

Jev Phase A: Safe, but Not Promotable

Across 65 unique postings, Jev rejected nothing worthwhile and was right on every reject it issued. The sample was too small to clear the promotion bar, so Phase A closes with Jev remaining a manual aid.

Verdict: insufficient evidence for promotion. No live routing, quarantine, holds, or automated triage. The protocol's 20 natural reject bar was not met.

65

unique postings
adjudicated, two batches

0

false negatives in Jev's
reject lane

13/13

Jev rejects confirmed
correct on evidence

13 of 20

natural rejects required by
the protocol

Agreement is not correctness.

The career copilot is a reference evaluator, not ground truth. Every disagreement was adjudicated against posting evidence and the corrected candidate snapshot, and the counts below report agreement, not who was right.

PREPARED FOR

Nikema and the Hermes chief of staff

SNAPSHOT

September 18, 2026

65 unique postings, blind shadow test

The pilot's recommendation was replaced, not followed

The pilot

31 postings, 58% exact agreement. Jev called 13 strong, 11 weak, 7 reject; the copilot called 8 strong, 11 weak, 12 reject. All 7 Jev rejects matched the copilot, with 0 observed false negatives against the copilot, but those were not evidence-adjudicated. Jev was more optimistic in 11 of 13 disagreements.

The protocol change

The chief of staff replaced the pilot's live Phase 2 with a blind shadow test: no live quarantine, evidence-based eligibility with quote plus URL plus capture date, Eligible, Ineligible, and Unknown kept distinct, 50 postings trigger review not deployment, and fewer than 20 natural rejects means insufficient evidence, not a pass.

Fixed collection criteria v1

LinkedIn Jobs and Indeed. Exact queries: developer advocate, developer relations, frontend engineer, full stack developer. Past 7 days, newest-first. No seniority, pay, type, remote, or geography filters. Dedupe by normalized employer plus exact title plus location scope, checked against canonical URL and requisition ID.

Blind by construction

The copilot assessed and locked columns V and W before Jev ran, in both batches. Jev verdicts were written to a sealed tab. The 24 title-word ineligible calls in batch 1 were not rerun; they were tested at adjudication against actual requirements.

What was collected

Batch 1 logged 50 rows, 11 duplicates, leaving 39 unique postings. The extension added 30 rows, 3 duplicates and 1 excluded search-page row, leaving 26 unique. Combined: 80 rows, 65 unique postings, all adjudicated in columns X, Y, and Z.

Superseded: the pilot report's recommendation of a 50-posting live pilot with high-confidence reject quarantine. Phase A ran fully in shadow, and nothing in it touched live routing.

Source: [Phase A experiment sheet](#), rows 2-81; [pilot seed sheet](#); amended Phase A protocol and version log.

Sixty-five postings, every disagreement adjudicated

65

unique postings across both batches

66%

agreements, 43 of 65

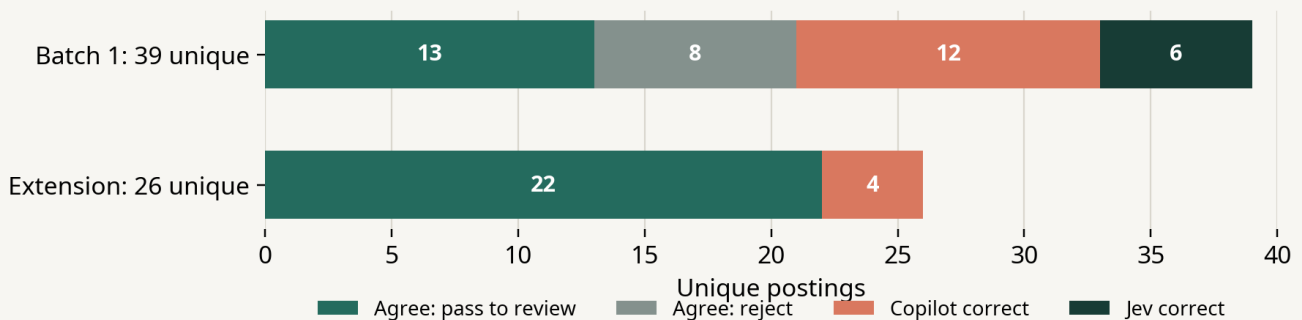
22

disagreements, each adjudicated on evidence

30 / 35

worthwhile yes / no after adjudication

Every disagreement was adjudicated against posting evidence



What agreement means here

Agree-pass (35) means both systems routed the role to review. Agree-reject (8) means both rejected on clear disqualifiers. Neither label says the shared verdict was correct; correctness was decided at adjudication.

What disagreement produced

Copilot-correct (16): the copilot's exclusion held up on quoted requirements. Jev-correct (6): Jev caught blockers the copilot missed, or preserved a role the copilot wrongly rejected. One copilot false negative total, in batch 1.

Source: [Phase A experiment sheet](#), columns X, Y, Z. Percentages are calculated from the 65 unique postings.

Jev caught what the title filter missed

In batch 1, the copilot marked five roles eligible that Jev rejected. Adjudication confirmed Jev in all five, each on a quoted mandatory blocker the baseline had missed:

POSTING	COPILOT SAID	THE BLOCKER JEV CAUGHT
Greptile, Frontend Engineer	Eligible	Text states an explicit preference for in-person over remote work.
L3Harris, Specialist Software Engineer	Eligible	Requires US security clearance; on-site in Colorado Springs.
Deloitte, Full Stack Developer	Eligible	Requires an active Top Secret / SCI clearance.
eSimplicity, Full Stack Engineer	Eligible	Requires passing a Public Trust clearance.
Piedmont Plastics, Full-Stack Software Engineer	Eligible	Charlotte, NC based; requires five or more years of experience.

The Cognition save: the copilot rejected the Developer Community Manager role on a title filter alone. The posting asks for 2+ years in developer relations or community, on the AI agents team behind Devin, a direct fit for the target transition. Jev kept it in weak fit. Adjudicated worthwhile: yes. One genuine save in 65 postings.

The value-add answer is thin.

Safe, yes: zero false negatives. Net value beyond the eligibility rules: one save in 65 postings barely demonstrates it. That is the honest reading, and it is why the gate stayed closed.

The weakness reproduced: geography

The pilot flagged geographic eligibility as underweighted: Jev called Supabase DevRel London a strong fit though London presence was required, and missed PlanetScale Support's EMEA-only restriction. Phase A reproduced the failure, this time on both sides:

Six extension roles, passed by both systems

Six agree-pass roles carried strict non-US physical residency requirements (Bengaluru, London, France). Both evaluators routed them to review. Adjudication marked all six not worthwhile. The miss was not a Jev-only problem; the copilot baseline passed them too.

Why it matters

Geography is the cheapest disqualifier to check and the most expensive to get wrong: every missed location restriction becomes a full review of a role that was never available. Six in 26 extension postings is not a corner case.

What would fix it

Structured extraction of geography, work authorization, clearance, and on-site requirements before any routing. Unknown routes to review. This is the first repair any Phase B would need, on both evaluators.

Standing rule, still unmet: unverified location counts as unknown, never assumed remote. Both systems need the extraction layer that enforces it.

No basis for strong-fit automation

Pilot: confidence routes, it does not authorize

13 calls at 0.70 and above produced 8 exact matches; 18 calls below produced 10. The gap is small, and five confident calls still disagreed. Confidence chooses the review path, never the outcome.

Phase A: the strong-fit lane barely exists

Batch 1 produced zero strong fits across 39 unique postings. The extension produced two, at confidences 0.53 and 0.29. Two low-confidence strong calls are not a lane; they are noise.

The two extension strong fits

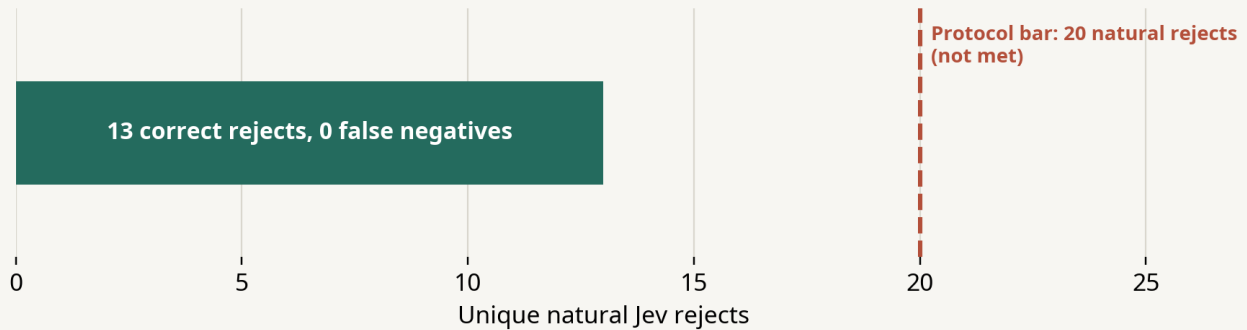
- Developer Relations Engineer (React), US: strong fit at 0.53
- Frontend Engineer, Web Applications (Remote): strong fit at 0.29

Operating rule: a strong-fit call must never authorize anything automatic: no tracker writes, no outreach, no suppression of review. Nothing in Phase A earned that trust.

Jev's verdict distribution also shifted hard between samples: 13 strong fits of 31 in the pilot, 2 of 65 unique in Phase A. Fit labels are sample-sensitive, which is another reason they cannot carry authority.

Thirteen is not twenty

The reject lane was perfect, but too small to promote



What the bar meant

Twenty naturally occurring rejects was the agreed minimum evidence for promotion consideration, set before collection started. Thirteen unique natural rejects, with zero in the extension batch, is insufficient evidence. That is the protocol working as designed, not a verdict on Jev's quality.

What a Phase B would need

Structured geography extraction first, on both evaluators. The same blind protocol and adjudication. The reject bar actually met. The copilot stays the evaluator. Any live role, routing, quarantine, or automation needs a separate explicit approval.

Phase A is closed. Final record: 65 unique postings, 0 Jev false negatives, 13/13 reject precision, 1 copilot false negative caught and preserved, 13 natural rejects against a 20 bar. Insufficient evidence for promotion.

Source record

ARTIFACT	LOCATION / VERSION
Experiment sheet	Phase A intake sheet: 80 rows, columns V/W (blind copilot assessments), X/Y/Z (adjudication), sealed Jev tab.
Jev screener	~/workspace/jev-experiment/screen_phase_a.py (single choice-question format; sealed reason records the matched criterion text).
Candidate snapshot	v1, corrected: X @NikemaCodes (~11K followers), Bluesky @nikema.dev (3.5K). Confirmed facts only.
Jev profile	v2: pilot v1 plus corrected social presence.
Collection criteria	v1, fixed before collection: LinkedIn Jobs and Indeed, four exact queries, past 7 days newest-first, no eligibility filters.
Pilot report	Corrected 2026-09-18. Its live Phase 2 recommendation was superseded by the Phase A protocol.
Version log	Sheet tab, records snapshot, profile, protocol, model, and extension authorization entries.

Process notes

- A mid-task model switch (gpt-5.6-terra to gemini-3.1-pro-preview), forced by a rate limit, caused context loss in the chief of staff's session. The switch is logged in the version log.
- The chief of staff's first adjudication summary misstated the agreement counts; the numbers were corrected to the sheet, which is the system of record.
- Copilot rationales described the candidate as "explicitly avoiding" senior roles. That preference is not in snapshot v1; the caveat is recorded in column Y, and outcomes were adjudicated on actual stated requirements.

Claim discipline: every number in this report traces to the sheet, the sealed verdicts, the pilot report, or the version log above. Percentages are calculated from the 65 unique postings unless stated otherwise.