Team 19

JobShield: Detecting Fraudulent Job Postings on Hiring Platforms

Yohannes Aklilu, Anna Roth, Anayochukwu Edwin Uche, Veronica Wang, Victoria Yang

Stanford
Computer Science

Problem Description

Attackers post deceptive jobs on platforms like LinkedIn and Indeed with the purpose of candidate data theft. For example, they may gather personal documents such as Visas or trick developers into downloading malware during interviews, stealing SSH & API keys, cloud credentials, and more, which not only exposes the victim, but can affect former employers as well.

Why Victims are Vulnerable:

- Fake companies may have a presence (domains, GitHub orgs, recruiter LinkedIn profiles) that look genuine
- Off-platform interviews are routine and expected
- Early-career applicants face pressure to apply fast and broadly

Why Platforms Fail Them:

- Existing defenses are predominantly reactive. Content removed after candidates report it
- When abuse is reported, credentials are already harvested
- Attackers can cheaply generate many postings

Real malware families using this vector, such as OtterCookie and FlexibleFerret, have been ongoing since 2024. Our system intercepts fraudulent postings before they reach applicants.

Evaluation

We tested three approaches on a balanced 100/100 labeled dataset of benign vs. fraudulent postings.

Approach	Precision	Recall	F1	Cost/1k
Logistic Regression	0.925	0.860	0.891	$0.00
LLM (Gemini 2.5 Flash)	0.591	0.810	0.684	$0.35
Hybrid (LR → LLM)	**0.880**	**0.950**	**0.913**	**$0.077**

The table shows precision, recall, F1, and cost. The table below breaks down where each model actually fails: fraud that slips through vs. legitimate jobs wrongly blocked.

	Fraud Caught	Fraud Missed	Legit Wrongly Flagged
LR alone	86/100	14	7
LLM alone	81/100	19	56
Hybrid	**95/100**	**5**	13

Headline: Hybrid leads with every metric: highest F1 (0.913), fewest missed fraud postings (5/100), and 4.5x cheaper than the LLM alone

Why The Hybrid Wins: LR clears obvious cases in 0.05ms. Only the uncertain 22% escalates to Gemini, where qualitative reasoning contributes

Why LLM alone underperforms: Gemini over-flagged 56 legitimate postings, having no calibrated sense of what a normal job posting looks like on the dataset

Policy Language

Our platform connects job seekers with real, not fraudulent opportunities. Suspicious postings are flagged and held for human review.

Banned Content:

- **Company Impersonation**: falsely claiming affiliation with/as a real organization as supported by fraudulent online channels
- **Deceptive Job Details**: misrepresenting salary, visa sponsorship or remote status to attract applications under false pretenses or benefits
- **Premature Data Extraction**: requesting SSNs, passport scans, or any payment (application fees, "processing costs") before required in a legitimate hiring stage
- **Malicious Interview Workflows**: repositories, coding challenges, or assessment links designed to install malware, access local files, or silently exfiltrate credentials (API tokens, SSH keys, cloud access keys, .env files)

Allowed: Small, or new companies (like startups) are welcome. A recently registered domain or minimal web presence is not grounds for removal. Suspicious behavior is flagged, not company size or reputation.

Why it matters: Developers reasonably trust that interview code is safe to run. A single malicious postinstall script can steal credentials accumulated across years of work/use, compromising not just the applicant, but former employers whose tokens were never revoked. Visa-dependent and early-career applicants are disproportionately harmed, often applying at high volume under time pressure, oftentimes skipping to vet each posting individually.

Enforcement: Flagged postings are held from publication and routed to human moderators. High-confidence cases are blocked automatically. Incorrectly flagged postings are may be appealed through the moderator.

Technical Backend

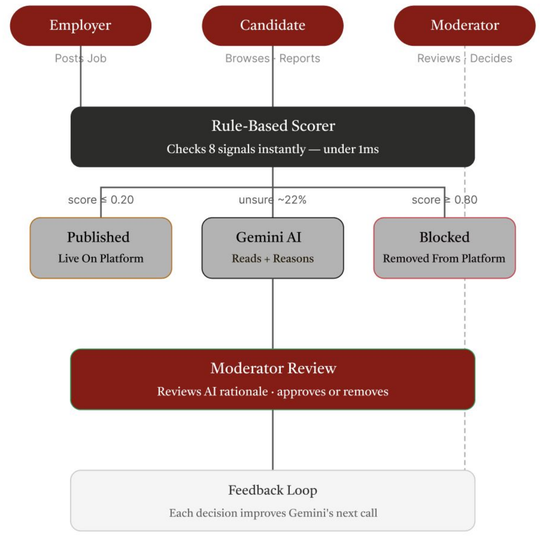
JobShield is a full-stack job posting platform with a three-layer fraud detection pipeline, live moderator interface, and an automatic feedback loop, all deployed on a single URL.

How A Posting Gets Classified:

1. Employer submits → system checks auth and requires a completed employer profile (prevents spoofing another company's identity)
2. Logistic regression + TF-IDF runs instantly on the posting text plus 8 metadata flags (e.g. missing salary, missing requirements)
3. If score ≤ 0.20 → auto-approved. If ≥ 0.80 → auto-blocked. If in between: Gemini reads the full posting and returns a verdict with plain-English rationale
4. If Gemini fails for any reason: routes to moderator as a safe
5. default. Errors are never auto-published.
6. Moderator reviews flagged postings, sees per-job report history and per-reporter history, and approves, removes, or escalates

Feedback loop: Every time a posting is approved or rejected, the decision is auto-injected as a labeled few-shot example in the next LLM call (on both posting-intake and report-analysis), with no retraining.

Stack	Why
Next.js	Server rendering, no separate backend needed
Supabase	Postgres + auth + row-level security in one service
Gemini 2.5 Flash	Fast, cheap ~$0.0006/call with few-shot
Scikit-learn	Free, sub-millisecond, interpretable weights
Vercel	Zero-config deploy, auto-deploys on push



Looking Forward

Next Steps:

- Tune Gemini's prompt with labeled examples to fix the 56 false positives on legitimate postings
- Grow the test set over the 200 examples, and add confidence intervals to confirm the hybrid's win is statistically significant
- Build an appeals flow for legitimate recruiters whose postings get incorrectly blocked

Scaling concerns:

- **Adversarial Evasion**: attackers will probe and adapt once the classifier is live, so keeping thresholds private and retraining frequently is essential
- **Concept Drift**: OtterCookie and FlexibleFerret both deployed within a couple months of each other–attack patterns evolve so maintenance cycles must reflect this
- **Disparate Impact**: small companies or those based in countries with less accessible web infrastructure may be disproportionately flagged

Techniques Not Pursued:

- **Repo Scanning**: directly analyzing linked GitHub repositories for malicious scripts would catch attacks invisible in the job description text
- **Recruiter Verification**: authenticating recruiters against company HR systems would eliminate impersonation, but requires platform partnerships that aren't accessible
- **Platform Expansion**: Handshake, Glassdoor, and internship boards face the same threat