# A Comparative Study of Reliability for Automated CV Screening

---

## Whitepaper

---

Mimo – AI Recruiter
Jl. Tulodong Atas No. 28
Kebayoran Baru
Kota Jakarta Selatan
DKI Jakarta
12190
Phone: +6281944247722
Email: hello@mimo.id

# Daftar Isi

# 1 Abstrak

Studi ini membandingkan reliabilitas dua pendekatan penyaringan CV berbasis kecerdasan buatan pada sistem Mimo.id *Embedding-Based Scoring* (EBS) yang menggunakan representasi vektor semantik, dan *Prompt-Based Scoring* (PBS) yang memanfaatkan *Large Language Model* (LLM) sebagai penilai otomatis. Penelitian eksperimental komparatif ini melibatkan 200 CV anonim, diuji melalui 30 pengulangan untuk mengestimasi konsistensi temporal, dengan analisis menggunakan Spearman's ρ, Intraclass Correlation Coefficient (ICC) [model two-way random, absolute agreement], dan Kendall's τ. Hasil menunjukkan EBS mencapai reliabilitas sempurna (ρ = 1.000; ICC = 1.000; τ = 0.998–1.000), mencerminkan determinisme, sementara PBS menunjukkan reliabilitas excellent (ρ = 0.76–0.85; ICC = 0.906; τ = 0.57–0.69), dengan variabilitas stokastik yang dapat diterima secara psikometrik. Temuan ini menegaskan bahwa pendekatan EBS lebih stabilitas untuk rekrutmen massal. Hasil ini merekomendasikan penggunaan teknik perhitungan EBS berbasis otomatisasi untuk mengedepankan aspek reliabilitas asesmen guna menghindari error.

Keywords: AI recruitment, CV screening, reliability, embeddings, large language models, psychometrics

# 2 Pendahuluan

Pertumbuhan angkatan kerja di Indonesia menimbulkan tantangan serius dalam praktik rekrutmen. Data Badan Pusat Statistik (BPS) pada Agustus 2025 mencatat bahwa jumlah angkatan kerja mencapai sekitar 140 juta orang, dengan tingkat pengangguran terbuka (TPT) berada di kisaran 5,5%. Kondisi ini menunjukkan tingginya prevalensi pencari kerja yang secara langsung berdampak pada beban operasional perusahaan dalam melakukan rekrutmen. Tahap awal seleksi, yakni penyaringan Curriculum Vitae (CV), kerap menjadi bottleneck karena volume lamaran yang tinggi. Survei internal HR menunjukkan bahwa 50–70% dari total waktu rekrutmen dihabiskan hanya untuk menyeleksi CV pada perusahaan skala menengah dan besar. Situasi ini berpotensi menghambat efektivitas proses rekrutmen serta meningkatkan risiko terlewatnya kandidat yang sebenarnya berkualitas

Untuk mengatasi tantangan tersebut, sejumlah applicant tracking systems (ATS) dan perangkat otomatisasi penyaringan CV telah banyak dikembangkan. Beberapa contoh di antaranya adalah Greenhouse, Workable, Lever, Zoho Recruit, SmartRecruiters, serta Mimo yang berasal dari Indonesia. Sistem ini dirancang untuk mempercepat proses seleksi melalui algoritma otomatisasi, mulai dari pengumpulan lamaran, penyaringan kata kunci, hingga pemberian skor. Meskipun demikian, tidak ada konsensus ilmiah mengenai metode skoring otomatis mana yang paling efektif serta reliabel jika ditinjau dari sudut pandang psikometrik. Dengan kata lain, ada kebutuhan mendesak untuk menguji properti reliabilitas dari metode skoring berbasis AI yang saat ini tersedia.

Perkembangan Large Language Models (LLM) memberikan alternatif baru dalam praktik evaluasi teks. Sejumlah riset menunjukkan potensi LLM untuk melakukan penilaian otomatis terhadap teks seperti esai, jawaban terbuka, atau dokumen profesional. Chafiq, et. al. (2025) misalnya, menyoroti bahwa LLM mampu menghasilkan evaluasi yang konsisten dengan standar penilaian manusia jika diarahkan melalui instruksi yang tepat (Fisher et al., 2024). Studi-studi lain dalam domain medis dan pendidikan juga menunjukkan kemampuan LLM dalam memberikan skor yang sebanding dengan penilai manusia, meskipun masih terdapat tantangan terkait reliabilitas dan sensitivitas terhadap desain prompt (Zheng et al, 2025; Dou et al., 2025).

AI Resume Screening adalah urutan komputasi yang secara efisien memvalidasi dan memberi peringkat pada ribuan lamaran, berfungsi sebagai gerbang pertama dalam corong penyaringan Mimo. Secara teknis, platform ini berbeda secara fundamental dari Applicant Tracking System (ATS) tradisional. Sementara ATS hanya menyaring berdasarkan pencocokan kata kunci, AI Mimo menggunakan Natural Language Processing (NLP) untuk melakukan analisis semantik. Kemampuan ini memungkinkan sistem untuk mengidentifikasi konsep, mengenali entitas seperti nama perusahaan atau keterampilan teknis, dan memahami konteks dari pengalaman kerja. Sebagai contoh, AI dapat memahami bahwa "memimpin tim pengembangan produk dari konsepsi hingga peluncuran" adalah indikator kuat dari pengalaman manajemen proyek, bahkan jika frasa tersebut tidak secara eksplisit disebutkan.

Hasilnya bukan sekadar daftar lolos/tidak lolos, melainkan peringkat dinamis seperti Very Fit atau Fairly Fit, yang disertai ringkasan naratif. Ringkasan ini menyoroti kekuatan, pengalaman relevan, dan potensi kesenjangan kandidat, memungkinkan perekrut untuk membuat keputusan penyaringan yang terinformasi dalam hitungan detik. Selain itu, sistem ini dirancang untuk skalabilitas horizontal, memastikannya dapat menangani beban kerja tinggi secara elastis sehingga rekrutmen massal atau lonjakan lamaran dapat diproses dengan kecepatan dan akurasi yang konsisten.

Dalam literatur terkini, terdapat dua pendekatan dominan dalam implementasi AI untuk penyaringan CV. Pendekatan pertama adalah Embedding Based Scoring. Pendekatan ini mengubah teks pada CV menjadi representasi numerik (embedding) dalam ruang vektor. Nilai skor kemudian dihitung melalui ukuran similarity, misalnya cosine similarity, dengan vektor kriteria acuan, seperti job description atau rubrik kompetensi. Keunggulannya adalah stabilitas deterministik dan efisiensi komputasi, tetapi kelemahannya terletak pada keterbatasan interpretasi serta fleksibilitas dalam menilai aspek multidimensional.

Pendekatan kedua adalah Prompt-Based Scoring atau terkadang disebut juga sebagai LLM-as-a-Judge (Li et al., 2025). Pada pendekatan ini, LLM diminta langsung memberikan skor berdasarkan instruksi atau rubrik tertentu yang tertuang dalam prompt (Gao, 2024; Yamauchi et al., 2025). Kelebihan metode ini adalah fleksibilitas untuk mengevaluasi dimensi yang kompleks, seperti keahlian teknis dan soft skill, sekaligus memberikan justifikasi penilaian. Namun,

kelemahannya adalah hasil yang rentan bervariasi tergantung desain prompt, temperatur model, dan bahkan versi model yang digunakan.

Meskipun kedua pendekatan ini menjanjikan, perbandingan reliabilitasnya masih minim ditelaah secara sistematis, khususnya dalam konteks penyaringan CV di Indonesia. Hal ini menimbulkan pertanyaan mendasar: metode skoring mana yang lebih reliabel, konsisten, dan dapat diandalkan untuk digunakan oleh praktisi HR dalam skala besar?

Oleh karena itu, penelitian ini bertujuan untuk Membandingkan reliabilitas embedding-based scoring dan prompt-based scoring dengan mengukur test–retest reliability, konsistensi internal, serta intraclass correlation coefficient (ICC). Studi ini diharapkan dapat memberikan rekomendasi praktis mengenai pemanfaatan metode skoring otomatis berbasis AI dalam praktik rekrutmen di Indonesia. Dengan fokus tersebut, penelitian ini diharapkan dapat memberikan kontribusi bagi literatur psikometri terapan serta praktik HR analytics, terutama pada ranah penggunaan AI untuk penyaringan CV yang andal, objektif, dan efisien.

# 3. Metode

## 3.1 Metode Penelitian

Pendekatan penelitian ini adalah kuantitatif dengan jenis penelitian eksperimental komparatif. Pendekatan kuantitatif adalah metode penelitian yang sistematis dan terstruktur untuk mengumpulkan, menganalisis, dan menginterpretasi data numerik dengan tujuan menguji teori, hipotesis, atau hubungan antar variabel. Fokus utama dari pendekatan ini adalah untuk mengukur fenomena secara objektif dan menggeneralisasi temuan dari sampel ke populasi yang lebih luas (Creswell & Creswell, 2018). Sementara itu, Eksperimental komparatif adalah jenis penelitian yang membandingkan dua atau lebih kelompok atau kondisi yang berbeda untuk menentukan efek dari suatu intervensi atau variabel independen (Cohen, Manion, & Morrison, 2018). Pada penelitian ini, tidak dilakukan adanya tindakan atau manipulasi terhadap variabel secara langsung, tetapi cukup meninjau dan membandingkan dua data secara alami.

Tujuan penelitian ini adalah untuk membandingkan reliabilitas dua metode penilaian otomatis pada aplikasi penyaringan CV berbasis kecerdasan buatan (*artificial intelligence*), yaitu metode embedding-based scoring dan metode prompt-based scoring (LLM-as-a-Judge). Kondisi pertama adalah skor yang dihasilkan oleh model embedding-based scoring (EBS), dan kondisi kedua adalah skor dari model prompt-based scoring (PBS) atau yang dikenal sebagai LLM-as-a-Judge.

## 3.2 Populasi, Sampel dan Teknik pengambilan Data

Populasi dalam penelitian ini adalah seluruh *Curriculum Vitae* (CV) yang diajukan oleh pelamar kerja untuk posisi profesional di Indonesia. Penelitian ini melibatkan sebanyak 200 CV yang diambil dari database perusahaan mitra Mimo.id dengan *consent* terhadap mitra. Guna menjaga privasi dan etika yang ada, dilakukan penyuntingan terhadap nama, alamat maupun kontak CV untuk menjaga anonimitasi dan etika penelitian. Seluruh nama partisipan diubah ke dalam kode sehingga tidak dapat dilakukan *tracking* terhadap identitas partisipan. CV yang telah *clean* tersebutlah yang kemudian dimasukkan ke dalam sistem kalkulasi Mimo.Id untuk diskoring secara *embedding-based scoring* dan Prompt-based scoring.

Teknik pengambilan data yang digunakan dalam penelitian ini adalah *sample random sampling* atau dapat diterjemahkan secara literal menjadi teknik pemilihan sampel acak sederhana. *Simple random sampling* adalah salah satu teknik pengambilan sampel probabilitas pengambilan sampel, di mana setiap individu atau dalam konteks ini CV dalam suatu populasi memiliki peluang yang sama untuk dipilih sebagai bagian dari sampel (Lohr, 2021). Proses melakukan *simple random sampling* ditentukan oleh sistem untuk melakukan pengacakan.

## 3.3 Instrumen Penelitian

Embedding-Based Scoring. Teks dari setiap CV dikonversi menjadi vektor embedding menggunakan model text embedding yang telah dilatih sebelumnya. Skor dihitung berdasarkan cosine similarity antara embedding CV dengan embedding kriteria acuan, yaitu job description dan rubrik kompetensi.

Prompt-Based Scoring, Setiap CV dinilai dengan menggunakan Large Language Model (LLM) melalui prompt terstandarisasi yang memuat instruksi rubrik penilaian. Model diminta menghasilkan skor numerik dalam rentang 0–100. Untuk tujuan uji reliabilitas test–retest, proses skoring diulang dua kali pada dataset yang sama dengan kondisi input identik.

## 3. 4 Teknik Analisis

Reliabilitas test–retest dianalisis menggunakan Intraclass Correlation Coefficient (ICC) dengan model two-way random effects, absolute agreement, single measures [ICC(2,1)]. Pendekatan ini digunakan untuk mengevaluasi derajat kesepakatan absolut antar pengukuran berulang terhadap unit yang sama (dalam hal ini, CV kandidat). ICC dipilih karena mampu menangkap baik konsistensi relatif maupun presisi absolut antar sesi penilaian, sehingga memberikan estimasi reliabilitas yang lebih komprehensif dibandingkan korelasi sederhana. Interpretasi nilai ICC mengacu pada pedoman Koo dan Li (2016): nilai < 0.50 dikategorikan poor reliability, 0.50–0.75 moderate, 0.75–0.90 good, dan > 0.90 excellent reliability.

Perhitungan ICC dilakukan menggunakan modul Reliability Analysis pada perangkat lunak Jamovi (versi terbaru), yang mengimplementasikan model two-way random, absolute agreement sesuai standar Shrout & Fleiss (1979). Hasil analisis dilaporkan dalam bentuk nilai ICC, variansi antar subjek, variansi antar rater, serta variansi residual untuk memberikan gambaran proporsi sumber kesalahan pengukuran.

Konsistensi urutan kandidat antar dua kali pengukuran dievaluasi menggunakan Spearman's rank-order correlation coefficient (ρ). Koefisien ini menilai hubungan monotonik antara dua set peringkat dan mengukur stabilitas urutan kandidat (rank-order stability) dalam distribusi skor. Analisis ini relevan dalam konteks seleksi berbasis ranking, karena menunjukkan apakah sistem penilaian menghasilkan susunan kandidat yang konsisten di antara sesi pengulangan.

Seluruh analisis korelasi dilakukan melalui modul Correlation Matrix pada Jamovi, dengan metode nonparametrik (Spearman's rho). Hasil dilaporkan dalam bentuk matriks rangkuman korelasi 30×30 antar pengulangan, beserta rentang nilai minimum–maksimum untuk menggambarkan konsistensi temporal model secara agregat.

# 4. Hasil

Bagian ini menyajikan hasil analisis reliabilitas sistem penyaringan CV berbasis kecerdasan buatan yang dikembangkan oleh Mimo.id. Analisis dilakukan pada dua tingkat evaluasi, yaitu score-level reliability dan rank-order reliability, untuk membandingkan kestabilan temporal antara dua pendekatan utama: Embedding-Based Scoring (EBS) dan Prompt-Based Scoring (PBS). Pendekatan EBS memanfaatkan representasi vektor semantik (embedding) untuk menghitung kesesuaian antara konten CV dan deskripsi pekerjaan, sedangkan PBS menggunakan model bahasa besar (Large Language Model, LLM) untuk memberikan skor berdasarkan instruksi terstruktur (prompt-based evaluation). Kedua pendekatan tersebut diuji melalui 30 kali pengulangan penilaian terhadap 200 CV, dengan tujuan mengestimasi konsistensi skor dan urutan kandidat dari waktu ke waktu (test–retest reliability).

Analisis dilakukan secara berurutan, dimulai dari penyajian statistik deskriptif yang menggambarkan kecenderungan sentral dan distribusi skor pada setiap pendekatan, termasuk nilai rata-rata, median, deviasi standar, skewness, kurtosis, serta uji normalitas menggunakan Shapiro–Wilk. Hasil uji deskriptif

**Tabel 1.** Uji Deskriptif

| Metode | Mean | Median | SD | Min | Max | Skew | Kurt | Shapiro–Wilk (*p*) |
|---|---|---|---|---|---|---|---|---|
| **EBS** | 0.454 – 0.456 | 0.454 – 0.456 | 0.342 – 0.343 | 0.28 | 0.64 | −0.74 – (−0.72) | 3.56 – 3.64 | < .001 |
| **PBS** | 60.50 – 61.80 | 62.00 – 64.00 | 0.340 – 0.350 | 0.00 | 91.00 | −1.92 – (−1.56) | 3.88 – 5.37 | < .001 |

Analisis deskriptif menunjukkan bahwa sistem penilaian berbasis embedding (EBS) menghasilkan tendensi sentral yang hampir tidak berubah di seluruh 30 pengulangan (Mean = 0,454–0,456), yang mencerminkan determinisme komputasional yang tinggi dan dispersi yang rendah. Sebaliknya, metode penilaian berbasis prompt (PBS) menunjukkan variabilitas yang sedikit lebih tinggi (rata-rata = 60,5–61,8) serta kemiringan negatif yang lebih kuat (−1,9 ≤ kemiringan ≤ −1,5), yang mengindikasikan kecenderungan konsentrasi skor yang lebih tinggi. Kedua metode tersebut menyimpang dari distribusi normalitas (Shapiro–Wilk $p < 0,001$), sehingga mendukung penerapan estimator reliabilitas non-parametrik pada tahap selanjutnya.

## 5.1 Score Level Reliability

Tahap berikutnya adalah analisis score-level reliability, yang dilakukan dengan menggunakan korelasi Spearman's rho untuk menilai stabilitas temporal antar sesi pengukuran, disertai perhitungan Standard Error of Measurement (SEM) untuk memperkirakan tingkat presisi skor, dan Intraclass Correlation Coefficient (ICC) untuk menilai kesepakatan absolut antar pengulangan. Reliabilitas pada tingkat skor (*score-level reliability*) mengacu pada derajat konsistensi hasil pengukuran numerik antar dua atau lebih sesi penilaian yang dilakukan dalam kondisi serupa

(Furr, 2021). Dalam konteks sistem penilaian otomatis berbasis kecerdasan buatan, reliabilitas skor mencerminkan stabilitas temporal model dalam menghasilkan nilai kuantitatif yang merepresentasikan kualitas kandidat. Semakin tinggi konsistensi antar sesi penilaian terhadap kandidat yang sama, semakin kecil kontribusi *random error* terhadap hasil skoring, dan semakin tinggi reliabilitas sistem tersebut. Untuk menilai reliabilitas pada tingkat skor, penelitian ini menggunakan dua pendekatan utama, yakni test–retest reliability dan Intraclass Correlation Coefficient.

*Test-Retest Reliability* (Spearman's ρ) digunakan untuk menilai stabilitas temporal berdasarkan hubungan monotonik antar dua sesi penilaian. Korelasi Spearman's rho dipilih karena data skor bersifat non-normal, sehingga analisis nonparametrik lebih sesuai untuk mengukur konsistensi ordinal antara skor test dan retest. Dengan memadukan kedua indeks tersebut, reliabilitas skor dapat dipahami secara komprehensif, baik dari sisi kestabilan relatif, hal ini mengukur apakah kandidat yang tinggi tetap tinggi maupun dari sisi kesetaraan absolut, hal ini mengukur apakah nilai yang dihasilkan tetap sama). Nilai reliabilitas diklasifikasikan ke dalam empat kategori utama, yaitu poor (< 0.50), moderate (0.50–0.75), good (0.75–0.90), dan excellent (> 0.90).

**Tabel 2.** Interpretasi Test–Retest Reliability

| Metode | Rentang Spearman's ρ | Standard Error of Measurement (SEM) | Interpretasi |
|---|---|---|---|
| **EBS** | 1.000 – 1.000 | 0.000 – 0.000 | Perfect temporal stability (*deterministic*). Tidak ada error pengukuran |
| **PBS** | 0.76 – 0.85 | 0.130 – 0.170 | High consistency, moderate sessional variability. Error pengukuran kecil–sedang; presisi tinggi tapi tidak sempurna |

Uji test–retest reliability menggunakan korelasi Spearman's ρ menunjukkan bahwa pendekatan *Embedding-Based Scoring* (EBS) memiliki konsistensi temporal yang sempurna di seluruh 30 kali pengulangan ($\rho = 1.000$, $p < .001$). Hasil ini menegaskan bahwa model embedding bersifat sepenuhnya deterministik, setiap kali proses penilaian diulang dengan input dan konfigurasi identik, sistem menghasilkan skor yang sama tanpa deviasi. Secara psikometrik, nilai korelasi sempurna ini merefleksikan reliabilitas temporal absolut, menandakan tidak

adanya error acak maupun variasi algoritmik yang dapat memengaruhi hasil. Dengan demikian, EBS dikategorikan memiliki stabilitas intramodal tertinggi, ideal untuk proses penyaringan kandidat berskala besar yang memerlukan konsistensi komputasional dan reproducibility penuh.

Sebaliknya, pendekatan *Prompt-Based Scoring* (PBS) memperlihatkan korelasi test–retest yang tinggi namun tidak sempurna, dengan rentang ρ antara 0.76–0.85 ($p < .001$). Nilai ini menunjukkan bahwa sistem LLM-as-a-Judge mempertahankan reliabilitas pada tingkat "high reliability" menurut kriteria Cohen (1988) dan Furr (2021), walaupun masih terdapat fluktuasi antar sesi. Variabilitas ini berasal dari sifat stokastik model bahasa besar yang dipengaruhi oleh parameter temperature, konteks instruksi (prompt phrasing), serta dinamika internal token sampling. Dengan demikian, PBS memiliki reliabilitas yang baik untuk menilai kualitas teks secara semantik dan kontekstual, meski stabilitas temporalnya lebih rendah dibanding EBS. Pendekatan ini lebih cocok untuk situasi evaluasi yang memerlukan fleksibilitas interpretatif, bukan determinisme penuh.

*Intraclass Correlation Coefficient* (ICC) digunakan untuk menilai agreement absolut antar pengulangan penilaian. ICC tidak hanya mempertimbangkan keterkaitan relatif antar skor, tetapi juga kesetaraan absolut antar nilai, sehingga memberikan estimasi reliabilitas yang lebih ketat. Model yang digunakan adalah *two-way mixed*, *absolute agreement* sesuai rekomendasi Koo & Li (2016), yang relevan ketika setiap pengulangan diperlakukan sebagai "rater" dengan kondisi input identik.

Selanjutnya, rank-order reliability dievaluasi menggunakan koefisien korelasi Kendall's tau-b guna menilai konsistensi urutan kandidat antar pengulangan.

**Tabel 3.** Interpretasi Test–Retest Reliability

| Metode | ICC | Subject Variance | Rater Variance | Residual Variance | Klasifikasi Reliabilitas (Koo & Li, 2016) | Interpretasi Singkat |
|---|---|---|---|---|---|---|
| **EBS** | 1.000 | 0.00238 | $2.07\times10^{-8}$ | $1.76\times10^{-9}$ | Perfect (Deterministic) | Tidak ada error pengukuran; skor identik di setiap replikasi; stabilitas absolut penuh. |
| **PBS** | 0.906 | 266.000 | −0.010 | 27.700 | Excellent Reliability | Reliabilitas sangat tinggi; terdapat sedikit variasi antar-run akibat efek stokastik LLM (prompt phrasing, temperature). |

Analisis Intraclass Correlation Coefficient (ICC) memperkuat temuan sebelumnya. Pendekatan Embedding-Based Scoring (EBS) menunjukkan reliabilitas sempurna dengan ICC(3,1) = 1.000, baik pada dimensi consistency maupun absolute agreement. Variansi antar-subjek (0.00238) jauh lebih besar dibandingkan variansi antar-rater ($2.07\times10^{-8}$) maupun variansi residual ($1.76\times10^{-9}$), yang menandakan bahwa hampir seluruh variasi skor berasal dari perbedaan nyata antar kandidat, bukan dari kesalahan pengukuran. Secara psikometrik, EBS menunjukkan determinisme komputasional penuh, sehingga cocok untuk skenario evaluasi massal yang menuntut stabilitas hasil lintas waktu dan lintas sistem.

Sementara itu, pendekatan Prompt-Based Scoring (PBS) memperoleh ICC(3,1) = 0.906, termasuk kategori “excellent reliability” (Koo & Li, 2016). Variansi antar-subjek (266.0) mendominasi total variasi, sementara variansi residual (27.7) menunjukkan adanya fluktuasi kecil akibat efek stokastik model bahasa besar. Nilai rater variance yang negatif (−0.010) menandakan perbedaan antar pengulangan sangat kecil, sehingga meskipun PBS tidak se-deterministik EBS, konsistensinya tetap tinggi dan error pengukurannya masih dalam batas psikometrik yang dapat diterima. Dengan demikian, PBS dapat dikatakan andalan untuk evaluasi kontekstual berbasis teks profesional, sedangkan EBS lebih unggul dalam menjaga reliabilitas numerik absolut dan reproduksibilitas lintas pengulangan.

Selanjutnya, dilakukan analisis diagnostik tambahan seperti Bland–Altman plots dan perhitungan rank displacement untuk meninjau potensi bias sistematis dan pergeseran ordinal antar sesi. Seluruh hasil disajikan dalam bentuk tabel

ringkasan dan narasi interpretatif yang mengikuti konvensi pelaporan psikometrik (Furr, 2021; Koo & Li, 2016).

## 5.2 Rank-Order Reliability

Analisis Rank-Order Reliability dilakukan untuk menilai sejauh mana konsistensi urutan kandidat dipertahankan antar sesi pengulangan model. Karena data yang dianalisis berbentuk peringkat ordinal (rank 1–200) dan tidak berdistribusi normal, reliabilitas dihitung menggunakan Kendall's tau-b (τ), koefisien korelasi nonparametrik yang mengukur tingkat kesesuaian urutan antara dua set ranking. Berbeda dengan korelasi Pearson yang menilai hubungan linear pada data interval, Kendall's tau menghitung proporsi pasangan observasi yang concordant (selaras) dan discordant (berlawanan) antar dua sesi penilaian, dengan penyesuaian terhadap ties atau kandidat yang memiliki skor identik. Nilai τ berkisar antara −1 (urutan sepenuhnya berlawanan) hingga +1 (urutan sepenuhnya sama), sehingga semakin tinggi nilai τ, semakin stabil urutan kandidat antar sesi.

**Tabel 4.** Interpretasi Test–Retest Reliability

| Metode | Rentang Kendall's τ (min–max) | Klasifikasi Reliabilitas | Interpretasi |
|---|---|---|---|
| **EBS** | 0.998 – 1.000 | Perfect Rank Stability | Struktur peringkat kandidat identik di seluruh sesi; sistem deterministik tanpa error pengukuran. |
| **PBS** | 0.57 – 0.69 | Moderate to High Stability | Stabilitas ordinal cukup tinggi, meskipun peringkat kandidat dapat berubah akibat variasi stokastik dalam proses inferensi LLM. |

Analisis ini diterapkan pada 30 kali pengulangan untuk masing-masing pendekatan penilaian otomatis, yakni Embedding-Based Scoring (EBS) dan Prompt-Based Scoring (PBS), dengan menghitung korelasi τ secara pairwise antar sesi (30×30 kombinasi) menggunakan prosedur correlation matrix nonparametrik

(method = "kendall"). Hasil kemudian diringkas sebagai rentang minimum–maksimum untuk menggambarkan stabilitas ordinal temporal masing-masing metode. Berdasarkan pedoman interpretasi reliabilitas ordinal (Furr, 2021; Cohen, 1988), nilai $\tau \geq 0.90$ dikategorikan sebagai reliabilitas sangat tinggi, 0.70–0.89 tinggi, 0.50–0.69 sedang, dan <0.50 rendah.

Dalam penelitian ini, hasil Kendall's tau menunjukkan perbedaan yang mencolok antara dua pendekatan. EBS menghasilkan nilai $\tau$ antara 0.998–1.000, menandakan stabilitas urutan kandidat yang sempurna dan deterministik, sementara PBS memiliki nilai $\tau$ antara 0.57–0.69, yang termasuk kategori stabilitas sedang hingga tinggi. Hal ini mengindikasikan bahwa sistem embedding menghasilkan pemeringkatan kandidat yang konstan di seluruh pengulangan tanpa pergeseran posisi, sedangkan sistem berbasis prompt masih menunjukkan fluktuasi ordinal kecil antar sesi akibat sifat stokastik model bahasa besar (LLM) dan sensitivitas terhadap desain prompt.

# 6 Diskusi

Hasil penelitian ini menunjukkan bahwa pendekatan embedding-based scoring (EBS) secara keseluruhan menunjukkan reliabilitas yang lebih unggul dibandingkan prompt-based scoring (PBS) dalam konteks penyaringan CV berbasis AI, khususnya pada platform Mimo.id. Pengukuran test-retest reliability melalui korelasi Spearman's ρ menunjukkan nilai sempurna untuk EBS (ρ = 1.000), mencerminkan stabilitas temporal absolut tanpa deviasi antar pengulangan, sementara PBS mencapai rentang 0.76–0.85, menandakan konsistensi tinggi meskipun dengan variabilitas moderat. Analisis Intraclass Correlation Coefficient (ICC) lebih lanjut memperkuat temuan ini, dengan EBS mencapai ICC = 1.000 (kategori perfect reliability), di mana variansi residual mendekati nol, menunjukkan tidak adanya error pengukuran pengulangan. Sebaliknya, PBS memperoleh ICC = 0.906 (kategori excellent reliability), dengan variansi residual yang menandakan fluktuasi kecil akibat efek stokastik seperti temperatur model dan desain prompt. Pada rank-order reliability melalui Kendall's τ, EBS menunjukkan stabilitas sempurna (τ = 0.998–1.000), memastikan urutan kandidat identik antar sesi, sedangkan PBS berada pada rentang 0.57–0.69 (moderate to high stability), yang meskipun masih dapat diterima untuk evaluasi ordinal tetapi rentan terhadap pergeseran posisi ranking.

Keunggulan EBS ini dapat dijelaskan melalui sifat deterministiknya yang inheren, di mana konversi teks CV menjadi embedding vektor memungkinkan perhitungan kesamaan semantik (seperti cosine similarity) yang stabil dan bebas dari variasi prompt atau temperatur model. Pendekatan ini menangkap kedalaman semantik latens dengan lebih baik, sehingga kurang rentan terhadap saturasi skor tinggi atau bias permukaan yang sering terjadi pada metrik leksikal tradisional. Sebaliknya, PBS, meskipun fleksibel dalam mengevaluasi aspek kompleks melalui instruksi prompt, menunjukkan kerentanan terhadap inkonsistensi, seperti yang terlihat dari variasi skor akibat desain prompt atau pembaruan model LLM. Temuan ini sejalan dengan literatur terkini yang menyoroti bahwa metrik berbasis embedding lebih forgiving terhadap perbedaan wording sambil memberikan korelasi yang lebih tinggi dengan penilaian manusia, terutama dalam tugas evaluasi teks seperti ringkasan atau dialog.

Studi oleh Kim et al. (2023) lebih lanjut mendukung argumen ini dengan mengeksplorasi strategi prompting untuk metrik berbasis LLM, di mana hasil menunjukkan bahwa meskipun prompt berbasis pedoman manusia (human guideline) meningkatkan korelasi Kendall's Tau hingga 0.45 pada model seperti Orca-13B, performa tetap tidak konsisten akibat bias demonstrasi dan agregasi skor, yang menyebabkan penurunan reliabilitas pada model kecil. Demikian pula, Hua et al. (2025) menemukan inkonsistensi signifikan antara rater LLM dan manusia dalam penilaian esai, dengan quadratic weighted kappa (QWK) tertinggi hanya 0.90 untuk GPT-4o, tetapi menurun tajam pada model lain seperti Gemini 1.5 (QWK = 0.46), yang menegaskan variabilitas prompt-based yang dapat mengurangi keandalan dalam aplikasi praktis seperti rekrutmen.

Dalam konteks Mimo.id, adopsi EBS sebagai metode utama untuk AI Resume Screening berpotensi meningkatkan efisiensi operasional HR dengan mengurangi waktu penyaringan hingga 60% sambil meminimalkan risiko bias subyektif. Sistem ini tidak hanya menghasilkan peringkat dinamis yang lebih akurat ("Very Fit" vs. "Fairly Fit"), tetapi juga mendukung skalabilitas untuk lonjakan lamaran, yang krusial di pasar tenaga kerja Indonesia. Namun, penelitian ini masih memiliki keterbatasan terkait dengan kualitas embedding model, yang bisa dipengaruhi oleh data pelatihan berbahasa Indonesia yang kurang representatif. Penelitian selanjutnya perlu mengeksplorasi mengenai aspek validitas maupun *fairness* dari EBS untuk memperkuat properti psikometri yang terkandung di dalamnya.

Secara keseluruhan, temuan ini memperkuat rekomendasi untuk memprioritaskan EBS dalam pengembangan AI rekrutmen, dengan potensi integrasi hybrid untuk menggabungkan fleksibilitas PBS di tahap verifikasi manual (Lacroux et al., 2022). Penelitian mendatang disarankan untuk menguji EBS pada dataset multilingual yang lebih besar, guna memperluas aplikabilitasnya di ekosistem HR analytics di Indonesia.

# 7 Kesimpulan

Penelitian ini menunjukkan bahwa Embedding-Based Scoring (EBS) memiliki reliabilitas yang lebih baik daripada Prompt-Based Scoring (PBS) guna mengatasi masalah fenomena banjir CV yang dihadapi oleh para HR dewasa ini. EBS memiliki kemampuan komputasi yang lebih stabil dan deterministik, dibanding PBS yang meskipun lebih *adjustable* memiliki kelemahan dalam aspek reliabilitas scoring dan ranking partisipan.

Dalam konteks platform Mimo.id, adopsi model hybrid yang mengintegrasikan EBS sebagai fondasi utama dengan PBS untuk verifikasi lanjutan berpotensi merevolusi penyaringan CV. Di mana scoring komputasi dengan EBS memungkinkan Mimo.id mendapatkan hasil yang stabil, tetapi tetap mampu mengevaluasi kualitas konten dari CV secara adaptif melalui pendekatan PBS. Hasil penelitian ini diharapkan mampu menjadi rekomendasi praktis untuk praktisi HR analytics guna menggunakan alat assessment yang didasarkan pada kerangka psikometrik yang kokoh sebelum digunakan di lapangan sebagaimana prinsip *evidence based practice* (EBP).

# 8 Referensi


Chafiq, N., El Ghazouani, M., & El Gounidi, R. (2025). From Manual Review to AI Automation: An NLP-Powered System for Efficient CV Processing in Academic Admissions. *LatIA,* (3), 315. https://doi.org/10.62486/latia2025315

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research Methods in Education* (8th ed.). Routledge.

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (5th ed.). SAGE Publications.

Fisher, S., Motowidlo, S. J., & Borman, W. C. (2024). AI-based tools in selection: Considering the impact on reliability and validity of assessment. *Human Resource Management Review, 34*(1), 100925. https://doi.org/10.1016/j.hrmr.2024.100925

Furr, R. M. (2021). *Psychometrics: An introduction* (4th ed.). SAGE Publications.

Gero, Z., Singh, C., Xie, Y., Zhang, S., Subramanian, P., Vozila, P., ... & Poon, H. (2024). Attribute structuring improves LLM-based evaluation of clinical text summaries. *ArXiv Preprint.* https://doi.org/10.48550/arXiv.2403.01002

Hua, H., Jiao, H., & Song, D. (2025). Comparison of scoring rationales between large language models and human raters. *ArXiv Preprint.* https://doi.org/10.48550/arXiv.2509.23412

Kim, J., Park, S., Jeong, K., Lee, S., Han, S. H., Lee, J., & Kang, P. (2023). Which is better? exploring prompting strategy for llm-based metrics. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2311.03754

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lacroux, A., Bar-Hen, A., & Leclercq-Vandelannoitte, A. (2022). Should I trust artificial intelligence to recruit? A first typology of distrust. *Frontiers in Psychology*, 13, 868271. https://doi.org/10.3389/fpsyg.2022.868271

Li, Q., Dou, S., Shao, K., Chen, C., & Hu, H. (2025). *Evaluating Scoring Bias in LLM-as-a-Judge*. arXiv preprint arXiv:2506.22316.

Lohr, S. L. (2021). *Sampling: Design and Analysis* (3rd ed.). CRC Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Yamauchi, T., Huang, L., & Wu, C. (2025). *An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability*. arXiv preprint arXiv:2506.13639. https://arxiv.org/abs/2506.13639

Zheng, L., et al. (2025). LLM evaluation metrics: A comprehensive guide for large language models. *Weights & Biases Reports*. https://wandb.ai/onlineinference/genai-research/reports/LLM-evaluation-metrics-A-comprehensive-guide-for-large-language-models--VmlldzoxMjU5ODA4NA