

ORLO / OPEN EVALS · SUPPORT ROUTING

What a score would have hidden: two support routing runs on Orlo

Two Orlo comparisons of GPT-5.4 and Claude Sonnet 4.6 on synthetic support tickets. Neither ended in a release. What they produced instead were problems a person could inspect, each one tied to the tickets, answers and rules behind it: an overlap in the written policy, replies that broke the answer format, escalations past a written rule, and one scenario family behind every miss by one model.

Both studies use a fictional software company's policy and synthetic tickets whose reference answers have not received independent human review. Counts measure agreement with those references, not verified production accuracy, and neither study identifies today's best model. The 500-ticket run used different cases and a different policy from the pilot. It is separate evidence, not a measured improvement caused by the policy changes. No release was created in either study.

34	11	6	0
Claude Sonnet 4.6 replies that broke the JSON contract, kept as failures	Valid answers that escalated after one earlier contact, where the policy requires two	GPT-5.4 misses, all from one scenario family	Releases on provisional labels and thin evidence

WHY THIS MATTERS

Had a team picked a model from a headline score, whichever one it picked would have brought one of these problems with it: format failures and escalation drift with one, a repeated priority reading with the other. The policy overlap would have stayed in the policy. Orlo kept each problem attached to the ticket, the answer and the rule behind it, matched every grade against an independent replay, and released nothing while the labels wait for a person to review them.

What is in this case study

Part 1, October 2026. The 500-ticket comparison on dataset 0.3.1: 1,000 answers, 91 scenario families, and what inspecting the saved answers showed.

Part 2, October 2026. The 20-ticket pilot that came first, where both models matched 19 of 20 references and disagreed with the same queue label, and Orlo reported ● **More evidence needed.**

What 500 support tickets revealed about answer format and routing decisions

A completed Orlo comparison produced 1,000 answers from two models. GPT-5.4 matched 494 of 500 provisional references; Claude Sonnet 4.6 matched 452. Inspecting the saved answers separated three issues: replies that broke the JSON contract, escalation at the wrong contact boundary, and a repeated priority disagreement in one scenario family. Every Orlo grade matched an independent offline replay.

This study ran in October 2026 against [dataset 0.3.1](#). The tickets and reference answers are synthetic and have not received independent human review. Counts below measure agreement with those references, not verified production accuracy.

The result, with its denominators

Measure	GPT-5.4	Claude Sonnet 4.6
Provider calls completed	500/500	500/500
Answers meeting the JSON schema	500/500	466/500
Whole answers matching, development	400/400	363/400
Whole answers matching, evaluation	94/100	89/100
Whole answers matching, all cases	494/500	452/500
Families where every variant matched	90/91	61/91
Unnecessary escalation in valid answers	0/349 negative references	11/349 negative references
Positive references with a usable escalation answer	151/151	133/151

There are **91 scenario families**, not 500 independent situations. Each family stays in one partition: 71 families in development and 20 in evaluation. Both partitions are public, and both ran with the same frozen instructions and settings. We did not tune between them or run a release test.

Whole answers matching provisional references, by reference queue

■ GPT-5.4 ■ Claude Sonnet 4.6



500 synthetic tickets, 91 scenario families, dataset 0.3.1, all cases. These bars measure all three fields together within each reference queue; they are not queue-only accuracy.

34 replies failed the answer format

Claude returned 34 replies that could not be parsed as a single JSON document. Some included reasoning or Markdown around an otherwise plausible answer. For example, `support-028` contains an active lockout plus a routine security question. Its reply explained the distinction and included the three reference values, but added prose before the JSON.

The declared contract required JSON alone. We retained those failures instead of extracting a convenient fragment after seeing the result. All 18 of Claude's positive-reference misses came from these invalid replies; they are not 18 valid `escalate:false` decisions. Another 16 invalid replies occurred on negative references. On the evaluation partition, five of 32 positive references lacked a usable escalation answer for this reason.

This run asked for JSON in the instructions. It did not enable provider-native schema-constrained generation, repair malformed answers, or retry them. A follow-up can test those choices under a newly frozen configuration. This result describes the tested setup, not the best achievable configuration for either model.

One earlier contact exposed an escalation boundary

Of Claude's 466 valid answers, 11 escalated tickets whose `prior_contacts` value was one. These span several queues. The policy requires at least **two earlier contacts**, excluding the current ticket; customer tier does not change that threshold.

`support-155` is one example. It says the request remains unresolved after one earlier contact and explicitly excludes the current ticket. Claude returned `escalate:true`; the reference is `false`. These are usable JSON answers that disagree with the written boundary, a different problem from invalid output.

The other three valid disagreements were queue choices in the planned IP-allowlist family (`support-105` , `106` and `109`). Claude chose Technical; the reference is Security, under the rule for security-purpose how-to questions.

Six priority disagreements were one family

All six GPT-5.4 disagreements belong to `scenario-expired-invite` , cases `support-153` through `158` , in the evaluation partition. Each describes someone who cannot enter a workspace or begin assigned work because the invitation has expired. The references assign high priority under the blocked-login/core-workflow rule; GPT-5.4 chose normal. Queue and escalation matched in every case.

That is one repeated interpretation across six related variants, not six independent discoveries. A domain reviewer should confirm whether the blocked-work wording is sufficient, adjudicate the reference, and add distinct access scenarios before drawing a broader conclusion.

Consistency alone also misses this issue: GPT-5.4 preserved all 390 declared relationships between variants, including consistent priority choices that disagreed with this family's references. The scorer reports consistency and correctness separately for this reason.

What the Orlo workflow established

The task and dataset were saved before execution. Provenance tied each sample to its case ID and artifact hash. The comparison completed within its declared limits, all pages could be collected through a scoped runtime key, and independent scoring matched all 1,000 native grades. A separate tutorial acceptance task imported the same 500 examples without redaction and passed preflight.

The run took approximately **19 minutes**. Both models combined reported **734,737 tokens**: 711,783 input and 22,954 output. Recorded token-price estimates totalled **about \$2.31**, reconciled between the comparison budget and call ledger. That is an estimate, not an audited provider invoice. The ledger records 500 successful calls for each model, with no failed or additional calls for this comparison.

No release was created. The task's required release policy remained in place; an ordinary comparison on provisional labels does not supply a passing release-test verdict.

Reproduce the grades

From the repository root, with Node.js 20 or later:

```
node --test verify-expanded.test.mjs
node cli.mjs score studies/support-routing-2026-10-17/answers-openai.json --dataset . --split evaluation
node cli.mjs score studies/support-routing-2026-10-17/answers-anthropic.json --dataset . --split
evaluation
```

These commands make no model calls. The test replays development, evaluation and full-dataset scores for both models.

Methods and limits

We chose the same requested models as the earlier pilot for continuity across two direct provider paths, not to rank the newest models. The provider-returned IDs were `gpt-5.4-2026-03-05` and `claude-sonnet-4-6`. Each model answered each case once. There was no judge, embedding grader, retrieval, early elimination or tuning between partitions.

The output limit was 4,096 tokens for both providers. OpenAI received `max_completion_tokens` with no temperature, reasoning-effort or seed parameter. Anthropic received `max_tokens` and temperature zero, with no thinking, effort or seed parameter. These settings are reconstructed from the caller used by the run; full outbound request bodies were not archived. Omitted settings are not measured provider defaults, and fresh generations can differ.

The plan and artifact hashes were saved before queueing. They were not independently timestamped as a public preregistration. The manifest records that limit and the recorded usage. Rich diagnostics use the previously declared scoring rules; the interpretation of individual disagreements above is a post-run inspection.

The proposed priority-loss score was 24 relative error points for GPT-5.4 and 272 for Claude. Invalid answers receive the maximum per-case penalty; all of Claude's priority loss came from invalid output. These weights are assumptions, not dollar costs or operationally calibrated harm estimates.

We report descriptive counts without significance tests or confidence intervals that assume independent rows. Synthetic templates, visible references, related variants and unreviewed labels limit generalization. GPT-5.4 had more matching answers in this setup; this does not establish a general model winner or production suitability.

The historical 20-case study in Part 2 used different cases and a different policy. This expanded run is separate evidence, not a measured improvement caused by the policy changes. Human adjudication and a separately designed follow-up remain the next steps.

Two models disagreed with our label on the same support ticket

An Orlo comparison returned 40 valid JSON answers. Both models matched 19 of 20 provisional references and disagreed with the same queue label. Inspecting that ticket exposed an overlap in our written policy. The next step is to clarify the policy and review the label before running another comparison.

This synthetic study ran in October 2026. It uses a fictional software company's policy and AI-authored labels that have not received human review. It demonstrates an evaluation workflow; it does not measure production accuracy or identify today's best model.

What we tested

Each ticket asks for a queue, a priority and an escalation decision. The policy covers security, access, technical and billing questions, including urgency and repeated contacts. It instructs the model not to raise priority because of customer tier or demands for urgency. The small dataset does not establish that models follow that rule in every situation.

There are 20 tickets in ten related scenario families. Fourteen are marked development. The other six, labelled regression in the original artifact, cover three different families: unclear requests, integrations and subscriptions. They provide a family-separated check, but all labels were visible and all 20 cases were evaluated together. There was no unseen holdout.

We selected GPT-5.4 and Claude Sonnet 4.6 from the models available through the local setup's direct provider connections to exercise the same workflow across two providers. This was a local integration pilot, not a survey of the latest models. Each model answered each ticket once. The exact returned model IDs and available settings are recorded in the [methods appendix](#).

An answer matched only if all its JSON values equalled the provisional reference. Key order and whitespace did not affect the grade. No model judge, embedding grader or retrieval was used. We independently re-scored the saved answers and reproduced all 40 grades.

Measure	GPT-5.4	Claude Sonnet 4.6
Answers returned	20/20	20/20
Valid JSON	20/20	20/20
Whole answers matching the reference	19/20	19/20
Queue matching the reference	19/20	19/20
Priority matching the reference	20/20	20/20
Escalation matching the reference	20/20	20/20

Both models combined used 21,920 tokens, at an estimated model cost of **about \$0.07**. That is Orlo's estimate from recorded token usage and pricing, not a provider invoice.

Why the shared disagreement matters

Case `support-008` asks how to rotate API keys during routine maintenance. There is no active compromise or blocked workflow. Our provisional reference routes it to Security. Both models routed it to Technical, while matching the reference's normal priority and no-escalation decision.

Case support-008: rotating API keys during routine maintenance			
	Queue	Priority	Escalation
Provisional reference	Security	Normal	No
GPT-5.4	Technical	Normal	No
Claude Sonnet 4.6	Technical	Normal	No

The original policy contains both of these clauses:

- security for security incidents or security questions
- technical for product/API/integration issues and product how-to questions

It orders active incidents and blockages, but does not explicitly settle which clause wins for a routine security-related API how-to. Case `support-011`, a CSV-export question, is labelled Technical. The rationale for `support-008` explains why the ticket is not an incident, but does not explain why Security takes precedence over Technical.

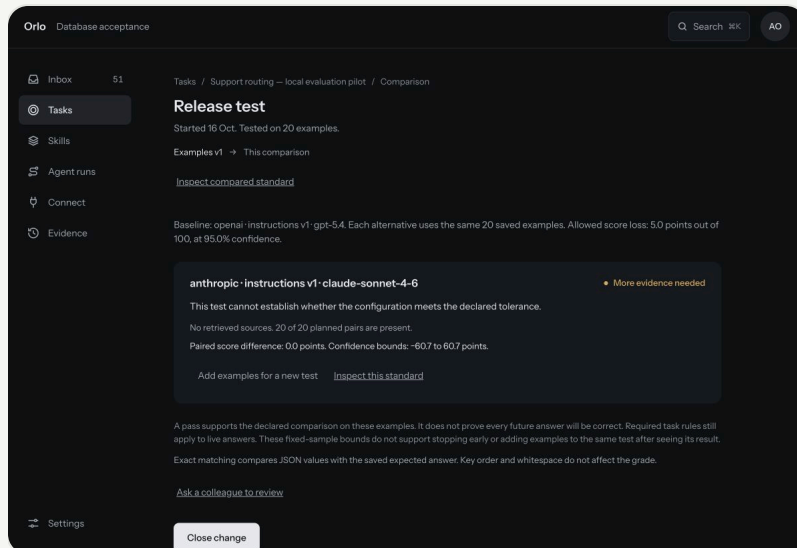
That makes this a policy-and-label review question. The saved answer disagrees with our reference; that alone does not establish that the model made a domain error. Nor does agreement between two models establish that the label is wrong.

We preserved the original policy, labels and answers, documented the overlap, and prepared a separate draft with explicit precedence and additional boundary cases. That draft still needs human review and a new comparison. We have not measured an improvement.

What the comparison could tell us

It could not distinguish the models on these cases. There were **zero tickets where one model matched the reference and the other did not**. Equal observed scores do not establish equivalence on future tickets, and this small pilot supplied no observed advantage for either model.

Orlo reported ● **More evidence needed**. The task also required at least 100 examples before release; this pilot had 20. Neither model was released from this experiment. Adding near-duplicate tickets just to reach the count would not establish broader coverage.



The comparison remains inconclusive. Orlo's release test against the GPT-5.4 baseline: paired score difference 0.0 points, confidence bounds of -60.7 to 60.7 points.

What a team can take from this

An evaluation can reveal an assumption in the task definition as well as a model failure. Here, the useful output was an inspectable disagreement: the original ticket, both model answers, the provisional label and the policy clauses behind it.

The next decision is concrete:

1. Have a domain owner resolve the queue overlap.
2. Review the affected references.
3. Test different cases under the revised policy.

Keep the earlier result so the reason for the change remains visible.

Check it yourself

500 tickets. [OpenAI answers](#), [Anthropic answers](#), [score reports](#), [frozen plan](#) and [run manifest](#).

20 tickets. [The cases](#), [the saved answers](#), [the dataset card](#) and the [methods appendix](#).

Run it yourself. The [support routing tutorial](#) in the Orlo docs repeats the 500-ticket comparison on the same dataset, against Orlo's hosted API or your own deployment. Every study is at github.com/useorlo/open-evals.

Published in Orlo's open-evals repository. The data and articles are CC BY 4.0; tools are MIT licensed. Reference labels remain unreviewed. Orlo is a fineSample product. hello@finesample.com