OpenAI Talent: AI-Native Job Board (Concept Case Study)

One-liner

A job marketplace where candidates prove skills with live, AI-assisted work samples and hiring teams get fewer, higher-signal applicants.

Context

Hiring for AI roles is messy. Titles are vague, resumes don't prove ability, and inbound overwhelms teams. Traditional boards optimize for clicks, not fit.

Users

- **Candidates**: AI or ML engineers, data folks, applied researchers, PMs.
- **Employers**: Startups and mid-market teams building with LLMs.

Problem

Both sides waste time. Candidates apply into black holes. Recruiters sift inflated claims with little evidence. There is no shared, fast way to verify if someone can actually do the work.

Goal and Single Metric

Goal: Raise match quality while lowering time to a meaningful conversation.
North-star: Qualified Intro Rate (QIR).
Guardrails: Candidate and employer first-response time, fairness across cohorts, assessment error rate, assessment cost.

Constraints

Small team, privacy expectations, bias risk, ATS integrations, and keeping run-costs low.

Options Considered

- Classic posts and resumes: Easy to ship, low signal.
- Skill badges or quizzes: Some signal, easy to game.
- Work samples in a sandbox with AI-assisted scoring: Highest signal, needs human review for edge cases.

Decision: Work samples with rubrics and human-in-the-loop.

MVP Scope (8 to 10 Weeks)

Candidate

- Profile with verified identity, links, and Work Samples (tiny tasks: retrieval prompt, eval design, small data clean, AI feature UX).
- “Proof” uploads: repo, Colab, Loom.
- Private Career Agent to draft tailored applications from profile plus samples.

Employer

- Structured job spec (skills, models, data constraints, salary band, location).
- Evidence-based shortlist: rank by skills plus sample proof, not keywords.
- ATS sync (Greenhouse or Lever) for source of truth.

Platform

- Secure task runner (containerized), deterministic scoring with rubrics.
- Bias controls: AI scoring is advisory, human decision required, explainable rubrics.
- Event tracking for end-to-end funnel.

What I’d Build (My Role)

- PRD, structured job schema, and starter task library (engineering, data, PM).
- Candidate flow: pick task → timebox → run in sandbox → auto-score → human review if needed → publish to profile.

- Employer dashboard: shortlist by evidence, one-click invite, SLA nudges.
- Analytics: QIR, time-to-first-response, candidate NPS, cost per assessment, human-review rate.
- Policy guardrails with Legal: explainability, appeals, privacy, retention.

Success Metrics (First 90 Days)

- QIR ≥ 35%
- Time-to-first-response ≤ 72h
- Assessment cost P95 ≤ $0.45
- Human-review rate ≤ 10%
- Candidate opt-in to work samples ≥ 60%

Experiment Plan

Hypothesis	Change	Metric Read	Decision Rule
Work samples improve match	Apply plus 1 task vs classic apply	QIR, first response time	Ship if QIR improves by 10% with neutral response time
Structured spec reduces noise	Structured spec vs free text	Invite quality, drop-off	Keep if invite-to-onsite improves by at least 15%
Agent improves applications	Career Agent on vs off	Invite rate, time to apply	Keep if invite rate improves by 10%
Process is fair	Audit by cohort	Invite rate parity	Investigate any gap greater than 5pp

System Sketch

- Web app (Next.js)
- Profiles, Jobs, WorkSamples (Postgres)
- Task Runner (per-task containers)
- Scoring Service (rubrics and model evals)
- Review Queue (human-in-the-loop)

- Results back to profile plus ATS sync

Risks and Mitigations

- **Bias**: AI scoring is advisory. Publish rubrics. Human decision is mandatory. Appeals path.
- **Gaming**: Rotate task pools, plagiarism checks, require repo history or Loom walkthrough.
- **Privacy**: Opt-in for training use, 90-day default deletion, clear retention policy.
- **Cold-start**: Curate early adopters, weekly “shortlist” mailers to hiring teams.

90-Day Results (Hypothetical Snapshot)

- QIR hit 38% by week 6.
- Median time-to-first-response 48h.
- Employers invited 2.1× fewer candidates to reach onsite.
- Candidate applications drafted via Agent were 1.4× more likely to get an invite.
- Assessment cost P95 $0.39, human-review 8%.

How measured: GA4 plus server events into Looker, weekly readouts.

Next Steps

- Proof stacks on profiles (code, Loom, rubric) reusable per job.
- PM case tasks: PRD critique, experiment design, metric tradeoffs.
- Deeper ATS actions: stage updates and rejection reasons.
- University and bootcamp partners to supply verified samples.

Artifacts (Future Adds)

- Structured job schema (JSON)
- Candidate task rubric example (image)
- Employer shortlist mock (screenshot)
- Tracking plan (table)
- 90-sec Loom of the candidate task flow