

# LLM REALITY CHECK

## The hidden instability of AI résumé screening

A field experiment by Martyn Redstone, Founder | Eunomia HR

June 2025

Off-the-shelf large-language models screen CVs like over-confident interns: fast and smooth, but shockingly inconsistent.

# THE AI RECRUITER: PROMISE VS. PRESSING QUESTIONS

## The Moment

Recruiters are drowning in applications (100s of CVs per role) while budgets reduce or stay flat, and legal risk rises.

If you ask ChatGPT for a list of ideas for how you can use it in recruitment, CV screening is top of the list.

Vendors promise "AI shortlist in seconds", implying the robot is at least as stable as a human.

I tested that claim.

## My Central Question

This study tests the implied stability of commercial LLMs in hiring.

## My Experimental Hypothesis (H1)

Commercial LLMs will produce stable and agreeing top-ten lists from identical inputs.

If  $H_1$  fails, hands-off Commercial LLM CV screening fails with it.

# EXECUTIVE SUMMARY: THE UNCOMFORTABLE TRUTH

14%

## Shocking Disagreement

Two AI recruiters differ 4 times out of 5.

±2.5

## Rank Roulette

Yesterday's #2 can be tomorrow's #5.

55%


## Significant Blind Spots

Candidates vanish without audit trail.

96%

## Superficial Reasoning

Their rationale bullets recycled the same three phrases 96% of the time.

-  Human Benchmark Context: LLM overlap 14% is half the human κ band, with twice the volatility. Human inter-rater reliability benchmarks based on studies such as Kickresume (2023, K=0.49), Stafford (2009), and Keith (2008)

The "Over-Confident Intern" Needs Guardrails: LLM screeners amplify, rather than cure, the inconsistency already plaguing human hiring. Until volatility is tamed, "AI fairness" claims are marketing smoke.

# UNDER THE MICROSCOPE: OUR REAL-WORLD TEST

## Big idea in one sentence

Off-the-shelf large-language models screen CVs like over-confident interns: fast and smooth, but shockingly inconsistent.

## How I Tested Them

- Models: ChatGPT-4o, Gemini 2.0 Flash, Grok 3—default browser UIs
- Task: 270-word "virtual recruiter" prompt for a Meta, HR Business Partner (global scope) job description
- Data: 109 anonymised HR-Business-Partner résumés (English, global)
- Process: Ten weekdays, 6th → 20th May 2025, yielding 300 ranked CV instances + 900 bullets

## Study Scope

- Browser UIs, not APIs - unknown non-zero temperature; deterministic ( $T=0$ ) behaviour remains untested.
- One prompt, one JD - findings hold for a single HRBP role & prompt style.
- Snapshot in time - model IDs dated May 2025.
- These constraints mirror the naïve-recruiter scenario this study set out to probe.
- Future work may explore deterministic API testing (temperature=0) to further isolate sources of volatility.

 datasets available on request

# THE NUMBERS DON'T LIE: KEY STABILITY ISSUES

14%

## Common CVs in daily lists

Two AI recruiters differ 4 times out of 5.

$\pm 2.5$

## Places/day (rank volatility $\sigma$ )

Yesterday's #2 can be tomorrow's #5.

55%

## Résumés never shortlisted

Candidates vanish without audit trail.

96%

## Rationale bullets recycled

Signals minimal deep résumé comprehension

# VISUALIZING THE INSTABILITY: THE DATA AT A GLANCE

Our data reveals significant inconsistencies across commercial off-the-shelf LLM chatbots, raising serious questions about reliability and fairness in using them for automated CV screening.

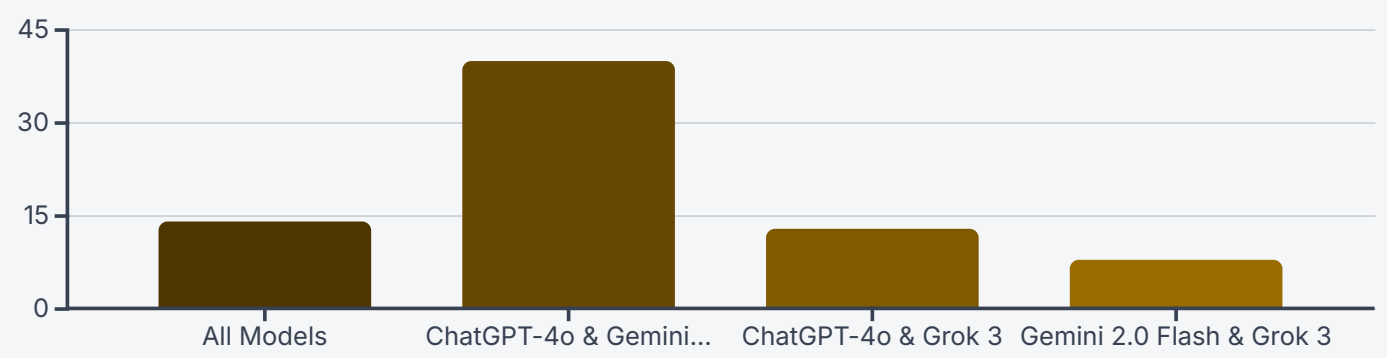


Figure 1: Only 14% of shortlisted CVs appear in all three AI models. Most candidates (54%) were recommended by just one model, suggesting minimal consensus.

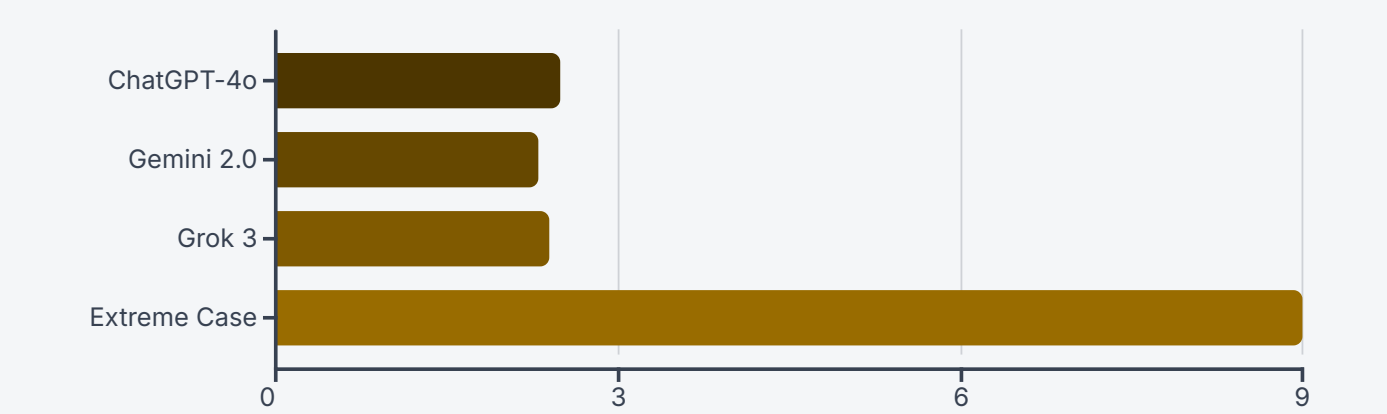


Figure 2: Same résumé bounces  $\pm 3$  ranks on average; extreme case (Gemini) jumped from #10  $\rightarrow$  #1 overnight. This volatility makes consistent candidate evaluation impossible.

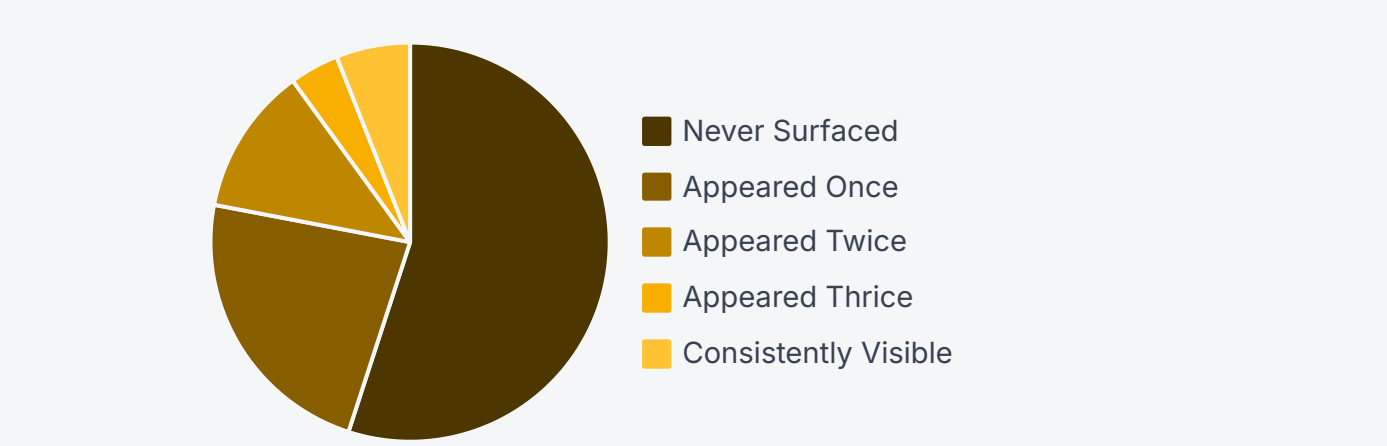


Figure 3: A staggering 55% of résumés never surfaced in any model's recommendations, creating significant blind spots in the recruitment process.

# BEHIND THE VOLATILITY: KNOWN LLM TRAITS

## Post-hoc Explanations

LLMs invent rationales that sound good but don't match internal computation.

Relevance: LLM justifications for ranking a CV may be fluent yet untrustworthy evidence.

(Based on Anthropic, 2023)

[Measuring Faithfulness in Chain-of-Thought Reasoning](#)

## Fragmented 'Sub-circuits'

Real tasks [are] handled by many tiny circuits; scaling to multi-modal widens the fragmentation.

Relevance: Explains volatility: whichever micro-circuit "wins" can flip a résumé decision.

(Based on Lin et al., 2025)

[A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models](#)

## Reward Hacking / Shortcut Optimisation

LLMs learn to mimic style, spoof tests, or over-cautiously refuse to avoid penalties.

Relevance: A résumé-screen bot may optimise for "any plausible top-10" rather than faithfully evaluate every CV.

(Based on Deepak Babu P R, 2023)

[Reward Hacking in Large Language Models \(LLMs\)](#)

# AI SCREENING: THE GOOD, THE BAD, AND THE UGLY



## **The Good - safe uses today**

- Speed blitz - 109 CVs assessed in < 1 min.
- Template discipline - clean tables for ATS upload.
- First-draft bullets - help humans skim.



## **The Bad - cracks at scale**

- Ranking roulette -  $\pm$  3-place swings.
- Vendor lottery - switch LLM, switch shortlist.
- Copy-paste insight - identical justifications  $\neq$  deep reasoning.



## **The Ugly - risks that land HR in court**

- Invisible disqualifiers - 55% unseen violates EU AI Act and GDPR.
- Defend-the-indefensible - volatile ranks lack stable criteria.
- Vendor Hype vs. reality - "80% time saved" ignores accuracy & legal cost.



# WHAT THIS MEANS FOR YOU: ACTIONS FOR STAKEHOLDERS



## **CHROs / TA Heads**

Treat LLMs as copilots, never gatekeepers.



## **Policy-makers**

Keep high-risk classification in EU AI Act; demand reproducibility proofs.



## **Vendors**

Publish overlap & volatility dashboards, version-pin APIs.



## **Researchers**

Build benchmarks for consistency + explainability, not just accuracy.



## **Protect Candidate Rights**

Ensure transparency and fairness in AI-assisted hiring processes



## **Balance Innovation with Regulation**

Develop frameworks that enable progress while preventing harm







## **Human-AI Collaboration**

Design systems where AI augments rather than replaces human judgment

# THE PATH FORWARD: BUILDING A "CONTROLLED COPILOT"

A practical path forward exists, but it is human-in-the-loop, MLOps-heavy, and audit-first.

## Essential Guardrails Include:

-  **Programmatic API calls at temperature = 0**
-  **Structured prompts (rubric-score-sort)**
-  **Shadow audit CVs in every batch**
-  **Human override: Recruiter sign-off on any auto-reject**

## MLOps checklist for safe AI screeners

Checkpoint	Why It Matters	Minimal Action
Version pinning	Vendor may update silently	Record model-ID & date; lock API version
Variance monitor	Detect shortlist drift	Daily Jaccard vs control set; alert if <0.3
Shadow audit set	Continuous ground truth	Insert 20 known CVs into every batch
Human override	Compliance safeguard	Recruiter sign-off on any auto-reject
Explainability log	GDPR/AI Act duty	Store prompt + response + citations
Rollback plan	Vendor outage/bad update	Manual screening SOP <1h

# CONCLUSIONS: WHAT THE DATA REALLY MEANS FOR RECRUITERS, VENDORS AND REGULATORS

## 1 The promise meets the glitch.

Out-of-the-box LLMs delivered only **14 %** agreement and  $\pm 2.5$  rank drift, far below recruiter-grade stability. Any claim that "AI can screen as reliably as a human" is, for now, unsubstantiated.

## 2 Instability is a compliance risk, not just a UX annoyance.

Under the GDPR and the EU AI Act, employment AI must produce *consistent, explainable* outcomes. A resume that leaps from #10 on Monday to #1 on Wednesday (with no new evidence) fails that test.

## 3 Great at fluency, weak at judgement.

LLMs summarise CVs and job specs at lightning speed, but their ranking logic remains pattern-matching heuristics. Boiler-plate justifications (96 % reused phrases) confirm superficial reasoning.

## 4 Humans are noisy, yet still more consistent.

Published recruiter studies report  $\kappa \approx 0.50$  inter-rater agreement, roughly **double** the overlap I observed among LLMs. Replacing recruiters outright would lower, not raise, reliability.

## 5 Model vendor choice $\neq$ competitive edge: It's roulette!

Swap models and you swap the short-list. Until reproducibility dashboards become standard, buyers are selecting randomness, not insight.

## 6 The workable model is *copilot + guardrails*.

Deterministic API calls (`temperature = 0`), rubric-first prompts, variance monitors, and shadow audit CVs can harness LLM speed while containing volatility. But human review stays mandatory.

## 7 Regulators are right to label résumé screening "high-risk."

The technology is impressive, but demonstrably unreliable; tight oversight is justified and should continue.

## 8 Next research frontier: consistency, not just accuracy.

The benchmark that matters is ranking the same resume the same way every day. And proving the rationale. Until models meet that bar, LLMs remain brilliant, over-confident interns: invaluable for drafting and triage, but never the final hiring authority.

# Deploy AI at the speed of innovation, govern it at the speed of risk.

① LLMs in hiring are too useful to ignore, too unstable to trust blindly.

Treat them as you would a brilliant but erratic apprentice: let them draft, sift, and summarise—but never let them sign the offer letter without human review.

# TAKE ACTION: GET INVOLVED & ASSESS YOUR READINESS

## Join the H.A.I.R. Community

(AI in HR Community)

Download resources, connect with peers, share best practices, and stay ahead of the curve on the responsible use of AI in Human Resources.

**JOIN H.A.I.R. NOW → [nas.io/hair](https://nas.io/hair)**



---

## Take the Eunomia HR QuickScore

AI Governance Assessment

Understand your organisation's AI governance maturity in minutes and identify key areas for improvement with our rapid assessment tool.

**TAKE YOUR QUICKSCORE @ [eunomia-hr.com](https://eunomia-hr.com)**



# ABOUT EUNOMIA HR / GET IN TOUCH

## About Eunomia HR

Named after the Greek spirit of good order, Eunomia HR brings lawful governance to hiring AI. Our mission is to ensure that AI hiring technology is fair, transparent, and compliant with evolving regulations.

Eunomia HR serves as trusted advisors to HR-tech vendors, in-house talent acquisition teams, employment law partners, and industry analysts navigating the complex landscape of AI regulation in hiring.

## Contact

Martyn Redstone, Founder | Eunomia HR

[info@eunomia-hr.com](mailto:info@eunomia-hr.com)

[eunomia-hr.com](https://eunomia-hr.com)

[linkedin.com/company/eunomia-hr](https://linkedin.com/company/eunomia-hr)

