L **Larkspur Outbound (sample)**

TEST REPORT · TRIAL 1

Clay ICP tiering: which model should score accounts?

Score one company from an enriched Clay table against our client's ideal customer profile, so the table only spends credits finding contacts at good-fit accounts. The client sells a shift-scheduling and…

25 test cases · 3 prompt and model combinations · 2 blind graders

November 17, 2026 · prepared by Larkspur Outbound (sample)

✓ Best prompt and model

Ship deepseek-v4.1-flash with prompt v1.

9.1
out of 10, likely range 8.7 to 9.4

$0.57
per 1,000 rows

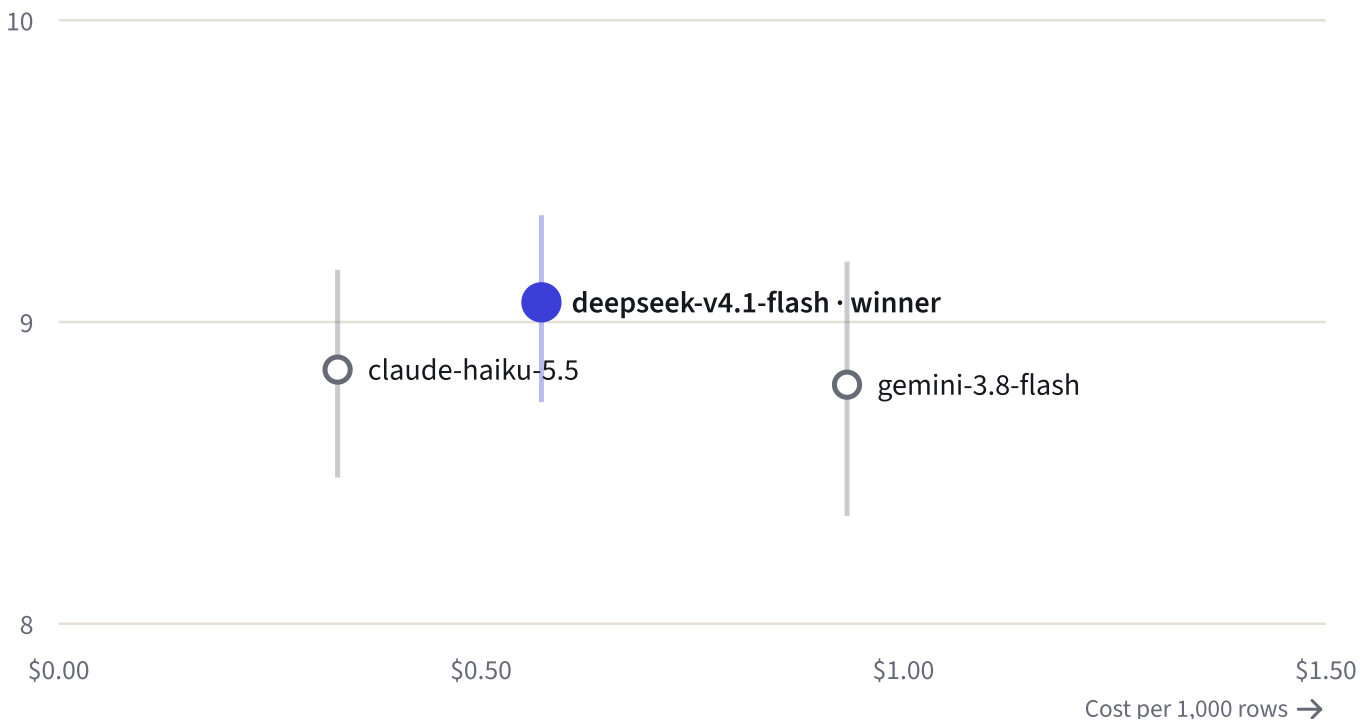
Its best answer · scored 10.0 out of 10 on one test case

"Tier 1 Food Manufacturing with 340 employees across 4 plants and most staff on rotating production shifts."

▲ **Too close to call:** deepseek-v4.1-flash and claude-haiku-5.5 are only 0.2 apart, too small to be sure of: on other test cases like these, the order could swap. Either is a fair pick; claude-haiku-5.5 costs less.

Results: quality against cost

Every candidate faced the same 25 test cases. Up is better, left is cheaper. Vertical lines show each score's likely range; where ranges overlap, the difference may not be real.



Scores by dimension

Each part of the grading guide on its own scale; overall out of 10. ⚑ marks a jury split.

Candidate	Overall	Tier correctness	Conservative handling	Exact two-line	Reason cites	Reason length	Grounding in	$/1k rows
`deepseek/deepseek-v4.1-flash` **Winner**	**9.1**	4.3 ⚑	4.6	5.0	4.6	4.5 ⚑	4.8	0.57
`anthropic/claude-haiku-5.5`	**8.8**	4.1 ⚑	4.6	4.9	4.8	4.0 ⚑	4.7	0.33
`google/gemini-3.8-flash`	**8.8**	4.3 ⚑	4.5	5.0	4.3	4.8	4.5	0.93

Too close to call: deepseek-v4.1-flash and claude-haiku-5.5 are only 0.2 apart, too small to be sure of: on other test cases like these, the order could swap. Either is a fair pick; claude-haiku-5.5 costs less.

What to test next

What the prompt didn't tell each model, drawn from the graders' reasons on its lowest-scoring answers. Each fix is a new prompt version: re-run the same test cases and keep it only if the score goes up.

deepseek-v4.1-flash saved as prompt v2

Most deductions come from the reason line not admitting doubt the row itself shows. In #1 the description mixes in salaried engineers and quality managers, and in #5 the industry column (Hospitality) disagrees with the description (staffing), with 220 employees against 1,800 placed workers. Both times the model said 'all Tier 1 criteria met' instead of…

```
+ Give Tier 1 only when every Tier 1 criterion is stated clearly and nothing in the row qualifies it; if the description mixes in salaried staff, or the industry column disagrees with the description, name that column, step down to Tier 2, and say in the reason which value decided the tier.
```

claude-haiku-5.5 saved as prompt v3

Main failures: (1) staffing firms placing hourly temps were wrongly excluded under 'agency' (#1). 'Agencies' in the ICP means marketing and creative agencies with no hourly workforce. A staffing firm managing hourly temps has hourly-workforce evidence, and healthcare staffing is even a Tier 1 industry, so staffing firms should be tiered and not excluded…

```
+ Staffing and recruiting firms that place or manage hourly temps are not the 'agencies' exclusion (that means marketing/creative agencies); tier them on headcount, sites and hourly evidence.
```

gemini-3.8-flash saved as prompt v4

The recurring failure is that the model treats the presence of some hourly or shift staff as proof that most staff are hourly. Cases #1, #10 and #9 show this, and #5 conflates placed workers with the company's own employees. The prompt never says that a mix of hourly and salaried staff, or workers placed at clients, leaves the hourly share unconfirmed and…

```
+ 'Most staff hourly' needs explicit row text saying so; if the row mentions salaried staff alongside hourly or shift staff, gives no share, or describes workers placed at clients rather than the company's own employees, the hourly share is unconfirmed: assign at most Tier 2 and say so.
```

How this was tested

01 **Success defined first**
6 grading dimensions approved before scoring: Tier correctness against ICP rules, Conservative handling of missing or conflicting data, Exact two-line output format, Reason cites specific column values, Reason length and single-sentence constraint, Grounding in the row only.

02 **Same test for every candidate**
3 prompt and model combinations each answered the same 25 test cases.

03 **Graded blind**
2 graders (claude-sonnet-5.5, gpt-5.4-mini) never saw which prompt or model wrote an answer, and none graded its own model.

04 **Average of the graders**
With fewer than three graders, scores are averaged.

05 **A range on every score**
From resampling the test cases 1,000 times. We also compare the top two on each test case; when the gap between them could vanish on other test cases like these, we call it too close to call.

The evidence: three test cases side by side

Same input, two candidates, graded blind. Check the scores against your own reading; every answer and reason is in the online verdict.

---

company: Carrow & Yates LLP; domain: carrowyates.com;…

Data quality reviewer finding a row where the industry tags contradict each other · Hard

```
company: Carrow & Yates LLP; domain: carrowyates.com; industry: Legal Services; employees: 1,200; locations: 'fifteen offices across the Midwest'; description: 'Full-service law firm. Support staff include paralegals and receptionists; attorneys are…
```

`deepseek-v4.1-flash` **7.8**

Not a fit Legal Services, 1,200 employees, 15 offices; description shows salaried attorneys and only hourly billing, no hourly workforce; industry_tag_2 Hospitality conflicts.

It is one sentence but about 27 words, over the 25-word limit, and uses semicolons to pack in clauses.

`claude-haiku-5.5` **9.0**

Not a fit Legal Services with 1,200 employees, salaried attorneys, and no hourly workforce; "hourly billing" refers to billing, and industry_tag_2 "Hospitality" conflicts with industry.

It is one sentence but about 26 words, slightly over the under-25 limit.

company: NorthBay Staffing Partners; domain: northbays…

Agency analyst reviewing staffing firms, where the business model overlaps with the client's space · Hard

```
company: NorthBay Staffing Partners; domain: northbaystaffing.com; industry: Staffing Agency; employees: 260; locations: 5 branches in Washington and Oregon; description: 'Places light industrial and warehouse temps with client employers, managing hourly…
```

deepseek-v4.1-flash **8.8**

Tier 1 260 employees, 5 branches, staffing agency placing light industrial/warehouse temps with hourly timesheets and shift assignments support Tier 1.

Tier 1 fits the row because 260 employees, 5 branches, and light industrial/warehouse temps with hourly timesheets and shift assignments…

claude-haiku-5.5 **7.8**

Tier 2 260 employees, 5 branches, staffing agency placing hourly warehouse temps with shift tracking; industry is not a listed vertical and headcount may exclude temps.

Tier 2 is defensible given 260 employees, 5 branches and hourly temp shift work, but it hedges on a 'healthcare staffing' vertical and the…

company: Lattice Cloud Analytics; domain: latticecloud…

Clay table operator screening accounts before the paid people search step · Easy

```
company: Lattice Cloud Analytics; domain: latticecloud.io; industry: Software; employees: 300; locations: 2 offices in Austin and Denver; description: B2B SaaS platform for embedded business intelligence dashboards; tech_stack: Salesforce, Slack, Gusto…
```

deepseek-v4.1-flash **9.7**

Not a fit Software industry with 300 employees and 2 offices; no hourly workforce evidence, so excluded under no-hourly-workforce rule.

The data is complete and consistent and the answer invents no doubt, though 'no hourly workforce evidence' slightly frames absence as a…

claude-haiku-5.5 **8.7**

Not a fit Industry is Software with 300 employees across 2 offices (Austin, Denver) and no hourly workforce described, so it fails the hourly-staff exclusion.

It is one sentence but about 26 words (Industry is Software with 300 employees across 2 offices (Austin, Denver) and no hourly workforce…