EVALPRO: A Comprehensive AI-Based Platform for Simulated Recruitment and Skill Assessment.

An intelligent system designed to automate end-to-end mock hiring, enabling realistic candidate evaluation and performance analysis

Sumaira A. Shaikh[1], Pratima Patil[1], Danish Bhat[1], Prathamesh Bharsakale*[1], and Ayush Bhagwat[1]

Department of Information Technology, Trinity Academy of Engineering, Pune, India
{sumairashaikh.tae,pratimapatil.tae}@kjei.edu.in
{danishbhat052,prathameshpb2004,bhagwatayush264}@gmail.com
* Corresponding author: prathameshpb2004@gmail.com

Abstract. EVALPRO is a platform that helps with the hiring process. It does this by simulating the process of finding and hiring someone. This includes testing how smart someone is, how good they are at their job and how well they do in interviews with people from the company and with the human resources department. The people looking for a job and the people doing the hiring both get feedback from EVALPRO. EVAL- PRO is also very secure. It can run code in an environment and tell the person what the code does. The system can look at kinds of information including text, audio and video. This helps the people doing the hiring get a sense of who the person is and if they are a good fit for the job. EVALPRO is about making the hiring process easier and more effective, for everyone involved. To enhance the reliability, safety, and consistency of LLM-driven components, EVALPRO incorporates runtime guardrails, including constrained decoding strategies, template-first generation, functional evaluation heuristics, and sandboxed execution with inverse overlay analysis. This paper presents the overall system architecture, core algorithms, runtime safety mechanisms, and an evaluation framework. Pilot study results demonstrate the feasibility of the proposed approach in realistic recruitment scenarios and highlight directions for future improvements in robustness, fairness, and scalability.

Keywords: AI-Powered Candidate Evaluation · Natural Language Processing · Adaptive Testing · Large Language Models · Sandbox Execution · Multimodal Assessment

M. Kaushik et al. (eds.), *Proceedings of the International Conference on Responsible, Risk-aware, and Regulated AI (RRRAI 2026)*, Advances in Intelligent Systems Research 210,
https://doi.org/10.2991/978-94-6239-723-1_3

1 Introduction

Traditional hiring processes are often slow, inconsistent, and vulnerable to hu- man bias. Although in-person interviews are considered the best way to evaluate candidates, they have limited scalability, high costs, and rely on subjective judgment. The rise of online assessments and video interviews partially addresses these issues by allowing remote and flexible evaluations. However, these methods also bring new challenges related to fairness, reliability, and measurement accuracy. Additionally, the growing availability of advanced AI tools raises concerns about candidates misrepresenting their true skills by bypassing or manipulating standard assessment methods. [1][3]

Recruitment preparation usually involves multiple skill areas, such as cognitive ability, technical skills, and communication skills. In reality, these areas are assessed using various tools that are disconnected and lack coherence. Candidates, especially recent graduates and early-career professionals, often do not have access to platforms that accurately simulate the full interview process. Meanwhile, recruiters struggle to evaluate large groups of applicants efficiently while maintaining consistency and objectivity. These challenges highlight the need for a scalable, practical, and data-driven solution that helps both candidates and recruiters.

EVALPRO fills this gap by combining interview practice and assessment in an AI-powered mock recruitment environment. The platform creates questions tailored to specific roles and resumes, assesses both coding and non-coding responses using a mix of clear rules and learned patterns, and offers understandable feedback enhanced with long-term trend analysis. In its standard setting, EVAL- PRO aims to support candidate development and provide insights for recruiters instead of making final hiring decisions, positioning the system as a tool for decision support and skill development.

A key aspect of this work is ensuring stability while integrating large language models (LLMs) and program execution into the assessment process. We use a generate-analyze-execute approach and implement runtime safeguards to reduce hallucinations, unsafe outputs, and unpredictable behavior. These protections allow for adaptive and personalized assessments while ensuring reliability, reproducibility, and safety in real-world recruitment settings.

2 Related Work

EVALPRO lies at the intersection of multiple research streams that collectively inform the design of automated recruitment, adaptive assessment, and AI-assisted evaluation systems. This section reviews the most relevant areas and situates EVALPRO within existing literature.

2.1 Adaptive Testing

Computerized Adaptive Testing (CAT) has long been used to tailor assessment difficulty based on an estimate of a candidate's ability level. Prior studies show that adaptive testing improves measurement precision, reduces test length, and mitigates candidate fatigue compared to fixed-form assessments. Recent work has also explored adaptive mechanisms for addressing cold-start scenarios in skill evaluation. Building on these foundations, EVALPRO adopts a lightweight, online difficulty controller that updates skill-specific difficulty levels after each response. This pragmatic approach enables personalization while remaining suit- able for real-time interview simulations. [1]

2.2 Resume Parsing and Question Generation

Natural language processing techniques for resume parsing have enabled the extraction of structured candidate attributes such as skills, experience, and education, which are widely used in job matching and candidate profiling systems. More recently, large language models have been applied to generate personalized interview questions conditioned on resumes and job roles. However, unconstrained generation often leads to hallucinations or irrelevant questions. To address this, recent research emphasizes prompt engineering, constrained decoding, and template-based generation. EVALPRO integrates these advances through a template-first, resume-aware question generation pipeline that prioritizes relevance, consistency, and controllability. [2]

2.3 Automated Code Evaluation

Automated code evaluation systems, including online judges and secure sandbox environments, are commonly used to assess programming correctness, efficiency, and adherence to constraints. While effective for correctness checking, these systems often provide limited insight into how or why a solution behaves as it does. EVALPRO extends traditional sandboxed execution by introducing effect summaries and inverse overlays, enabling explainability and post-execution analysis. This design supports both candidate learning and recruiter auditability while maintaining strong security isolation. [5]

2.4 Multimodal Interview Analysis

Recent research has demonstrated that combining textual, visual, and acoustic signals can improve the fidelity of automated interview analysis compared to unimodal approaches. Such multimodal systems can capture verbal content, facial expressions, and vocal cues, offering a richer representation of candidate behavior. However, they also introduce challenges related to computational cost, interpretability, and privacy. EVALPRO incorporates multimodal signals as supplementary indicators rather than primary decision factors, emphasizing transparency and optional use to balance evaluation depth with ethical and practical considerations. [8][10]

2.5 Bias Mitigation in AI-Based Hiring

Bias in AI-driven recruitment systems has been extensively documented, prompting research into fairness-aware models, auditing frameworks, and bias mitigation techniques. Studies consistently show that explicit bias-aware design im- proves equity in candidate evaluation. EVALPRO aligns with this body of work by focusing on fair feedback generation and recruiter insight, rather than fully automated hiring decisions. This design choice reduces the risk of unchecked algorithmic bias while still leveraging AI for scalable assessment support. [3][4][7]

2.6 Runtime Guardrails for LLMs

As LLMs are increasingly integrated into production systems, recent systems research advocates the use of runtime guardrails such as constrained decoding, template-first generation, validation pipelines, and post-generation checks. These techniques improve reliability, reduce unsafe outputs, and increase reproducibility. EVALPRO adapts these principles within an assessment context, applying guardrails to both question generation and response evaluation to ensure outputs remain within predefined semantic and operational bounds while preserving adaptability. [9]

3 System Architecture

Figure 1 shows EVALPRO's high-level architecture. The system is modular and cloud-ready, with decoupled AI services, sandbox execution, and a secure persistence layer.

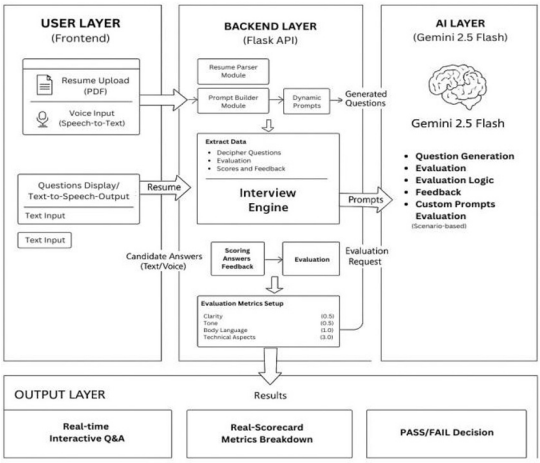

Fig. 1. High-level architecture of EVALPRO showing the four-layer design.

3.1 Layers and Components

EVALPRO is designed as a layered system to ensure modularity, scalability, and clear separation of concerns. Each layer encapsulates a specific set of responsibilities, enabling independent development and robust integration of user interaction, backend orchestration, and AI-driven assessment.

User Layer

The User Layer provides an interactive frontend through which candidates engage with the platform. Implemented using React with TypeScript and styled using Tailwind CSS, the interface includes dedicated views for Home, Authentication, Mock Tests, Interview Sessions, and Performance Dashboards.

For technical assessments, the system integrates the Monaco code editor to offer a professional, IDE-like coding experience. Audio and video inputs are captured using standard browser APIs, enabling seamless participation in multimodal interview simulations. The interface supports both text-based and voice- based interactions, improving accessibility and accommodating diverse user preferences and interview formats.

Backend Layer

The Backend Layer is implemented as a set of RESTful ser- vices using Flask and is responsible for session orchestration, state management, and request routing across system components. Secure authentication and authorization mechanisms are employed to protect user data and maintain session integrity throughout assessment workflows.

Key backend modules include:

- **Resume Parser Module:** Processes uploaded resumes and extracts structured candidate information, including skills, educational background, work experience, and certifications.
- **Prompt Builder Module:** Constructs dynamic, context-aware prompts by combining resume-derived attributes, target job roles, and adaptive difficulty parameters.
- **Interview Engine:** Manages the end-to-end assessment flow, controlling question sequencing, response collection, and transitions across aptitude, technical, and interview stages.

This layer acts as the central coordinator, ensuring consistent interaction between the frontend and AI-driven components.

AI Layer

The AI Layer encapsulates the core intelligence of the system and performs computationally intensive evaluation tasks. Its primary components include:

- **Dynamic Question Generator:** An LLM-augmented module that pro- duces role-specific and resume-aware questions using constrained decoding and template-guided generation to ensure relevance and consistency.
- **Scenario-Based Assessment Engine:** Generates realistic interview scenarios designed to assess problem-solving ability, analytical reasoning, and situational judgment.
- **Evaluator Engines:** A collection of evaluators, including a secure code sandbox for program execution, embedding-based semantic scoring for textual responses, and a speech-to-text (STT) with NLP pipeline for analyzing spoken answers.
- Scoring Logic: Aggregates multiple evaluation signals—such as keyword matching, semantic similarity, and test-based correctness—into structured, multi-dimensional performance scores. [8][9]
- **Feedback Generator:** Produces interpretable and actionable feedback, highlighting strengths, identifying skill gaps, and recommending targeted areas for improvement.

Together, these components enable adaptive, explainable, and multimodal assessment while maintaining operational safety.

Persistence and Reporting Layer

The Persistence and Reporting Layer manages long-term data storage and analytics. PostgreSQL is used for structured, relational data such as user profiles, assessment metadata, and scores, while object storage services handle large media artifacts including audio and video recordings. Reporting services aggregate evaluation results to generate composite scorecards, performance trends, and analytical summaries for both candidates and recruiters.

4 Runtime Guardrails and Safety

The integration of large language models (LLMs) into assessment pipelines introduces several risks, including hallucinated content, biased or unsafe outputs, and inconsistent formatting. In a high-stakes evaluation setting, such failures can undermine trust, fairness, and reproducibility. EVALPRO addresses these challenges through a layered runtime guardrail strategy designed to constrain generation, validate outputs, and prevent unsafe artifacts from entering live assessment sessions.

4.1 Template-First Generation

EVALPRO adopts a template-first generation strategy in which curated templates are maintained for each role and skill category. These templates define the structural and semantic boundaries of assessment questions, including expected difficulty, scope, and response format. During question generation, the pipeline selects an appropriate template and invokes the LLM only to populate constrained fields through controlled slot-filling. This approach significantly reduces variability and hallucination while preserving contextual adaptation based on candidate profiles.

4.2 Constrained Decoding and Post-Filtering

All LLM invocations in EVALPRO employ constrained decoding techniques and schema-guided output formats, such as JSON-based specifications. Generated outputs are subsequently subjected to a series of post-generation validation checks, including:

- Verification of keyword and concept overlap with resume-extracted skills
- Consistency checks between generated content and the intended difficulty level
- Safety screening to detect offensive, irrelevant, or out-of-scope material
- Structural and format validation against predefined schemas

Only outputs that satisfy all validation criteria are admitted into the assessment workflow, ensuring consistency and reliability across sessions.

4.3 Generate–Analyze–Execute Loop

To further enhance robustness, EVALPRO employs a generate–analyze–execute loop before deploying generated artifacts in live interactions. In this loop, candidate questions or tasks are first analyzed by a verification module that simulates expected responses or executes short code stubs in a secure environment when applicable. Artifacts that fail analysis are either discarded or flagged for human review. This additional verification stage acts as a final safeguard, preventing erroneous or unsafe content from reaching end users while maintaining system autonomy.

5 Core Methods

This section describes the core methodological components that enable controlled question generation, reliable evaluation, and adaptive personalization within EVALPRO.

5.1 Constrained Question Generation and Functional Evaluation Heuristics

Two-Stage Generation EVALPRO employs a two-stage question generation pipeline to ensure both relevance and reliability. In the first stage, an appropriate template is selected based on the target role, skill category, and desired difficulty level. In the second stage, a large language model is used to populate constrained template slots through controlled slot-filling. Generated questions are subsequently subjected to post-generation filtering and validation to enforce structural correctness, topical relevance, and safety constraints. This staged approach limits uncontrolled variability while preserving contextual adaptation.

Functional Evaluation Heuristics (FEH) To move beyond surface-level matching, EVALPRO introduces Functional Evaluation Heuristics (FEH), which prioritize functional correctness and behavioral validity over lexical similarity. For programming tasks, submitted code is executed within a secure sandbox against hidden unit tests, and evaluation metrics include test pass ratio, execution time, and static quality indicators such as code structure and rule violations. For non-code responses, FEH combines multiple signals, including embedding- based semantic similarity, presence of required technical or behavioral keywords, grammar and fluency scores, and sentiment analysis for behavioral interview questions. These heterogeneous signals are aggregated into a multi-dimensional score vector, enabling fine-grained assessment across different competency dimensions. [8][9]

5.2 Adaptive Difficulty Controller [1]

EVALPRO incorporates an adaptive difficulty controller that personalizes assessment complexity on a per-skill basis. Each skill is associated with a difficulty index Dt, which is updated after every attempt according to observed candidate performance: [1]

$$D_{t+1} = clip\,(D_t + \alpha \cdot (c_t - \theta),\ D_{\min},\ D_{\max}) \quad (1)$$

Here, c_t represents the observed correctness or confidence score for the current attempt, θ denotes the target success threshold, and α is a learning rate controlling the responsiveness of adaptation. The clipping operation bounds the difficulty within predefined limits $[D_{\min}, D_{\max}]$, ensuring stability and preventing extreme oscillations. This mechanism enables gradual personalization while maintaining assessment fairness and comparability.

6 Sandboxed Execution and Effect Summaries

Secure and interpretable execution of candidate submitted code is critical for reliable technical assessment. EVALPRO addresses this requirement through a tightly controlled sandboxing mechanism coupled with structured execution summaries that support transparency and auditability.

6.1 Execution Sandbox

All executable submissions are evaluated within isolated Docker containers configured with strict CPU and memory limits, disabled network access, and sandboxed filesystem overlays. Each container is instantiated ephemerally for a single execution session, ensuring that no state persists across runs. During execution, the system instruments the container to capture standard output and error streams, execution time, resource usage, and any files created or modified within the sandboxed environment. This design ensures secure, reproducible, and resource-bounded evaluation of untrusted code.

6.2 Effect Summaries and Inverse Overlays

Following execution, EVALPRO generates a structured effect summary that records key outcomes, including the number of tests passed, generated artifacts, raised exceptions, and detailed runtime statistics. In addition, the system maintains an inverse overlay that captures all filesystem changes relative to a clean base image. This inverse overlay enables deterministic rollback, post-hoc inspection, and audit trails for each execution. Together, effect summaries and inverse overlays provide a transparent and explainable view of program behavior, supporting both candidate feedback and system-level accountability.

7 Implementation

7.1 Prototype Stack

- **Frontend:** React + TypeScript + Tailwind + Monaco editor
- **Backend:** Flask + Gunicorn; Auth via Auth0
- **AI Services:** Gemini (or equivalent LLM), embedding models, Whisper for STT
- **Execution:** Docker-based isolated runtimes
- **Storage:** PostgreSQL + S3-compatible object storage

7.2 Data and Privacy

Explicit user consent is obtained before any audio or video capture begins, with clear disclosure of what is recorded and why. All recordings are encrypted in transit (TLS) and at rest (AES-256), and retention periods are governed by the platform's data policy. Users retain the ability to opt out of multimodal capture at any point without affecting core platform functionality; in such cases, the assessment falls back to text and code-based interaction only. [10]

Data retention periods are governed by the platform's data retention policy. Recordings and session data are automatically purged once the retention window expires, unless a legal or compliance hold is in place. Users can request early deletion of their data through a self-service mechanism in line with applicable data protection regulations such as GDPR and applicable local privacy laws[6].

Users retain the ability to opt out of multimodal capture (audio and/or video) at any point without affecting core platform functionality; assessments in such cases fall back to text and code-based interaction only. Where a candidate opts out, this is noted in the session record without penalising their evaluation. [10]

Personally identifiable information (PII) extracted during sessions — such as voice data or facial imagery — is not used for model training without separate, explicit consent. The platform does not share candidate data with third parties beyond what is necessary for core service delivery (e.g. cloud infrastructure providers), and all such third parties are bound by appropriate data processing agreements.

EVALPRO is evaluated along four primary dimensions: functional correctness, semantic alignment with human judgment, runtime safety, and overall user experience. The evaluation is designed to assess both the technical validity of system outputs and their perceived usefulness in realistic recruitment preparation scenarios.

8.1 Metrics

We employ the following quantitative and qualitative metrics:

- **Functional Accuracy:** The percentage of coding submissions that successfully pass hidden unit tests executed within the sandboxed environment.
- **Semantic Correlation:** Pearson correlation coefficient between system- generated semantic scores and human evaluator ratings for non-code responses.
- **Question Relevance:** Human-rated relevance of generated questions on a 1–5 Likert scale, measuring alignment with resume content and target roles.

- **FEH Error Rates:** False acceptance and false rejection rates of the Functional Evaluation Heuristics (FEH), measured against human adjudication.
- **Safety Rate:** The fraction of generated outputs flagged by automated safety and validation filters prior to deployment.

Together, these metrics capture correctness, alignment with human judgment, robustness, and user-facing quality.

8.2 Pilot Study

We conducted a pilot study involving 30 participants across diverse technical backgrounds. Participants interacted with the system through mock assessments and interview simulations, and their submissions were evaluated using the pro- posed pipeline. Human evaluators independently rated question relevance and response quality to provide ground-truth comparisons.

Table 1. Pilot evaluation summary.

Metric	Mean	Std Dev	Notes
Question relevance (1–5)	4.2	0.5	Human rating
Coding pass rate (%)	76	11	Hidden testcases
Semantic correlation (r)	0.70	0.09	vs. human scores
FEH FPR (%)	6.5	–	False positives
User satisfaction (1–5)	4.1	0.4	Survey

9 Discussion

The development and pilot deployment of EVALPRO yield several insights into the design of AI-assisted assessment systems.

- **Controlled generation improves reliability.** The combination of template-first generation and post-generation filtering substantially reduces hallucinations, unsafe content, and formatting inconsistencies in LLM-generated questions. By constraining the generation space before output and applying structured validation afterward, the pipeline produces more predictable and trustworthy assessment artifacts — a meaningful improvement over unconstrained prompting approach.

- Functional evaluation aligns better with human judgment. The Functional Evaluation Harness (FEH), which integrates embedding-based semantic similarity, keyword presence checks, and executable test-based validation, demonstrates stronger agreement with human evaluators than lexical or surface-level scoring alone. This suggests that combining multiple complementary signals — semantic, syntactic, and behavioral — yields a more robust proxy for expert assessment than any single method. [8][9]

- **Multimodal signals require cautious use.** While audio and visual cues can enrich the assessment picture, they are inherently noisy and may encode demographic or environmental bias. Treating multimodal signals as supplementary indicators rather than primary decision drivers helps mitigate fairness and interpretability concerns and ensures that no candidate is disadvantaged by factors unrelated to their technical ability — such as background noise, accent, or hardware quality.

- **Adaptive difficulty enhances engagement.** Skill-wise difficulty adjustment improves candidate engagement and reduces frustration during the assessment. This finding suggests that adaptive mechanisms are valuable even in relatively short, interview-style assessments — not just in lengthy standardized tests — and that responsiveness to candidate performance contributes meaningfully to a fairer and more informative evaluation experience.

10 Limitations and Future Work

Despite promising results, EVALPRO has several limitations. First, the system currently relies on cloud-hosted large language models, which introduces dependencies related to latency, cost, and data governance. Second, the pilot evaluation was conducted at a limited scale, restricting the statistical power and generalizability of the findings. Third, the speech-to-text (STT) pipeline is sensitive to background noise and accent variability, which can affect the reliability of spoken-response evaluation.

Future work will address these limitations through several directions:

- Conducting larger-scale user studies, including systematic fairness and bias audits across demographic groups
- Improving robustness against adversarial behavior, including detection of AI-assisted or scripted candidate responses
- Exploring on-device and lightweight generation techniques to enhance privacy and reduce reliance on cloud services
- Expanding domain-specific test banks and adding multilingual assessment support
- Integrating with applicant tracking systems (ATS) and investigating blockchain- based credential verification for enhanced trust and portability

These extensions aim to improve scalability, fairness, and real-world applicability.

11 Conclusion

EVALPRO presents a unified and practice-oriented platform for realistic inter- view preparation and assessment. By combining adaptive testing, constrained LLM-driven question generation, reproducible sandboxed execution, and multi- modal evaluation, the system supports both candidate skill development and recruiter insight. The integration of runtime guardrails and Functional Evaluation Heuristics enables a balance between personalization, safety, and explainability. Overall, EVALPRO demonstrates the feasibility of deploying robust AI-assisted assessment systems in recruitment contexts and provides a foundation for future research and development in this space.

Acknowledgements The authors thank Prof. S. A. Shaikh for guidance through- out this project, Prof. S. A. Joshi for coordination support, and Prof. P. R. Patil, Head of Department, for encouragement. We also acknowledge Trinity Academy of Engineering, Pune, and Savitribai Phule Pune University for support.

References

1. Ebrahim, S.S., Rajab, H.A.: The future of hr: The role of ai-powered recruitment in shaping the modern workforce. Open Access Library Journal **12**(01), 1–22 (2025)
2. Bevara, R.V.K., Mannuru, N.R., Karedla, S.P., Lund, B., Xiao, T., Pasem, H., Dronavalli, S.C., Rupeshkumar, S.: Resume2vec: Transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching. Elec- tronics **14**(4), 794 (2025)
3. Fabris, A., Baranowska, N., Dennis, M.J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F., Biega, A.J.: Fairness and bias in algorithmic hiring: A multidisciplinary survey. ACM Transactions on Intelligent Systems and Technology **16**(1), 1–47 (2025)
4. Ip, E.: Fair ai in hiring: Experimental evidence on how biased hiring algorithms and different debiasing methods affect the quality and diversity of applicants. Be- havioral Science Policy, 1–27 (2025)
5. Kim, H., Park, S.: Proctoring with ai: Real-time detection of cheating behavior in online assessments. IEEE Transactions on Learning Technologies **17**(1), 45–57 (2024)
6. Fernandez, R., Lopez, E.: Ethical risks in ai-powered talent acquisition systems. AI Ethics **5**(1), 112–129 (2024)
7. Rigotti, C., Fosch-Villaronga, E.: Fairness, ai recruitment. Computer Law Security Review **53**, 105966 (2024)
8. Roy, N., Banerjee, T.: Using ai to measure soft skills: An empirical evaluation. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2024)
9. Suresh, A., Raman, K.: Llm-powered resume screening: Accuracy, risks, and improvements. Information Processing Management **61**(2), 103483 (2024)
10. George, T., Philip, A.: Ai-based personality assessment using multimodal behav- ioral signals. IEEE Access **12**, 76523–76535 (2024)