V3 TUNED ENGINE VALIDATION · MARCH 2026

The Phantom Workforce

A Deterministic Ghost Job Detector With 0.904 Correlation to Llama 3.3 70B

1,000 stratified listings · Llama 3.3 70B & Gemini 2.5 Flash baselines · thesubspace.io

Download V3 PDF ↓ V2 PDF (archive)

0.904
Correlation with Llama 3.3 70B

98.2%
Within 30pp of Llama

0.877
Correlation with Gemini

18
Major disagreements / 1,000

1. The Ghost Job Economy: February 2026

The global labor market of February 2026 is defined by a growing disequilibrium between digital signal and operational intent. Job postings increasingly serve secondary functions — market signaling, talent stockpiling, and internal compliance — rather than immediate recruitment. The “hires-per-job-posting” ratio has collapsed from 0.75 in 2019 to below 0.5 in late 2025, meaning fewer than half of all active listings correspond to a genuine, funded vacancy.

Macroeconomic Drivers

- **The Hiring Freeze Paradox.** Entities such as Salesforce, Amazon, and Heineken announced aggressive hiring freezes in early 2026, yet ATS platforms like Workday keep postings live for weeks after internal freeze orders. This “Zombie Job” window creates critical risk: listings are technically active but operationally dead.

- **The Internal Compliance Trap.** Government and Education sectors exhibit ghost rates as high as 60% and 50%, driven by regulations mandating public posting even when internal candidates are pre-selected. Ontario enacted legislation in 2026 requiring employers to disclose whether a posting is for an existing vacancy.
- **Talent Hoarding & Evergreen Requisitions.** The Technology sector (~48% ghost rate) uses generic titles to continuously harvest resumes. Data indicates ~45% of employers post listings without an immediate plan to hire, building pipelines for hypothetical future quarters.
- **The ATS Amplifier Effect.** This study draws directly from corporate Applicant Tracking Systems — Greenhouse, Ashby, Lever, and SmartRecruiters. When companies freeze hiring but fail to close requisitions, the ATS continues serving the listing as active. ATS-direct data reveals the raw “corporate intent signal” — making ghost detection both more precise and more consequential.

2. Why Deterministic Scoring Matters

Ghost job detection tools broadly fall into two categories: AI-based systems that use large language models to analyze listings, and deterministic algorithms that apply structured rules to observable signals. The V2 engine proves these approaches can converge — achieving identical mean scores with fundamentally different architectures.

- **Reproducibility.** Given the same listing, the Subspace V2 engine produces the same ghost score on every evaluation. AI models exhibit run-to-run variance of ±10-15pp on the same listing due to sampling temperature, context window effects, and model updates.
- **Auditability.** Every V2 score decomposes into 7 nutrition categories, each derived from a transparent DAG of enrichment signals. When a job is flagged as high risk, the system explains exactly which categories drove the assessment.
- **Cost at Scale.** Scoring 1,000 listings through two LLMs costs API fees and takes processing time. Subspace scores the same dataset in seconds at zero marginal cost. The V2 validation ran at $0.00 total inference cost for the deterministic engine.
- **Multi-Model Validation.** This study validates against two independent LLMs (Llama 3.3 70B and Gemini 2.5 Flash Lite) rather than a single baseline. If a deterministic algorithm converges with two different AI architectures simultaneously, the underlying signal structure is robust — not an artifact of one model's biases.

> **V2 Architecture:** 51+ enrichment signals flow through a directed acyclic graph (DAG) to produce 24 L2 composites and 4 L3 meta-patterns, which are then weighted into 7 nutrition categories and a final Ghost Probability Score (GPS). This layered architecture captures compound risk patterns that simpler rule-based approaches miss.

3. The Three Detection Models

This study compares three ghost job scoring methodologies against a stratified sample of 1,000 real job listings drawn from the Subspace database. The dataset spans multiple ATS sources and a full range of posting ages, providing a comprehensive test of each model's scoring behavior.

Subspace V3 Tuned (Deterministic): A tuned multi-signal scoring engine that processes 51+ enrichment signals through a DAG architecture into 6 nutrition categories — freshness, authenticity, listing completeness, employer quality, compensation, and role substance — with positive-only signal handling and softened post-processing. Mean: 38%, median: 35%, std dev: 13.

Llama 3.3 70B Instruct (AI Baseline 1): Meta's 70-billion parameter instruction-tuned model, accessed via OpenRouter. Evaluates each listing holistically with non-linear reasoning, contextual language analysis, and compound risk assessment. Acts as the primary AI baseline for correlation and agreement measurement. Mean: 53%, median: 58%, std dev: 20.2.

Gemini 2.5 Flash Lite (AI Baseline 2): Google's efficient frontier model, accessed via OpenRouter. Provides a second independent AI perspective on each listing, enabling three-way cross-validation. Tends to score slightly higher than both Subspace and Llama, particularly on older listings and Lever-sourced jobs. Mean: 59%, median: 65%, std dev: 18.

Model	Mean	Median	Std Dev	Type	Architecture
Subspace V3	38%	35%	13	Deterministic	51+ signals → DAG → 6 nutrition categories → GPS
Llama 3.3 70B	53%	58%	20.2	AI / Stochastic	70B parameter LLM, holistic listing assessment
Gemini 2.5 Flash	59%	65%	18	AI / Stochastic	Frontier LLM, efficient multi-modal evaluation

4. Three-Way Model Agreement

The headline result: Subspace V3 Tuned achieves a 0.904 correlation with Llama 3.3 70B across 1,000 listings — up from 0.789 in V2. Agreement within 30pp reaches 98.2%, with only 18 major disagreements (1.8%) exceeding the 30pp threshold. The mean gap of 15pp (Subspace 38% vs Llama 53%) reflects deliberate calibration: the tuned engine is more conservative, flagging fewer false positives.

Pair	Agree (≤15pp)	Minor (16-30pp)	Major (>30pp)	Correlation	Mean Gap
Subspace vs Llama	532 (53%)	450	18	0.904	15pp

Pair	Agree (≤15pp)	Minor (16-30pp)	Major (>30pp)	Correlation	Mean Gap
Subspace vs Gemini	282 (28%)	609	109	0.877	21pp
Llama vs Gemini	988 (99%)	12	0	0.975	6pp

> **Three-Way Convergence:** All three pairwise correlations exceed 0.87 (0.877-0.975), and the Llama-Gemini pair shows near-perfect agreement (r=0.975, 0 major disagreements). The V3 tuned engine's 0.904 correlation with Llama — up from 0.789 in V2 — shows that positive-only signal handling and weight rebalancing brought the deterministic engine significantly closer to LLM consensus.

5. Confusion Matrix & Classification Analysis

With V3 tuning, Subspace scores lower on average (38% vs 53%), meaning 0 listings reach the 60%+ threshold. The risk tier distribution tells a clearer story: Subspace places 621 listings in Low risk (<40%) and 379 in Moderate (40-59%), while Llama places 459 in High risk (≥60%). The tuned engine is deliberately more conservative — it avoids over-flagging.

621	379	0
Subspace Low	Subspace Moderate	Subspace High

Conservative by Design. The V3 engine's positive-only signal handling means absence of data (no salary, no interview process) is no longer penalized — only confirmed negative signals drive up ghost scores. This eliminates the false-positive problem identified in V2 where 47 legitimate listings were over-flagged.

High Correlation Despite Different Means. Despite the 15pp mean gap, the 0.904 correlation shows that Subspace and Llama rank listings in nearly the same order. The engine correctly identifies the same relative risk — it just uses a tighter scale.

> **V2 → V3 Improvement:** Correlation with Llama improved from 0.789 to 0.904 (+14.6%). Correlation with Gemini improved from 0.655 to 0.877 (+33.9%). Major disagreements with Llama: 6 → 18 (slight increase due to tighter scoring window, but still only 1.8% of the dataset).

6. Score Distribution Analysis

The V2 engine produces a well-distributed scoring curve. Subspace scores span 20-97%, with the bulk concentrated in the 30-69% range (784 of 1,000 listings). This spread avoids the pathological clustering seen in earlier validation sets — neither bunching at 0% nor at 60%+.

Llama shows a bimodal distribution with peaks at 20-39% and 60-69%, while Gemini concentrates heavily at 60-69% (358 listings). Subspace's more Gaussian-shaped distribution suggests the V2 DAG architecture produces finer-grained risk differentiation than either LLM approach — scoring the “muddy middle” of moderate risk rather than forcing binary low/high classifications.

Risk Tier	Subspace V2	Llama 3.3 70B	Gemini 2.5 Flash
Low Risk (<30%)	260 (26%)	268 (27%)	211 (21%)
Moderate (30-59%)	412 (41%)	273 (27%)	134 (13%)
High Risk (≥60%)	328 (33%)	459 (46%)	655 (66%)

7. Ghost Probability by Posting Age

All three models produce a monotonically increasing risk curve with posting age — the strongest single ghost signal. On fresh listings (0-7 days), Subspace averages 39%, Llama 27%, and Gemini 30%. On stale listings (90+ days, n=337), the models converge at 67-73%. The 28pp spread between fresh and stale confirms robust age discrimination across the V2 engine.

Age Band	N	Subspace	Llama	Gemini	Sub-Llama	Sub-Gemini
0-7 days	75	39%	27%	30%	+12pp	+9pp
8-14 days	99	40%	28%	32%	+12pp	+8pp
15-30 days	186	41%	40%	46%	+1pp	-5pp
31-60 days	182	48%	52%	65%	-4pp	-17pp
61-90 days	114	59%	61%	68%	-2pp	-9pp
90+ days	337	67%	72%	73%	-5pp	-6pp

> **The Convergence Pattern:** Subspace scores 12pp above Llama on fresh listings (39% vs 27%) — this reflects V2's structural risk detection (ATS source penalties, transparency gaps) that

> fresh-biased AI models tend to discount. As listings age, the gap closes: at 15-30 days all three models cluster within 6pp, and at 90+ days the spread narrows to just 6pp (67-73%).

8. ATS Source Analysis: A New Detection Dimension

The V2 engine introduces ATS source-level analysis — a dimension invisible to language-only AI models. By identifying which Applicant Tracking System serves a listing (Greenhouse, Ashby, Lever, SmartRecruiters, Amazon), the algorithm detects platform-specific risk patterns that correlate with ghost job prevalence.

ATS Source	N	Subspace	Llama	Gemini	Key Insight
Greenhouse Direct	604	47%	49%	57%	Lowest risk; largest ATS. Three-way convergence.
Ashby Direct	240	59%	56%	59%	Mid-range. Startup/growth-stage ATS.
Lever Direct	113	72%	67%	69%	Highest risk across all models. Pipeline-heavy culture.
SmartRecruiters	24	53%	54%	51%	Tight convergence: 51-54%.
Amazon Direct	10	51%	42%	43%	Subspace +9pp over AI — structural risk detection.

> **Lever Is the Highest-Risk ATS.** All three models agree: Lever-sourced listings average 67-72% ghost probability vs 47-57% for Greenhouse. This 20-25pp gap is consistent across all models and reflects Lever's prevalence among startups that maintain “always open” pipeline requisitions. V2's ATS detection gives it a structural advantage: it can differentiate a Lever listing from a Greenhouse listing before reading a single word of the description.

9. Case Studies

The following cases illustrate model behavior across a range of scenarios — from fresh genuine listings to ancient zombie posts, including individual matched scores and aggregate patterns from the 1,000-job validation set.

#	Case Type	Sub	Llama	Gem	Key Finding
1	Genuine Fresh — Specific Title	20%	20%	15%	All three converge: fresh, salary-posted, specific title = genuine

#	Case Type	Sub	Llama	Gem	Key Finding
2	Genuine Fresh — ML Role	25%	10%	15%	Strong convergence at low risk; AI models even more confident than Subspace
3	Subspace Catches Structural Risk	57%	26%	30%	Subspace detects Lever source + transparency gaps that AI models overlook
4	Moderate Age — Three-Way Convergence	48%	52%	65%	All three flag 31-60d risk; Gemini highest due to sector penalty
5	Stale Listing — High Convergence	67%	72%	73%	All three converge at 67-73% on stale listings — risk is unambiguous
6	Ancient Zombie — Near-Certain Ghost	97%	98%	85%	5+ year old listing: all models max-flag. Subspace catches via temporal + Lever forensics
7	Lever Source Premium	72%	67%	69%	Lever listings avg 72% — Subspace detects ATS-level risk patterns
8	AI Catches Semantic Risk	35%	85%	80%	"Talent" in title signals pipeline role; AI catches linguistic intent Subspace misses
9	Gemini Overscore on Fresh	23%	20%	65%	Fresh, specific tech role — Subspace/Llama correct at ~22%. Gemini +42pp too high.
10	Salary Posted — Universal Signal	42%	41%	53%	Salary disclosure lowers scores across all models: 16pp Subspace, 18pp Llama

Where Subspace Leads

On ATS-level risk (Row 3: VRChat, Lever source) and temporal patterns (Row 6: Peakgames zombie), Subspace detects risk before AI models can. The V2 engine's enrichment pipeline surfaces signals — source classification, staleness curves, transparency gaps — that are structurally invisible to language-model analysis.

Where AI Leads

On semantic-intent signals (Row 8: “Talent” in title at Wrike), Llama and Gemini catch pipeline-signaling language that the deterministic engine misses. This is the primary enhancement target for V3: integrating lightweight NLP signals without sacrificing determinism.

10. Salary Transparency & Signal Strength

Salary disclosure remains one of the most powerful ghost signals across all three models. Listings with salary posted average 42% on Subspace vs 58% without — a 16pp gap. Llama shows an 18pp gap

(41% vs 59%). Gemini is less salary-sensitive with an 8pp gap (53% vs 61%), consistent with its heavier weighting on structural factors over listing-level transparency.

Only 30% of listings post salary (303 of 1,000). The remaining 70% receive a transparency penalty from both Subspace and Llama. This aligns with industry research: salary omission correlates with organizational opacity, which in turn correlates with ghost job prevalence. New regulatory requirements in Ontario and proposed bills in other jurisdictions may shift this ratio — and the V2 engine's modular architecture can adapt its salary weighting as market norms evolve.

16pp
Subspace salary gap

18pp
Llama salary gap

8pp
Gemini salary gap

11. Ghost Job Taxonomy

Industry research identifies six distinct categories of ghost listings. The V2 engine's multi-signal architecture detects each through different combinations of its 7 nutrition categories, while AI models rely on holistic assessment of each listing's text.

Ghost Type	Description	V2 Detection Vector	AI Detection Approach
Evergreen Requisition	Perpetual listing harvesting resumes; generic title	Source forensics + age curve + description entropy	Catches "generic" language and repetition
Zombie Post	Filled/frozen but ATS keeps it live	Age signal + source staleness patterns	Detects description decay over time
Compliance Artifact	Posted for legal reasons; candidate pre-selected	Sector risk + specific requirement density	Analyzes language formality patterns
Talent Hoarding	Building a bench for future quarters; no budget	Transparency gaps + temporal patterns	Catches conditional/tentative language
Market Tester	Gauging salary expectations; no approved budget	Salary absence + vague requirement signals	Detects "exploratory" framing

Ghost Type	Description	V2 Detection Vector	AI Detection Approach
Genuine Listing	Active vacancy with funded headcount	All nutrition scores low → low GPS	Full-text validation confirms activity

12. Conclusions & Strategic Implications

The V2 engine achieves the most compelling validation result in Subspace's development history: zero mean gap with a 70-billion parameter LLM (53% vs 53%), 78% agreement, 0.789 correlation, and only 6 major disagreements across 1,000 stratified listings. This is not a fluke of dataset composition — the sample spans 7 age bands, 5 ATS sources, and the full range of industries and seniority levels.

1. **Zero mean gap with Llama 3.3 70B.** Subspace V2 (53%) and Llama (53%) produce identical average ghost scores on 1,000 stratified listings. 78% agree within 15pp. Correlation: 0.789.
2. **Second AI baseline confirms convergence.** Gemini 2.5 Flash (59%) is 6pp higher but maintains 66% agreement and 0.655 correlation with Subspace. Three independent approaches, one conclusion.
3. **ATS source analysis is a new detection dimension.** Lever listings average 72% vs Greenhouse at 47% — a 25pp gap confirmed by all three models. Only Subspace can use this signal deterministically.
4. **Salary transparency is the strongest universal signal.** 16pp gap (Subspace), 18pp (Llama), 8pp (Gemini). All models agree: salary omission correlates with ghost risk.
5. **Age discrimination works.** From 39% (fresh) to 67% (stale 90+), the V2 engine produces a 28pp graduated risk curve. AI models show similar curves, confirming the signal is real.
6. **The remaining gap is semantic.** Llama catches linguistic intent (“Talent” in title, tentative language) that Subspace misses. This is the V3 development target: integrating lightweight NLP without sacrificing determinism.

> **Bottom Line:** The V2 engine proves that deterministic ghost job detection can match frontier AI accuracy. Zero mean gap, 78% agreement, 0.789 correlation — validated against two independent LLMs on 1,000 stratified real-world listings. Same input, same output, every time, at zero cost per listing. Explore the live scoring engine at thesubspace.io.

Questions about this research? vivek@thesubspace.io