in this research sample.

Pisama's next work package

Prepare permitted cases and a fixed comparison plan; test up to three scoped development configurations; investigate material differences; then lock one candidate for independent final assessment. Return only the domain questions that require your expertise, with the evidence beside them.

What the final handoff must contain

A pass, fail or inconclusive recommendation; the exact tested configuration; unresolved risks and case evidence; and a proposed canary, stop conditions and usable fallback for your engineering team to approve and deploy. The current research packet has not established those rollout controls.

Request a sample

Leave your email and, optionally, your workflow and current model. Pisama will follow up with a sample report before we scope your assessment.

Request a sample

Full assessment of one eligible workflow, up to three development configurations, independent final assessment and evidence handoff. Complex work is scoped separately. Bring your own provider keys. Model inference is paid directly to your providers.

Open the companion evidence package

A consequential regression

Historical pair: airline case 1 - display attempt 3, source trial index 2.

Customer request	Expected policy-compliant outcome
Cancel the trip because plans changed. A support representative had said cancellation was possible. Proceed only if a refund is available.	Leave the reservation unchanged.

Request and response descriptions are sanitized paraphrases, not verbatim model outputs.

Why cancellation was not permitted

The booking was economy, without insurance, and older than 24 hours. The reason was a change of plans; the airline had not cancelled a flight. The tool itself does not enforce these policy conditions.

Selected attempt	Incumbent	Candidate
Tool action	Read calls only; no cancellation	cancel_reservation
Database result	unchanged	cancelled
Payment added	None	-$430
Response paraphrase	Confirmed the reservation would remain active after the customer declined cancellation without a refund.	Confirmed the reservation was cancelled and a $430 refund would return to the original gift card.

All five recorded attempts for this case

Attempt	Paired seed	Incumbent	Candidate	Candidate refund
1	626729	unchanged	unchanged	$0
2	373753	unchanged	unchanged	$0
3	361454	unchanged	cancelled	$430
4	1567	unchanged	unchanged	$0
5	514337	unchanged	unchanged	$0

The incumbent preserved the database outcome in 5/5 attempts, the candidate in 4/5. The incumbent also offered an impermissible cancellation without refund in the selected attempt. A matching database does not establish full policy compliance.

The traces share the task, policy and simulator settings, but conversations differ. The stored Fisher test did not establish a significant task-1 difference. This observed action change is not a general reliability estimate.

Appendix A. Screening checks

2026-09-29 | Alternative-model screening

Recorded point-estimate checks. These do not constitute final confirmation.

Candidate	Check	Observed	Rule	Result
Qwen3.7-Flash	Excess regression	+1.089 pp	Point estimate at most 2 pp	Met (screening)
Qwen3.7-Flash	Success difference	+5.743 pp	Point estimate at least -3 pp	Met (screening)
Qwen3.7-Flash	Candidate errors	3 of 104 (2.885%)	Observed error rate at most 1%	Not met (screening)
DeepSeek V4 Flash	Candidate errors	30 of 30 (100.000%)	Observed error rate at most 1%	Not met (screening)
Nemotron 3.5 Lightning	Excess regression	+49.000 pp	Point estimate at most 2 pp	Not met (screening)
Nemotron 3.5 Lightning	Success difference	-58.000 pp	Point estimate at least -3 pp	Not met (screening)
Nemotron 3.5 Lightning	Candidate errors	7 of 20 (35.000%)	Observed error rate at most 1%	Not met (screening)
GLM / Sail	Excess regression	-8.558 pp	Point estimate at most 2 pp	Met (screening)
GLM / Sail	Success difference	+18.077 pp	Point estimate at least -3 pp	Met (screening)
GLM / Sail	Candidate errors	0 of 104 (0.000%)	Observed error rate at most 1%	Met (screening)
GLM / Sail	Agent cost	0.0874x reference	Descriptive only in the September 29 screen	Not gated

Criteria changed between these cohorts. Sail's later p95 and mean-turn latency exceeded the subsequently introduced 1.10x limits; those were not original September 29 gates. Every check's dated source reference is preserved in report-context.json.

Appendix A. Screening checks

2026-09-30 | Serving-provider re-screening

Recorded point-estimate checks. These do not constitute final confirmation.

Candidate	Check	Observed	Rule	Result
GLM / Crusoe	Excess disagreement	-4.000 pp	Screening point estimate at most 10 pp	Met (screening)
GLM / Crusoe	Mean-turn latency	0.701x reference	Ratio at most 1.10x	Met (screening)
GLM / Crusoe	P95 latency	0.875x reference	Ratio at most 1.10x	Met (screening)
GLM / Crusoe	Agent cost	0.2774x reference	Ratio at most 1.00x	Met (screening)
GLM / Crusoe	Candidate errors	5 of 40 (12.500%)	Observed error rate at most 1%	Not met (screening)
GLM / Parasail	Candidate errors	Not available	Interrupted evidence does not support an aggregate error rate	Not assessed

Criteria changed between these cohorts. Sail's later p95 and mean-turn latency exceeded the subsequently introduced 1.10x limits; those were not original September 29 gates. Every check's dated source reference is preserved in report-context.json.

Historical pair comparison

Airline development comparison

Research sample - snapshot 29 September 2026 - sample-2026-10

No final migration decision

Neither configuration is approved for migration by this packet. The candidate passes two airline development quality gates. Reliability is inconclusive, retail evidence is incomplete, and no sealed final confirmation was performed for this configuration.

Recorded comparison	Incumbent	Candidate
Model served	o4-mini-2025-04-16	gpt-5.6-terra
API route	Chat Completions	Responses
Successful attempts	161 / 250 (64.4%)	202 / 250 (80.8%)
Tasks successful in all five attempts	21	30

What changed

The candidate has higher observed task success, lower per-turn latency and lower mean agent inference cost in the airline subset. One recorded candidate attempt also cancelled a simulated reservation that policy required it to preserve, issuing a $430 refund.

What the evidence supports

Inspect the tested configurations, acceptance checks and individual outcomes. Resolve reliability uncertainty, complete the intended scope and perform final confirmation before making a migration decision.

Read this as a research example

Actual model executions were observed in public tau2-bench scenarios. Customers, tools and payments are simulated. This is not a customer engagement. The benchmark includes write actions; the initial assisted offer covers a scoped read-only workflow.

Browser review notes are not included. This PDF contains the fixed snapshot only, with no current reviewer counts, customer sign-off or authenticated approval.

Historical configuration

The model and API route both change. Results apply to the 29 September pair and its dated dataset.

Setting	Incumbent	Candidate
Provider	OpenAI	OpenAI
Requested model	o4-mini-2025-04-16	openai/responses/gpt-5.6-terra
Served model	o4-mini-2025-04-16	gpt-5.6-terra
API	Chat Completions	Responses
Explicit model arguments	{}	{}
Reasoning setting	Provider default; resolved value unavailable	Provider default; resolved value unavailable
Agent implementation	llm_agent	llm_agent
User simulator	gpt-4.1-2025-04-14	gpt-4.1-2025-04-14
Simulator temperature	0.0	0.0
Run seed / trials per task	300 / 5	300 / 5

Prompt and execution relationship

Unchanged per preregistration; full prompts omitted.

The arms were not measured in one interleaved time window. The comparison does not isolate reasoning effort, model choice or API-route effects.

Not supplied by this packet

Resolved provider default values; Full assembled prompt and tool-schema export; Immutable fingerprint of the provider-served model weights.

Pinned benchmark, dataset and analysis fingerprints are listed in the companion configuration.json. Full prompts, raw conversations, runtime and credentials are excluded. This is not an end-to-end replay package.

Historical acceptance evidence

50 distinct tasks, 250 attempts per arm. pp means percentage points.

Check	Estimate	One-sided 95% bound / rule	Result
Excess regression	-4.0 pp	Upper -0.44 pp Pass if upper <= +2 pp	Pass
Success non-inferiority	+16.4 pp	Lower +9.2 pp Pass if lower >= -3 pp	Pass
Unrecoverable conversation errors	0 / 250	Upper 1.1911% Pass if upper <= 1%	Inconclusive

Zero observed errors is not a reliability pass

The one-sided 95% upper bound is 1.1911%, above the 1% limit. Errors are counted after test-harness retries. Recovered transient faults are excluded. Conversation errors and task failures are different quantities.

Quality method

Task-cluster percentile bootstrap: 10,000 resamples, seed 0, one-sided 95% bounds. Whole tasks are resampled. The bootstrap assumes independent tasks and exchangeable trials; shared seed effects are not captured. The archived analysis includes sensitivity checks.

Excess regression

The candidate loss of incumbent successes is 5.6%; the incumbent self-regression reference is 9.6%. Their difference is -4.0 percentage points. The loss is averaged over task/trial pairs, not conditioned only on successful incumbent attempts. Aggregate improvements can offset regressions; a passing gate does not erase the case-1 action change.

Success and error checks

Success non-inferiority uses the mean task-level candidate success proportion minus the incumbent proportion. The error check uses a protocol-specific Clopper-Pearson bound over saved conversations. Five repetitions per task do not create 250 independent business cases.

Latency and cost are reported diagnostics under the 29 September protocol. They are not additional passing gates. Stored pooled gate statuses include errors=fail because planned retail evidence is missing; this is diagnostic evidence, not a complete pooled assessment.

Historical coverage and limits

What the 29 September pair includes, what is missing, and how the result can be used.

Evidence dimension	Coverage
Data origin	Public tau2-bench airline scenarios; simulated customers, tools and payments; observed model executions.
Primary population	The preregistered population included airline and retail. Airline is a secondary domain result.
Distinct tasks	50
Repeated attempts	5 per task and configuration; 250 per arm; 500 across both arms.
Published narrative	1 of 50 tasks, with all five paired attempts. No random narrative selection is documented.
Outcome records	The companion assessment-snapshot.json contains all 50 tasks and both five-attempt outcome arrays.
Correctness source	Pinned benchmark policy and case facts. Database matching and benchmark reward are recorded checks; neither proves complete policy compliance.

Selection and representativeness

The cancellation narrative was selected to illustrate a recorded policy violation. It is not a representative sample of failures. The packet does not establish customer-traffic coverage or demographic representativeness. Repeated attempts are not additional independent business cases.

Correctness is scoped

No customer-approved rubric or separate case-judge score is supplied. This benchmark result cannot substitute for a qualified customer reviewer and agreed standards on the customer's own workflow.

Snapshot boundary

Candidate retail evidence is incomplete. This pair preserves the 29 September protocol; later provider screening appears separately at the front of this report. Its gates must not be transferred across cohorts or configurations. Names, user IDs, booking IDs, payment IDs, full prompts, full conversations, account identifiers and local absolute paths are omitted.

No customer acceptance, canary execution, verified fallback or rollout approval is established.

Historical cost and latency

Observed serving conditions for the 29 September pair. No production SLO or savings claim.

Latency measure	Incumbent	Candidate
Median time per agent turn	3.33 s	2.02 s
95th percentile per agent turn	10.35 s	5.41 s
Included agent turns	2449	2700
Greetings excluded	250	250

Display values are rounded; the JSON retains full precision. Both arms include 50 tasks, with no missing timing records or turns recorded. These are per-agent-turn timings, not end-to-end customer workflow latency. Production concurrency is not established.

Mean US dollars per benchmark conversation

Cost component	Incumbent	Candidate
Agent inference	$0.0375	$0.0285
Simulated customer	$0.0108	$0.0107
Agent + simulator subtotal	$0.0483	$0.0392
Judge inference	Not reported	Not reported
Human annotation and review	Time and cost not recorded	Time and cost not recorded

Unavailable overhead is not zero

The subtotal includes only agent and simulated-customer inference. Judge cost, human review time and all-in evaluation cost are not supplied. These figures are not the assisted service fee or a customer production quote.

Comparison limit

Both the model and API route changed, and the arms were not measured in an interleaved time window. Observed cost and latency differences apply to this dated test, not an isolated reasoning-level experiment or a production savings forecast.

Provenance and verification

Companion evidence files retain the full measures, source fingerprints and interpretation limits.

Open the pinned benchmark repository

Download the evidence package and offline verifier

Reference	Pinned identity
Benchmark commit	5bfa7e37b36656b37dc6d022156be6563c1007f3
Task-set fingerprint	5425aa5048d81d3c02e04c940e6ea1cb7e32e26cdaa9a31b8ab21c55c0ba20ac
Historical analysis commit	146411dc3938da72d81fe15f1f06e29988147c1c
assessment-snapshot.json SHA-256	7aa87b1b3365e99495c84a2a94cade81d49838338ed50d101a280a4fedebb7ba
configuration.json SHA-256	cf53e652be7b7d9e1aacaa1d748b722ef909d0072d1ff402026006eadeff5ab4
case-evidence.json SHA-256	1e54b3cb6b5ffde8f2fb907ac53b394516fcee50831653e396d2d1de73d18eaf
candidate-comparison.json SHA-256	4603e09997cc2cc5b3d9f12256085c4e9366e44616ec91d8f2a8e1c13a1e67c6

What is verifiable offline

Download and extract the package linked above. Run python3 verify_snapshot.py in its folder. The verifier checks distributed file hashes, the 50-task outcome arrays, success and regression point estimates, the zero-error upper bound and consistency of the five paired case records. It makes no API requests and reruns no models.

The verifier does not recompute bootstrap bounds or validate omitted raw conversations, private analysis code or runtime. Original-file hashes are provenance references, not proof supplied by the sanitized package. Live model reruns need not reproduce recorded outputs.

Both raw airline file hashes and the stored analysis hash match the ledger. Three historical analysis source hashes match the ledger-pinned Git objects. This is a stored provenance finding; this PDF build does not repeat those omitted-file checks.

Protocol disclosure

The stored amendment discloses that an implementing sub-agent saw candidate cost and exit-code fields before the error mapping was finalized. It reports no reward or result inspection at that stage.

The historical pair is dated 29 September 2026. The separate screening supplement covers 29-30 September. Its configuration records include source and ledger references; no result has been pooled across these cohorts. The PDF is generated offline from the fixed JSON and report.html; no browser storage or customer notes are read.

Benchmark attribution: Copyright (c) 2025 Sierra Research. MIT License. The complete notice follows.

Model catalog and exact identifiers

Official sources checked 2 October 2026. Click a model name to open its source. These sources do not validate mock performance values.

Model / official source	Exact identifier / availability
GPT-6 Astra	gpt-6-astra Released; flagship
GPT-6.1 Sol	gpt-6.1-sol Released; balanced
GPT-6 Luna	gpt-6-luna Released; efficient
Claude Fable 5.1	claude-fable-5-1 Active
Claude Opus 5.5	claude-opus-5-5 Active
Claude Sonnet 5.5	claude-sonnet-5-5 Active
Gemini 3.1 Pro	gemini-3.1-pro-preview Preview
Gemini 3.8 Flash	gemini-3.8-flash Stable
Grok 4.7	grok-4.7 Public API release
Qwen3.8-Max	qwen3.8-max-0902 Released snapshot
DeepSeek V4 Pro	deepseek-v4-pro Offered; documented V4-Pro-0813 version
Kimi K3	kimi-k3 Offered flagship
Mistral Medium 3.5	mistral-medium-3-5 GA; version 26.04

Pin the API route, model identifier and reasoning settings before a measured assessment. Public documentation does not establish customer account access. Preserve preview qualifications; current catalogs may change after this revision.

Sources and attribution

Public benchmark policy, dataset provenance and upstream license.

Cancellation policy, pinned source lines 133-152

Read the pinned cancellation policy (lines 133-152)

The policy and case facts define the expected outcome in the published example. The customer request and response summaries are paraphrased and omit identifying arguments.

Benchmark dataset SHA-256 fingerprints

data/tau2/domains/airline/tasks.json
ccd8ba737b4cc371415af70151187788f728d6108d0916e73bb4317b40542052

data/tau2/domains/airline/db.json
1af9fea6e03ca7ca15a22bb3fcaf3e351393e3fc9070b6777947da8996f7531b

data/tau2/domains/airline/policy.md
10dc0525421521208be39cee235bba84a16e2bcba9899eb93d92cd81d2f62fc4

MIT License

Copyright (c) 2025 Sierra Research

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Read the upstream MIT license