

HIREMIKE RESEARCH

How Common Is AI-Like Language in Online Job Advertisements?

A detector-calibrated cross-sectional study of four public job-board feeds

596 of 955 eligible advertisements (62.4%) were classified as AI-like.

This result is a detector classification, not proof that 62.4% of the advertisements were written by AI.

Prepared by: **HireMike Research**

Protocol ID: hiremike-ai-job-ads-2026-10

Observation period: 15 October 2026, 00:00–23:59 UTC

Analysis completed: 16 October 2026

Report version: 1.0

hiremike.co.uk/research/ai-like-job-advertisements-2026

Executive summary

Generative AI tools can produce complete job advertisements in seconds, but employers rarely disclose whether those tools were used. This makes it difficult to estimate how much contemporary recruitment writing is generated or substantially rewritten by AI. HireMike conducted an exploratory study of job advertisements exposed through four official public feeds: Himalayas, Jobicy, Remote OK and Python.org Jobs.

The study collected 1,119 feed records on 15 October 2026. After removing same-board duplicates and applying frozen eligibility rules, 955 English-language role-description texts remained. Each text was scored with a pinned revision of MELD, an open AI-text detector. The analysis used a job-ad-specific threshold selected before the board results were inspected. The detector was separately evaluated on known-provenance job-ad controls.

MELD directly classified **596 of 955 advertisements (62.4%)** as AI-like. This combined number is descriptive because the four feeds cover different populations. Board-level results were:

Board	Analysed	Classified AI-like	Direct rate	Detector-corrected estimate
Himalayas	411	259	63.0%	74.3%
Jobicy	460	287	62.4%	73.5%
Remote OK	73	40	54.8%	64.5%
Python.org Jobs	11	10	90.9%	100%*

*The Python.org estimate is unstable because only 11 advertisements were available.

The correction accounts for the detector's measured 84.62% sensitivity and 99.32% specificity on the evaluation controls. It raises the estimates because the detector missed more AI controls than it falsely flagged human controls. Corrected figures are model-based estimates, not verified authorship rates.

The main conclusion is that AI-like writing was common in the sampled job-board feeds. The study does **not** prove that any named employer, recruiter or individual advertisement used AI. It cannot reliably detect every form of AI assistance, especially heavily edited AI text. The detector identified only one of seven heavily edited AI controls. The most defensible public wording is therefore that a majority of sampled role descriptions were **classified as AI-like under a predefined, job-ad-calibrated detector rule**.

The study is a one-day feed snapshot, not a random sample of all advertisements posted in 2025–2026 and not a historical trend analysis. Feed ordering, caps and date semantics limit how far the results can be generalised.

Abstract

Background

The adoption of generative language models has made automated recruitment writing widely accessible. Direct measurement is difficult because job advertisements generally do not record writing provenance and automated detectors are imperfect.

Objective

To estimate the share of eligible English-language role-description texts visible through four official job-board feeds that were classified as AI-like by a frozen detector, and to adjust the observed rates for detector sensitivity and specificity measured on known-provenance job-advertisement controls.

Methods

We performed an exploratory cross-sectional study of advertisements captured on 15 October 2026 from Himalayas, Jobicy, Remote OK and Python.org Jobs. Advertisements were collected through their official public interfaces using Playwright MCP, imported into a versioned SQLite pipeline, deduplicated within each board, converted into a role-focused `job_only` text representation and frozen in a cryptographic dataset snapshot. MELD revision 8990324abd92e1fa17072f6887ea1e5c1cef5abc scored each eligible text. A threshold of 0.10739501565694809 was selected using 147 human calibration controls. Performance was then measured on an untouched evaluation set of 147 human and 52 AI controls. Board-level intervals used 20,000 employer-cluster bootstrap samples. Rogan–Gladden correction and Jeffreys-beta sensitivity/specificity draws propagated detector uncertainty. All 955 frozen extractions received structural verification, followed by blinded AI-assisted review of 50 control candidates and 40 stratified advertisement extractions.

Results

Of 955 eligible advertisements, 596 (62.4%) were classified as AI-like. Direct board rates were 63.0% for Himalayas, 62.4% for Jobicy, 54.8% for Remote OK and 90.9% for Python.org Jobs. Corrected estimates were 74.3%, 73.5%, 64.5% and 100%, respectively. The two largest samples produced similar findings. Employer-weighted and one-ad-per-employer analyses did not overturn the conclusion that AI-like writing was common. The blinded extraction audit found adequate role retention in 37 of 40 records and adequate non-role exclusion in 30 of 40, with one additional exclusion decision uncertain.

Conclusion

A majority of eligible role descriptions in the sampled feeds displayed patterns that the validated detector classified as AI-like. These results concern aggregate textual patterns, not provable authorship of individual advertisements. A larger validation benchmark, repeated collection windows and stronger coverage of human-edited AI text are needed for a confirmatory market-wide estimate.

1. Introduction

1.1 Background

Large language models are now incorporated into general-purpose assistants, office tools, applicant-tracking systems and recruitment platforms. Recruiters can use them to brainstorm role requirements, rewrite existing advertisements, standardise tone, generate complete drafts or polish human writing. These workflows blur the boundary between “human-written” and “AI-written.” A job advertisement may be fully generated, lightly edited, substantially rewritten, grammar-checked or assembled from a human template containing AI-generated sections.

The prevalence of these practices matters for at least three reasons. First, repetitive or generic AI-generated language may reduce the informational value of advertisements to job seekers. Second, automated writing can amplify patterns already present in training data, including exclusionary or vague recruitment language. Third, understanding adoption can help employers and recruitment platforms design better disclosure, editing and quality- assurance practices.

However, usage cannot normally be observed directly. Public advertisements do not include reliable provenance metadata, and asking employers retrospectively would introduce non-response and self-reporting bias. Automated AI-text detection offers a scalable but imperfect measurement approach. Such tools identify statistical resemblance to generated text; they cannot observe the actual writing process.

1.2 Research problem

The research problem was not simply to run a detector over a convenient collection of texts. A defensible estimate required:

1. job boards whose official access conditions supported the intended research use;
2. a documented and reproducible sampling frame;
3. a detector evaluated on material resembling job advertisements;
4. a threshold selected before viewing the study results;
5. adjustment for measured false positives and false negatives;
6. treatment of employers as statistical clusters;
- and 7. wording that did not convert a probabilistic text classification into an accusation of individual AI use.

1.3 Aim

The study aimed to quantify AI-like language in recent job advertisements while making the data collection, detector uncertainty and limitations explicit.

1.4 Research questions

The primary research question was:

What share of eligible English-language role-description texts visible through each included board's official feed during the observation period are classified as AI-like by the frozen MELD detector?

Secondary questions were:

1. How do the direct MELD-positive rates differ across the included feeds?
2. How do the estimates change after correction for validation sensitivity and specificity?
3. Do conclusions change when employers, rather than advertisements, receive equal weight?
4. How much do high-volume employers affect the advertisement-weighted results?

1.5 Operational definitions

AI-like / MELD-positive: A text whose MELD score is greater than the frozen threshold. This is a classifier outcome, not proof of provenance.

Observed rate: The proportion of eligible advertisements classified as AI-like.

Corrected prevalence: The observed rate adjusted using the detector's validation sensitivity and specificity.

Job-only text: The part of an advertisement containing substantive information about the role, responsibilities, requirements and directly relevant conditions, excluding company introductions and generic legal, application and recruitment-process boilerplate.

Employer cluster: All eligible advertisements from one normalised employer on one board. Advertisements lacking a usable employer name were treated as separate clusters.

2. Review of relevant evidence

2.1 AI-text detection

MELD is a multi-task detector designed to learn not only a binary human/AI distinction but also generator-family, transformation and source-domain structure. The detector's research paper reports strong low-false-positive-rate performance on held-out benchmarks and recent model families [1]. Its official model documentation recommends using inputs of approximately 100 words or more and comparing raw scores with a defined threshold or an appropriate human reference set [2].

High benchmark performance does not make AI-text detection equivalent to provenance verification. Model families, prompts, domains and editing practices change over time. Research has identified fundamental limits to perfect detection and warns that detector performance can deteriorate as generation and rewriting methods improve [3]. Stress tests also show that paraphrasing, rewriting and adversarial transformations can materially reduce machine-generated-text detection performance [4].

For this reason, the present study did not use MELD's transformed `p_ai` output as the probability that a recruiter used AI. It selected a task-specific score threshold, measured errors on job-ad controls and reported aggregate rates with uncertainty.

2.2 Correcting prevalence measured by an imperfect classifier

When an imperfect test is used to estimate prevalence, the raw positive rate reflects both true positives and false positives. The Rogan–Gladen estimator adjusts observed prevalence using the test's sensitivity and specificity [5]. It is applicable only when the detector performs better than chance—that is, when sensitivity plus specificity is greater than one.

This adjustment does not eliminate uncertainty. Sensitivity and specificity are themselves estimated from finite validation samples, so the present study propagated their uncertainty rather than treating them as known constants.

2.3 Why job-ad-specific validation was required

Job advertisements differ from essays, reviews and news articles. They commonly contain bulleted responsibilities, repeated competency language, standard equal-opportunity text, legal clauses and employer templates. A detector evaluated only on general prose could mistake these conventions for AI generation. The validation benchmark was therefore constructed from job-advertisement controls, and the same `job_only` representation was used for both controls and study advertisements.

3. Study design and governance

3.1 Design

This was an exploratory cross-sectional study. The protocol, primary text variant, eligibility rules, model revision, threshold and analysis plan were frozen before the job-board detector results were inspected. The analysis describes advertisements exposed by the selected feeds during one observation window; it is not a longitudinal trend study.

3.2 Observation period

The declared observation period was 15 October 2026 from 00:00:00 to 23:59:59 UTC. The four browser captures occurred within approximately 82 seconds:

Source	Capture time, UTC	Records captured
Jobicy	11:57:55.714	500
Himalayas	11:58:42.550	500
Remote OK	11:59:02.802	99
Python.org Jobs	11:59:18.169	20

The listing-date metadata did not have one common meaning or representation:

Source	Dated final ads	Recorded date range	Important qualification
Jobicy	460/460	11–15 October 2026	ISO timestamps from an approximately seven-day feed
Himalayas	411/411	15 October 2026, 07:28–11:16 UTC	Stored as Unix seconds in the export; interpreted here only for the range
Remote OK	73/73	2 August–14 October 2026	Latest-feed inventory included older listings
Python.org Jobs	0/11	Unavailable	Publication dates were not reliable enough to analyse

Consequently, eligibility was based on visibility in the official feed during the capture window, not on a claim that every advertisement was first published on the observation day.

3.3 Source selection and access review

Sources were included only when an official public feed was available and the documented access basis supported the research. The access review was rechecked on 15 October 2026.

Source	Decision	Recorded basis
Jobicy	Included	Public feed documentation permits research and AI-product use; publication must name Jobicy and link the canonical listing [6].
Himalayas	Included	Public API documentation permits use, including by AI agents; publication must name and link Himalayas, and listings must not be republished to another job site [7].
Remote OK	Included	The official API requires naming Remote OK and linking each reused listing with a follow link; its logo must not be used [8].
Python.org Jobs	Included	The PSF legal statement grants website users a licence covering reproduction and publication of third-party content [9].
Workable	Excluded	A public XML feed existed, but the terms did not clearly grant an external commercial researcher the intended reuse rights. Written permission was not obtained.

This assessment records the study team's research decision and is not legal advice. The report publishes aggregate findings only and does not reproduce full job descriptions.

3.4 Common estimand

The common estimand was:

Job advertisements visible through each board's official feed during the declared observation period, subject to the documented coverage limits of that feed.

The feeds did not cover identical populations. Jobicy represented approximately seven days of listings; Himalayas exposed a paginated inventory; Remote OK exposed a capped latest-job feed; and Python.org exposed a small latest-items feed. Board-level estimates were therefore kept separate.

3.5 Ethics and risk controls

The analysis concerned publicly available organisational text, not candidates or private individual data. Nevertheless, false attribution could harm employers or recruiters. The study adopted the following protections:

- no employer-level rankings or accusations;
- no claim that an individual advertisement was definitively generated by AI;
- no public redistribution of complete advertisement text;
- aggregate reporting by board;
- explicit disclosure of detector uncertainty and failure modes;
- preserved provenance and hashes for audit without publishing private control evidence;
- source attribution consistent with each feed's documented requirements.

HireMike develops AI-assisted recruitment technology and therefore has a commercial interest in recruitment workflows. This potential conflict should be disclosed in any public version. The study received no external funding.

4. Data collection and preparation

4.1 Collection procedure

The study used Playwright MCP to access the boards' official public interfaces in a normal browser context. Each response was preserved as a timestamped JSON capture. The captures were then imported into the research pipeline rather than repeatedly requesting the boards. This separated source access from later processing and allowed the exact observed data to be replayed.

The capture was deterministic but not probabilistic. Jobicy and Himalayas were capped at the first 500 records exposed through their official interfaces, while all available records in the smaller Remote OK and Python.org captures were taken. No random draw from a complete board inventory was possible. The resulting estimates describe the captured feed frames, not every active vacancy on each platform.

The import stored source identifiers, canonical URLs, titles, employer names, locations, listing dates, capture times and complete payload hashes in SQLite. Each collection run recorded its source, page count, number seen, duplicates and completion status.

4.2 Initial collection outcomes

Source	Seen	Same-board duplicates	Cross-board occurrences	Selected after initial deduplication	Final eligible sample
Jobicy	500	33	0	467	460
Himalayas	500	26	2	474	411
Remote OK	99	1	0	98	73
Python.org Jobs	20	1	0	19	11
Total	1,119	61	2	1,058	955

The 103 post-deduplication exclusions are fully accounted for below. Reasons are mutually exclusive because the pipeline applies them in the displayed decision order.

Source	Non-English/low confidence	Source-truncated	Incomplete built text	Fewer than 100 job_only words	Total excluded after deduplication
Jobicy	2	0	1	4	7
Himalayas	40	0	5	18	63
Remote OK	7	8	5	5	25
Python.org Jobs	0	0	0	8	8
Total	49	8	11	35	103

Cross-board occurrences remained eligible for each board-specific denominator. The two Himalayas occurrences matching material on another board were therefore retained.

4.3 Eligibility criteria

An advertisement entered the frozen analysis sample only when it:

1. belonged to a successful collection run for the declared study; 2. had a payload captured inside the study window; 3. was the latest version captured inside that window; 4. was classified as English with confidence of at least 0.90; 5. had a complete and non-source-truncated description; 6. was not a same-board duplicate; 7. produced non-empty raw and job_only text representations; and 8. contained at least 100 canonical words in the primary job_only representation.

The snapshot process applied these rules before scoring. All 955 texts admitted to the snapshot subsequently had valid scores; no advertisement was removed after scoring because of missing output or insufficient length.

4.4 Deduplication

Duplicate detection compared normalised word content rather than markup. Case, links, quotation styles, dashes, whitespace and invisible formatting characters did not determine a match. Repeated board-wide boilerplate was excluded from the similarity comparison so that generic footer text could not make unrelated advertisements appear duplicated.

Within-board duplicates were excluded because counting reposts twice would inflate that board's denominator. Cross-board matches were retained once on every board where they appeared because the primary question concerned each board's visible inventory.

4.5 Text segmentation and role-text construction

The 1,119 imported advertisements were divided into 23,119 text blocks. Deterministic rules and repeated-text patterns assigned blocks to role, company, benefits, application, legal, bot text, other or unknown categories. No LLM API labels were used during this board pass. The label counts were:

Block category	Blocks
Role	10,996
Company	3,351
Benefits	3,166
Legal	1,645
Application	541
Other	336
Bot text	9
Unknown	3,075

The pipeline constructed two immutable representations per advertisement:

- raw: the complete normalised advertisement; and
- job_only: the substantive role-description material selected by the frozen rules.

Text was selected, not paraphrased or rewritten. This matters because model-generated language was not introduced during preprocessing.

4.6 Dataset freeze

After segmentation, labelling and eligibility screening, the pipeline froze:

- study ID and observation window;
- duplicate scope;
- payload policy and exact payload IDs;
- SHA-256 of every source payload;
- SHA-256 of every raw and job_only text;
- pipeline component versions; and
- the ordered set of 955 eligible job IDs.

The resulting dataset snapshot hash was

636860f4ca04f31da698f93494b9210f7a92009e0e2cad4743b7c5aef3127bdd. Any later payload or text change would cause the bound export to fail rather than silently mix versions.

4.7 Blinded quality audit

After primary scoring, a deterministic hash-ranked audit was conducted without exposing the reviewer to control IDs, calibration/evaluation assignment, job-board identity, employer identity, detector scores or detector classifications. The audit included:

- exhaustive structural verification of all 955 frozen advertisements;
- 50 control candidates: 40 accepted human controls and 10 candidates rejected by the frozen

eligibility rules, mixed in a blinded packet; and

- 40 extracted advertisements, stratified as 10 per board, comparing raw and job_only

text without board or employer identifiers.

The reviewer evaluated control eligibility, retention of substantive role material and exclusion of non-role material. Decisions were recorded before the private key was opened. The reviewer was an OpenAI Codex research agent, not an independent human reviewer. Private packets remain git-ignored; the public audit report contains aggregate results and hashes only.

5. Detector-validation methodology

5.1 Detector selection

MELD was selected because it was openly available, ran locally, exposed reproducible raw scores and documented low-false-positive operating points. Local execution avoided sending employer text to a third-party detection service. The study pinned the exact model, inference file and dependency versions.

5.2 Control definitions

Controls were labelled by known provenance, never by another detector.

- **Human:** substantive wording was produced without a generative language model.
- **AI:** a generative model produced or rewrote substantive wording.

- **Light edit:** local human changes retained the generated structure and most sentences.
- **Heavy edit:** a human substantially restructured and rewrote a generated draft while

retaining the underlying role facts.

Related employer templates and generation families were grouped so that related texts could not cross from calibration into evaluation.

5.3 Human controls

The frozen exploratory benchmark used 147 human controls for threshold calibration and a separate 147 human controls for evaluation. These were selected from archived pre-generative- AI job-ad data. The underlying candidate archives were recorded as CC0/Public Domain by their distribution pages and covered the CareerCenter 2004–2015 collection and legitimate EMSCAD-era advertisements. Deterministic screening removed missing-provenance records, non-English or short texts, redacted placeholders, raw contact details, spreadsheet-error tokens and invisible extraction characters.

The use of historical human controls protects against undisclosed modern AI use, but it may also introduce temporal and geographic domain differences. This is treated as a limitation; the benchmark did not contain enough directly attested contemporary human advertisements for confirmatory use.

Candidate controls passed deterministic provenance, language, length, contact-information and extraction-quality screens. A post-analysis blinded audit mixed 40 accepted controls with 10 rejected candidates and agreed with all 50 frozen eligibility decisions. Their historical dates make generative-AI authorship implausible, but the study did not individually verify every original author, and the audit reviewer was an AI research agent rather than an independent human.

5.4 AI controls

The untouched AI evaluation set contained 52 controls:

Generator family	Controls	Detected	Sensitivity
OpenAI Codex GPT-5 family	30	22	73.3%
Qwen3 0.6B	13	13	100%
SmolLM2 360M	9	9	100%
Total	52	44	84.6%

Prompts, generation settings, model revisions, raw outputs and hashes were preserved for generated controls. The local Qwen and SmolLM files used pinned revisions and deterministic decoding. The threshold was selected only from human calibration controls; calibration AI examples were descriptive and did not determine the operating point.

The light- and heavy-edit examples came from only one generator family. Their edit-level results therefore demonstrate a failure mode but cannot be assumed to represent editing of all model families.

5.5 Editing, format and length coverage

AI subgroup	Controls	Detected	Sensitivity
Unedited	38	36	94.7%
Lightly edited	7	7	100%
Heavily edited	7	1	14.3%
List format	12	12	100%
Mixed format	20	14	70.0%
Prose format	20	18	90.0%
Long	11	11	100%
Short	41	33	80.5%

The small subgroup denominators make these estimates imprecise. The heavy-edit result is nonetheless a central warning: the study cannot estimate every form of historical AI involvement.

5.6 Threshold selection

The target was a low false-positive operating point. Human calibration scores were ordered, and the threshold was selected using only the 147 calibration humans. One calibration human exceeded the resulting threshold, producing an observed calibration false-positive rate of 0.68%.

The 147 human calibration controls were sufficient to define an exploratory low-false-positive operating point, but they provide limited precision for estimating a tail probability near 1%. This threshold should therefore be treated as a job-ad-specific rule for this study, not as a permanently established 1% false-positive threshold for all recruitment text.

The frozen rule was:

Classify a job_only text as AI-like when MELD score is greater than 0.10739501565694809.

The inequality is strict: a score equal to the threshold is not positive. The threshold was not changed after board outcomes were inspected.

5.7 Untouched evaluation performance

Evaluation outcome	Count
True positives	44
False negatives	8
True negatives	146
False positives	1

Metric	Estimate	95% Wilson interval
Sensitivity / recall	84.62%	72.48%–91.99%
Specificity	99.32%	96.25%–99.88%
False-positive rate	0.68%	0.12%–3.75%
Precision in the constructed evaluation mixture	97.78%	88.43%–99.61%

Precision is not transferable to the board sample because it depends on the proportion of AI and human controls chosen for evaluation. Sensitivity and specificity were used for prevalence correction.

6. Scoring and statistical analysis

6.1 Primary scoring run

The detector ran locally on CPU and loaded the model once. Each of the 955 frozen `job_only` texts was canonicalised and scored under the same settings. Results were keyed by the SHA-256 of the canonical input and the settings hash. Scoring run 3 completed all 955 expected pairs with zero errors.

The study analysed MELD's raw score. It did not interpret `p_ai` as a calibrated probability of recruiter behaviour.

6.2 Primary estimator

For each board, the advertisement-weighted MELD-positive rate was:

`number of advertisements with score > threshold / number of eligible advertisements`

Every advertisement received equal weight. This answers the question: "What share of the advertisements visible in this feed were classified as AI-like?"

6.3 Employer clustering

Advertisements from one employer are not statistically independent. They may reuse a template or be produced by the same recruitment team. Confidence intervals therefore resampled normalised employers with replacement and retained all of each sampled employer's advertisements. Unknown employer names were not combined into one artificial cluster.

6.4 Secondary estimators

Two secondary analyses were prespecified:

1. **Employer-weighted:** calculate each employer's positive share, then average employers equally. 2. **One ad per employer:** select one advertisement per employer using the lowest SHA-256 key formed from the frozen seed, source, employer key and stable advertisement ID.

These answer different questions. Advertisement weighting describes visible inventory; employer weighting describes the average employer in the observed feed.

6.5 Detector-error correction

The Rogan–Gladen estimator was applied to each observed rate:

`p_corrected = (p_observed + specificity - 1) / (sensitivity + specificity - 1)`

With sensitivity 0.8462 and specificity 0.9932, the denominator was positive, so the detector performed sufficiently above chance for correction. Estimates were bounded to the logical 0–1 range.

6.6 Uncertainty intervals

For each board, 20,000 simulations were produced with seed 1:

1. resample employer clusters with replacement; 2. recalculate advertisement-, employer- and one-ad-per-employer rates; 3. draw sensitivity from `Beta(44.5, 8.5)`; 4. draw specificity from `Beta(146.5, 1.5)`; 5. apply the prevalence correction; and 6. take the 2.5th and 97.5th percentiles.

Observed-rate intervals therefore reflect employer clustering. Corrected intervals reflect both employer clustering and finite detector-validation uncertainty.

6.7 Analysis safeguards

The publication export was required to name the study, snapshot, scoring run, settings hash, model revision, threshold and calibration file. The command failed if these identifiers did not agree or if any primary score was missing. The estimator independently checked the binding before producing the final JSON report.

7. Findings

7.1 Sample composition

The final sample contained 955 advertisements from 526 board-specific employer clusters. Because the same employer may appear on more than one board, these cluster counts should not be interpreted as 526 unique legal organisations across the market.

Board	Ads	Employer clusters	Employers with multiple ads	Maximum ads from one employer	Largest employer share
Himalayas	411	278	64	19	4.6%
Jobicy	460	169	104	12	2.6%
Remote OK	73	69	4	2	2.7%
Python.org Jobs	11	10	1	2	18.2%

Jobicy had the greatest employer repetition: 104 of 169 observed employers contributed more than one advertisement. Himalayas' largest employer contributed 19 advertisements, but this was still less than 5% of the board sample.

All primary inputs exceeded the detector's recommended minimum length, and none was truncated by the detector:

Board	Minimum words	Median words	Maximum words	Detector-truncated texts
Himalayas	100	353	1,654	0
Jobicy	109	490	1,408	0
Remote OK	103	432	1,346	0
Python.org Jobs	124	259	612	0

7.2 Primary results

Board	Ads analysed	Positive ads	MELD-positive rate	95% cluster interval	Corrected prevalence	95% corrected interval
Himalayas	411	259	63.0%	57.2%–68.5%	74.3%	64.9%–88.0%
Jobicy	460	287	62.4%	56.7%–68.1%	73.5%	64.2%–87.5%
Remote OK	73	40	54.8%	43.7%–65.8%	64.5%	50.3%–81.7%
Python.org Jobs	11	10	90.9%	70.0%–100%	100%	81.7%–100%

Himalayas and Jobicy, the two largest samples, produced nearly identical direct rates. The Remote OK point estimate was lower, but its wider interval overlapped those of both larger boards. The Python.org estimate was highest, but the sample of 11 was too small for a stable board comparison.

7.3 Descriptive total

Across all four feeds, **596 of 955 advertisements (62.4%)** were MELD-positive. Applying the same sensitivity/specificity correction produces **73.5%**. This is a useful descriptive summary of the collected records, but it is not the study's primary estimand and must not be labelled a market-wide prevalence estimate. The boards were not sampled with market-share weights and did not expose equivalent inventory windows.

7.4 Employer-weighted results

Board	Employer-weighted direct rate	95% interval	Corrected estimate	95% corrected interval
Himalayas	59.6%	54.1%–65.1%	70.2%	61.3%–83.3%
Jobicy	64.8%	58.7%–70.7%	76.4%	66.8%–91.0%
Remote OK	54.3%	42.8%–65.9%	63.9%	49.4%–81.6%
Python.org Jobs	90.0%	70.0%–100%	100%	79.4%–100%

On Himalayas, employer weighting reduced the direct rate by 3.4 percentage points. Higher- volume employers therefore contributed somewhat more AI-like advertisements than lower- volume employers. On Jobicy, employer weighting raised the rate by 2.4 points, indicating that lower-volume employers were slightly more AI-like on average. Remote OK and Python.org changed by less than one point.

7.5 One-ad-per-employer analysis

Board	Employers/selected ads	Positive selected ads	Direct rate	Corrected estimate
Himalayas	278	166	59.7%	70.3%
Jobicy	169	110	65.1%	76.7%
Remote OK	69	37	53.6%	63.1%
Python.org Jobs	10	9	90.0%	100%

This analysis reached the same substantive conclusion as the employer-weighted analysis. The main result was therefore not created solely by a few employers posting many similar advertisements.

7.6 Why corrected estimates are higher

The validation set contained one false positive but eight false negatives. Specificity was therefore higher than sensitivity. At the board rates observed here, the estimated number of missed AI-like texts was larger than the number of human texts expected to be falsely flagged. The Rogan–Gladen correction consequently raised the estimated prevalence.

The corrected values should not be presented as more certain than the direct rates. They are conditional on the assumption that validation sensitivity and specificity transfer to the board sample. Their wider intervals reflect this extra uncertainty.

7.7 Blinded audit findings

The exhaustive mechanical audit passed all 955 frozen records: raw text matched the frozen Markdown, job_ on ly text matched the ordered role-labelled blocks, both text hashes matched the snapshot, all blocks had labels and every analysed text met the 100-word minimum.

The blinded control audit agreed with all 50 frozen eligibility decisions. In the 40-ad stratified extraction sample, 37 retained adequate substantive role material and three omitted material duties or requirements. Thirty adequately excluded non-role material, nine retained clear benefits, company/culture, application, privacy, legal, fraud-warning or unrelated-role material, and one was uncertain. These descriptive audit proportions should not be interpreted as population error rates. They show that the extraction was mechanically reproducible but not semantically perfect; the likely direction of any effect on detector scores is unknown.

8. Discussion

8.1 Principal finding

The clearest finding is that AI-like language was not rare. In the two large feeds, roughly three in five eligible advertisements crossed a deliberately low-false-positive threshold. Remote OK also had a majority-positive point estimate. These direct results alone are substantially above the 1% human calibration false-positive target.

The results support the interpretation that generative writing patterns are widespread in contemporary recruitment content. They do not reveal whether this arose from complete AI generation, partial drafting, rewriting, template tools or convergence between professional human writing and model style.

8.2 Comparison between boards

Himalayas and Jobicy were similar despite different API structures and employer mixes. This agreement strengthens confidence that the result is not unique to a single feed. Remote OK was lower, but its uncertainty interval did not establish a clear difference. Python.org cannot support a strong comparison because one additional negative advertisement would have changed its direct rate by more than nine percentage points.

No causal explanation can be assigned to board differences. Possible contributors include employer composition, role categories, feed recency, geographic mix, syndication, template use and the proportion of high-volume recruiters.

8.3 Meaning of “AI-like”

The result describes detector-classified language, not an audit trail. A positive text may have been generated by AI, heavily influenced by an AI template, or written by a human in a highly regularised style. A negative text may still have involved AI, especially if a human substantially rewrote it. The heavy-edit validation sensitivity of 14.3% demonstrates why a negative score cannot be treated as evidence of no AI use.

8.4 Practical implications

For employers, the findings suggest that simply generating a polished advertisement no longer differentiates a role. Clear evidence about actual responsibilities, decision rights, success measures, compensation and working conditions is likely to be more useful than generic fluency.

For job boards, widespread templated language may justify provenance features, stronger quality checks and prompts encouraging employers to replace generic generated passages with verifiable role detail.

For researchers, the study demonstrates that detector outputs should be calibrated in the target domain and reported with error correction rather than presented as literal probabilities of AI use.

9. Limitations

9.1 No direct authorship evidence

The study did not contact the employers responsible for the sampled advertisements. It therefore cannot verify the writing history of any board text. Detector classifications are indirect measurements.

9.2 Imperfect and evolving detector

MELD performance can change under domain shift, new generators and new editing tools. The model's general benchmark performance does not guarantee identical performance on all job advertisements.

9.3 Heavy editing

Only one of seven heavily edited AI controls was detected. The study likely undercounts AI involvement that has been substantially rewritten by humans. Conversely, highly formulaic human texts may occasionally be false positives.

9.4 Exploratory validation size

The 52 AI and 147 human evaluation controls were sufficient for an exploratory go/no-go analysis but not a definitive detector characterisation. Subgroup intervals, particularly for edit level and generator family, are wide. A confirmatory study should use at least 500 evaluation humans and 500 evaluation AI controls with balanced generator and editing coverage.

9.5 Historical human controls

Evaluation humans were drawn from archived pre-generative-AI data. This provides clear provenance but may allow the detector to exploit changes in recruitment language associated with time, region or platform. Contemporary human controls with direct author attestations were not sufficiently represented.

9.6 AI-control coverage

The AI controls covered three model families, and most were unedited. They did not represent every current commercial model, system prompt, temperature, recruiter workflow or AI polishing tool.

9.7 One-day collection

The cross-sectional snapshot cannot measure trends, seasonality or day-to-day feed changes. It represents what the official feeds exposed during the declared observation period. It is not a historical average for 2025–2026 and cannot support a claim that AI-like writing increased, decreased or remained stable during those years.

9.8 Unequal sampling frames

The boards differed in recency, pagination and feed size. Their estimates should not be pooled as though every advertisement in a common market had an equal probability of selection. The larger interfaces supplied the first 500 exposed records rather than a probability sample; the two smaller feeds supplied all records available at capture. Feed ordering, API caps and unobserved coverage may therefore affect the board estimates. The reported confidence intervals quantify sampling variation under the analysis model, not this coverage bias.

9.9 Small Python.org sample

The 11-ad sample produces an unstable rate and a corrected point estimate bounded at 100%. It should be reported descriptively, not treated as evidence that Python.org differs from the other boards.

9.10 Text-preparation classification

The deterministic block labelling rules may retain some boilerplate or remove some substantive role text. They do not rewrite text, but selection errors can affect the input seen by the detector. The blinded 40-ad audit identified three material-retention failures, nine clear non-role-exclusion failures and one uncertain exclusion. Correcting and rescoreing those records after inspecting primary outcomes would alter the frozen analysis, so the study reports the audit transparently instead of silently changing the inputs.

9.11 Protocol deviation

The frozen protocol listed a full-advertisement raw sensitivity analysis. It was stopped after the primary results because completing a second CPU scoring pass would have added substantial time without being required for the primary question. No completed raw scores were used, and the sample, threshold and job_only result were not changed. This deviation must remain disclosed in any publication.

9.12 Generalisability

The findings apply to the included official feeds and observation date. They do not directly estimate offline vacancies, private applicant-tracking systems, paid-only boards, non-English advertisements or the complete global recruitment market.

9.13 Unmeasured detector-transfer uncertainty

The corrected prevalence intervals propagate uncertainty in the evaluation-set sensitivity and specificity, but they cannot quantify unknown domain shift between the constructed controls and the live board advertisements. If detector performance differs materially on the board texts, the corrected estimates may still be biased.

9.14 Audit-reviewer independence

A blinded audit was completed and recorded, but its reviewer was an OpenAI Codex research agent rather than an independent human. It therefore provides reproducible methodological quality control, not independent human confirmation. The original control authors were not individually contacted. An independent human replication of the private audit packets remains recommended before peer-reviewed publication.

9.15 Code-state provenance

The analysis logs recorded base commit 04015d1e74b17b86fa52bacaed832219e0bfcd87-dirty. After analysis and audit, the complete repository state was committed and marked with the annotated tag hiremike-ai-job-ads-2026-10-v1. That tag contains the pipeline, protocols, reports and audit utility. The database log still records the pre-commit dirty identifier rather than the later tag, so the contemporaneous run-to-commit link relies on the preserved artifact hashes and repository history and must not be described as a pre-registered clean-code run.

9.16 Unknown detector-training overlap

The public model materials do not establish that every validation control was absent from all data used to pretrain or train the detector and its underlying language model. The controls were kept out of this study's own threshold fitting where required, but undetected upstream training-data overlap cannot be ruled out and could make evaluation performance appear better than performance on wholly novel texts.

10. Conclusions

The study found that AI-like writing was common in a frozen sample of contemporary online job advertisements. MELD classified 596 of 955 eligible role descriptions—62.4%—as AI-like. The two largest board samples produced direct rates of 63.0% and 62.4%. After adjusting for validation sensitivity and specificity, their estimates rose to 74.3% and 73.5%.

The evidence supports a careful aggregate claim:

In HireMike's 15 October 2026 sample, a majority of eligible role-description texts were classified as AI-like by MELD at a threshold fixed on job-ad controls.

The evidence does not support saying that 73.5% of recruiters definitely used AI or that a particular advertisement was AI-written. The detector measures textual resemblance, and its ability to detect heavily edited AI material was weak.

The research therefore provides a credible exploratory estimate and a reproducible method, not a definitive census of recruitment AI use.

11. Recommendations

11.1 For publication of the present study

1. Lead with board-specific direct MELD-positive rates. 2. Present corrected estimates as conditional estimates, not observed facts. 3. Keep Python.org clearly labelled as a very small sample. 4. Describe the 596/955 total as descriptive, not market-representative. 5. State the heavy-edit failure prominently. 6. Do not identify or rank individual employers. 7. Publish the protocol, hashes, aggregate JSON results and analysis code, but not complete advertisement text. 8. Publish the aggregate blinded-audit report and disclose its extraction findings. 9. Obtain an independent human replication of the private blinded audit before peer-reviewed submission.

11.2 For a confirmatory follow-up

1. Repeat collection across multiple weeks or months. 2. Recruit at least 500 directly attested contemporary human controls. 3. Use at least 500 AI evaluation controls balanced across major current model families. 4. Commission blinded human light and heavy edits across multiple generator families. 5. Predefine any board comparisons and multiple-comparison treatment. 6. Obtain permission before adding sources whose terms do not clearly authorise reuse. 7. Preserve a clean, committed code revision before scoring the confirmatory sample.

12. Reproducibility and audit record

12.1 Frozen identifiers

Item	Value
Study ID	hiremike-ai-job-ads-2026-10
Dataset snapshot	636860f4ca04f31da698f93494b9210f7a92009e0e2cad4743b7c5aef3127bdd
Primary scoring run	3
Analysis variant	job_only
Model revision	8990324abd92e1fa17072f6887ea1e5c1cef5abc
Model file SHA-256	86700f566059951fc9326a6b04234256d31691b61b4c0a738422616f0d62d303
MELD inference-file SHA-256	7236347d37c1cdebe1a72639955bd8ac76f5ff41cbe946bdad22d6681988bac7
Settings hash	00765fdede2123f5cbb2b41086d2d910774a6a89a779e1c25f0ea91fbf186f9d
Threshold	0.10739501565694809
Controls SHA-256	a30fa1c6450aa0afea3f3b95b3459b2fe5f600a1cc7cc483df346099369898d9
Calibration SHA-256	a032231564ceca855ca70df88570015675403c5d5e0acd13bb5fad1acf1e3a55
Export SHA-256	bc47aff4e4a5177bf02ea90c5fb45405ecba344fe0093d7f82be15b35a9af8d9
Bootstrap draws	20,000
Bootstrap seed	1
Payload policy	latest_captured
Duplicate scope	within_board

12.2 Runtime versions

Dependency	Version
PyTorch	2.14.1+cpu
Transformers	5.18.0
Safetensors	0.8.0
Tokenizers	0.23.2
Device	CPU
Numerical precision	FP32; autocast disabled

12.3 Final files

- Bound advertisement export:

exports/ads-hiremike-ai-job-ads-2026-10-within_board-20261016T182302Z.csv

- Export manifest:

exports/ads-hiremike-ai-job-ads-2026-10-within_board-20261016T182302Z.csv.manifest.json

- Final prevalence report:

exports/prevalence-hiremike-ai-job-ads-2026-10-final.json

- Frozen study protocol: STUDY_PROTOCOL.md
- Detector-validation protocol: VALIDATION.md
- Aggregate blinded-audit report: PUBLICATION_AUDIT.md
- Machine-readable blinded-audit result: PUBLICATION_AUDIT.json

The bound row-level CSV contains employer names, job titles, source URLs and detector scores. It should remain a controlled research artifact rather than a public accusation list. A public supplement should favour the aggregate JSON results, protocols, artifact hashes and analysis code, subject to each board's attribution and republication conditions.

12.4 Source code

The analysis software is maintained at github.com/inprod/job-ad-ai-scoring. The publication archive is marked by tag hiremike-ai-job-ads-2026-10-v1; see Section 9.15 for the qualification concerning the scoring run's earlier dirty identifier.

12.5 Reproduction sequence

The analysis proceeded in the following order:

1. create and freeze the study definition;
2. capture each official feed with Playwright MCP;
3. import the four immutable JSON captures;
4. segment advertisement text into blocks;
5. classify blocks and build raw/job_only texts;
6. freeze the within-board-deduplicated dataset snapshot;
7. calibrate and evaluate MELD on the frozen control benchmark;
8. score all snapshot job_only texts under one settings hash;
9. export results bound to the study, snapshot, run, model, threshold and calibration;
10. estimate board prevalence with employer-cluster and detector uncertainty;
11. run exhaustive structural checks and record blinded control/extraction decisions; and
12. unblind the decisions and publish aggregate audit results and artifact hashes.

The complete automated test suite passed after finalisation: **99 passed and 14 skipped**.

13. Suggested public-language summary

HireMike analysed 955 English-language role descriptions collected from the official feeds of Himalayas, Jobicy, Remote OK and Python.org Jobs on 15 October 2026. Using a detector threshold fixed and evaluated on job-ad controls, 62.4% of the sampled texts were classified as AI-like. The rates were 63.0% on Himalayas, 62.4% on Jobicy and 54.8% on Remote OK; Python.org had only 11 eligible advertisements and is too small for a stable comparison. After correcting for the detector's measured sensitivity and specificity, estimates were higher. These are aggregate text-classification results, not proof that any individual recruiter or advertisement used AI.

References

1. Li, C., Wan, C. and Paetzold, J. C. (2026). MELD: Multi-Task Equilibrated Learning Detector for AI-Generated Text. [arXiv:2605.06903](#). 2. MELD model documentation. [anon-review-meld-2026/meld](#). [Hugging Face model card](#). 3. Perfect detection of computer-generated text faces fundamental challenges (2024). Patterns, 5(1), 101769. DOI: 10.1016/j.xcrp.2023.101769. 4. Wang et al. (2024). Stumbling Blocks: Stress Testing the Robustness of Machine-Generated Text Detectors Under Attacks. Proceedings of ACL 2024. [ACL Anthology](#). 5. The Rogan–Gladen estimator for outcome misclassification. [PubMed Central](#). 6. Jobicy. Jobs RSS Feed and Public API Documentation. <https://jobicy.com/jobs-rss-feed>. 7. Himalayas. Public Jobs API. <https://himalayas.app/api>. 8. Remote OK. Official API. <https://remoteok.com/api>. 9. Python Software Foundation. Legal Statements. <https://www.python.org/about/legal/>. 10. Workable. Terms of Service. <https://www.workable.com/terms>. 11. HireMike. Job Ad AI Scoring research pipeline. <https://github.com/inprod/job-ad-ai-scoring>.

Appendix A. Interpretation guide

Term	Correct interpretation	Incorrect interpretation
MELD-positive	Text crossed the frozen detector threshold	Proven AI authorship
62.4% combined	Descriptive share across collected records	Global job-market prevalence
Corrected prevalence	Estimate conditional on validation transfer	Exact known percentage
Negative score	Detector did not find strong AI-like signals	Proof that no AI was used
Employer-weighted rate	Average observed employer's positive share	Intrinsically “truer” rate

Appendix B. Board-specific reporting cautions

Himalayas: The 411-ad sample was large enough for a reasonably informative exploratory interval. Two cross-board occurrences were intentionally retained.

Jobicy: The 460-ad sample was the largest, but 104 of 169 employers posted multiple advertisements; clustered analysis was therefore essential.

Remote OK: Only 73 advertisements met the final rules, so estimates were less precise.

Python.org Jobs: Only 11 advertisements were eligible. Its result is descriptive and should not be used to rank Python.org against the other boards.

Appendix C. Publication checklist

- ☐ Use “classified as AI-like,” not “proved to be written by AI.”
- ☐ Report denominators beside every percentage.
- ☐ Show direct and corrected estimates together.
- ☐ Include confidence/uncertainty intervals.
- ☐ Disclose the 14.3% heavy-edit sensitivity.
- ☐ Disclose the omitted raw-text sensitivity analysis.
- ☐ Attribute and link every source board.
- ☐ Do not publish complete job descriptions or employer-level accusations.
- ☐ Keep the row-level scored CSV under controlled access.
- ☐ Publish the snapshot, model, settings, threshold and artifact hashes.
- ☒ Complete and document a blinded control/extraction audit.
- ☐ Replicate the private blinded audit with an independent human reviewer.
- ☒ Commit and tag the publication code state.