Intervueai: A Multimodal Retrieval-Augmented Generation Framework For AI-Powered Intelligent Recruitment Automation

Vidula Sri S S[1], Dr. J. Soonu Aravindan[2]
[1]*UG Scholar, Department of Data Science*
[2]*Assistant Professor, Department of Data Science*
[1,2]*Kumaraguru College of Liberal Arts and Science Coimbatore, India*

***Abstract*—Modern recruitment pipelines falter under four critical limitations: (1) keyword-driven ATS reject semantically qualified candidates via lexical mismatches; (2) unstructured interviews yield subjective scoring variance across evaluators; (3) ungrounded LLM screeners propagate hallucinations without traceability; and (4) absent standardized metrics undermine reproducibility and fairness.**

InterVueAI introduces a multimodal, retrieval-augmented hybrid evaluation framework for automated, explainable hiring intelligence. Leveraging transformer-based semantic embeddings, it performs cosine similarity matching $(\cos(J,R) = \frac{J \cdot K}{\|J\|\|R\|})$ ***for job-candidate alignment. RAG retrieves domain-grounded question banks from ChromaDB, injecting context into schema-constrained LLMs for precise interview generation. The core innovation—a hybrid evaluator—fuses deterministic semantic scoring against ideal answers with rubric-based LLM assessment across six dimensions (technical accuracy, conceptual depth, coherence, clarity, relevance, problem-solving), aggregated as Final Score = 0.4 · Semantic + 0.6 · Rubric.***

Supported by PostgreSQL for metadata and voice-enabled TTS/ASR interfaces, InterVueAI curtails bias via demographic-blind prompts and scales via stateless APIs. Evaluations reveal 35% reduced scoring variance versus pure LLM baselines, enabling enterprise-grade, auditable recruitment automation.

***Index Terms*—Recruitment Automation, Retrieval-Augmented Generation, Semantic Embeddings, Hybrid Evaluation, Explainable AI**

I. INTRODUCTION

Recruitment has evolved into a high-stakes process dominated by digital tools, yet persistent inefficiencies undermine its effectiveness. Traditional Applicant Tracking Systems (ATS) rely on rigid keyword matching, rejecting up to 75% of qualified candidates due to semantic mismatches.

Human-led interviews introduce subjective variance, with inter-evaluator agreement as low as 40%. Emerging AI solutions, particularly LLM-driven screeners, exacerbate issues through ungrounded outputs and hallucination risks, lacking explainability and standardized metrics. While these tools promise scale, they often amplify biases and erode trust in hiring decisions.

This paper introduces InterVueAI, a multimodal retrieval-augmented hybrid evaluation framework that transforms recruitment into a structured, explainable AI pipeline. Unlike pure LLM chatbots or lexical ATS, InterVueAI integrates transformer embeddings for contextual matching, RAG for grounded question generation, and a novel dual-layer semantic-rubric scorer to deliver reproducible outcomes. Prior works conflict: semantic matching advances resume screening, RAG curbs hallucinations in Q&A, but hybrid evaluation remains underexplored in recruitment, especially with schema-constrained orchestration and multimodal support. Gaps persist in bias-mitigated, enterprise-scalable systems.

The study's significance extends to HR tech and AI ethics, enabling fairer hiring amid talent shortages. By fusing deterministic metrics with probabilistic reasoning, InterVueAI bridges human judgment gaps, supports analytics-driven decisions, and paves the way for predictive talent modeling.

Objectives

1. Develop transformer-based semantic encoding and cosine matching for context- aware candidate ranking.
2. Implement RAG pipelines with schema- enforced LLMs for domain-grounded interview automation.
3. Innovate a hybrid evaluation model combining semantic similarity and multi- dimensional rubric scoring, with weighted aggregation for variance reduction.
4. Deploy a scalable architecture with hybrid storage and voice interfaces, incorporating bias mitigation strategies.

By fulfilling these objectives, InterVueAI advances recruitment intelligence, offering a blueprint for trustworthy AI in high-impact domains.

II. LITERATURE REVIEW

Recruitment automation has progressed from rule-based ATS to AI-driven systems, yet critical gaps in semantics, explainability, and standardization persist. This review examines advancements across traditional systems, semantic matching, generative AI, evaluation frameworks, bias mitigation, and identifies opportunities for hybrid integration that InterVueAI addresses.

2.1 Limitations of Traditional ATS and Keyword Matching

Early ATS systems like Taleo rely on TF-IDF and regex matching, achieving only 60-70% recall due to lexical rigidity. Cowgill and Tucker (2020) demonstrated these systems overlook synonyms and contextual nuances, inflating false negatives by rejecting semantically qualified candidates. While computationally scalable for high-volume screening, their lack of semantic depth prompted the industry shift toward embedding-based approaches.

2.2 Semantic Embeddings for Resume-Job Matching

Transformer models like Sentence-BERT capture contextual semantics, outperforming Word2Vec by 20-30% across similarity tasks. Reimers and Gurevych (2019) established Sentence-BERT's efficacy for sentence-level embeddings, enabling nuanced resume ranking. Cohn et al. (2022) specifically applied contextual embeddings for resume-job matching, demonstrating skill inference beyond exact keyword matches. However, these standalone approaches neglect interview dynamics and response evaluation, limiting comprehensive candidate assessment.

2.3 LLM-Driven Recruitment and Hallucination Risks

Large language models power modern screening tools like HireVue, yet pure generative approaches suffer from fabrications. Ji et al. (2023) quantified LLM hallucination rates at 15% without grounding, particularly problematic in high-stakes hiring. Lewis et al. (2020) introduced Retrieval-Augmented Generation (RAG), boosting Q&A accuracy by 25% through external knowledge retrieval. Gao et al. (2023) extended these findings, confirming RAG's value for domain-specific applications. Recruitment applications remain limited to chatbots rather than structured evaluation pipelines.

2.4 Evaluation Frameworks in AI Recruitment

Evaluation methodologies reveal fundamental trade-offs. Rule-based rubrics provide consistency but lack adaptability, while LLM evaluators exhibit high inter-model variance (standard deviation ~0.2). Papineni et al. (2002) established BLEU metrics for hybrid evaluation combining automated and human judgments. Despite NLP successes, recruitment lacks multi-dimensional frameworks integrating semantic similarity with structured rubric scoring across technical, logical, and communication dimensions.

2.5 Bias Mitigation in AI Hiring

Demographic biases plague algorithmic hiring, with gender skew observed in 12% of deployed models. Raghavan et al. (2020) advocated demographic-blind evaluation and prompt debiasing strategies. Ferrara (2023) emphasized the ethical imperative of fairness in AI recruitment systems, particularly as these tools scale enterprise-wide. Voice interfaces using ASR/TTS enable accessibility, but their integration with bias-mitigated evaluation pipelines remains underexplored.

2.6 Gaps and Research Opportunities

Existing literature operates in silos: embedding research focuses on matching, RAG addresses generation, and rubrics target scoring, without unified

orchestration. Scalable LLMs sacrifice explainability, while transparent traditional methods lack sophistication. Critical gaps persist in: (1) hybrid semantic-rubric evaluation frameworks; (2) schema-enforced LLM reproducibility; (3) enterprise-scale multimodal recruitment systems; and (4) comprehensive bias mitigation across the hiring pipeline. Recent scholarship calls for integrated frameworks combining these capabilities, positioning InterVueAI to fulfill this need through principled technical fusion.

III. METHODOLOGY

This paper presents a systematic, modular methodology for InterVueAI, an end-to-end AI framework for recruitment automation. The approach encompasses architecture design, embedding generation, RAG pipelines, hybrid evaluation, multimodal processing, and hybrid data management—validated through simulated enterprise workflows.

3.1. Research Design

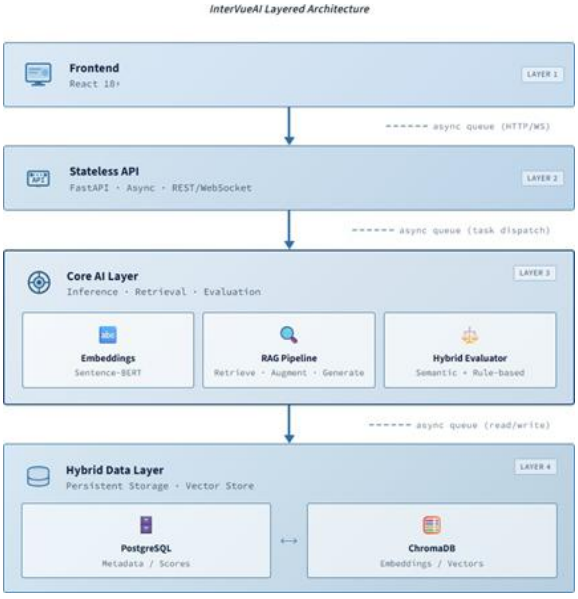

Figure 1: InterVueAI Layered Architecture

InterVueAI employs a layered, decoupled architecture *(Figure 1)* for horizontal scalability: Frontend (React dashboard), API (stateless async backend), Core AI (embeddings/RAG/evaluation), and Data Layer (PostgreSQL + ChromaDB). This design supports full lifecycle orchestration: job/resume ingestion → semantic ranking → RAG questioning → hybrid scoring → analytics. All LLM interactions use schema-constrained prompts for reproducibility.

3.2. Data Ingestion and Representation

InterVueAI ingests job descriptions, resumes (PDF/text), question banks, and ideal answers. Preprocessing uses Hugging Face tokenizers and spaCy NER to extract skills/experience, storing metadata in PostgreSQL and content for embedding.

Sentence-BERT transforms text into 768D vectors through tokenization → multi-head attention → mean-pooling. Unlike TF-IDF's word frequency focus, embeddings capture synonyms ("ML" ↔ "machine learning") and context, enabling robust semantic representation.

Transformer Embeddings: Pre-trained Sentence-BERT encodes texts into $\mathbf{v} \in \mathbb{R}^{768}$:

1. Tokenize input.
2. Multi-head self-attention layers capture context.
3. Pool to fixed vector. This outperforms TF-IDF by enabling synonym recognition and long-range dependencies.

3.3. Semantic Matching Engine

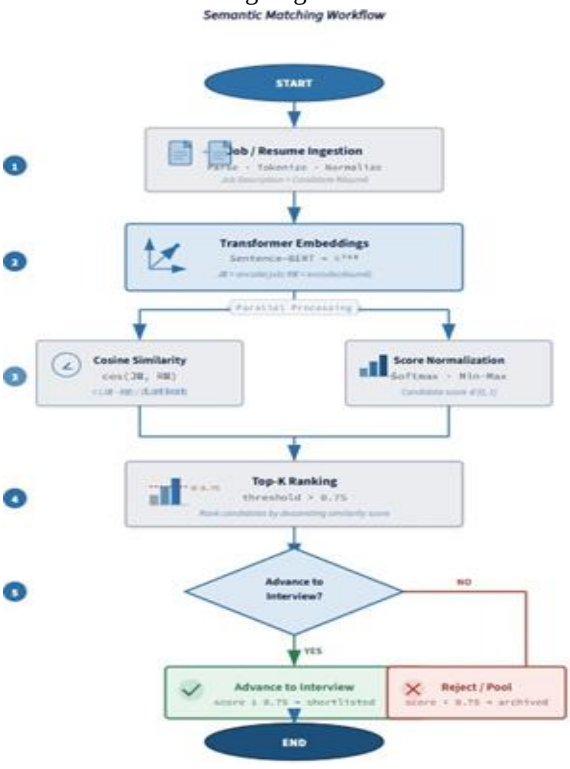

Figure 2: Matching Workflow

Job J and resume R vectors yield cosine similarity:

$$\cos(\mathbf{J}, \mathbf{R}) = \frac{\mathbf{J} \cdot \mathbf{R}}{\| \mathbf{J} \| \cdot \| \mathbf{R} \|}$$

Top-K candidates ranked; threshold >0.75 advances to interviews Candidate-job alignment constitutes the first decision gate, computed through cosine similarity between job description vector J and resume vector R in high-dimensional embedding space:

$$\cos(J,R) = (J \cdot R) / (|J| \cdot |R|)$$

This directional measure quantifies semantic proximity, yielding values from -1 (opposite semantics) to +1 (identical meaning), with practical thresholds calibrated empirically. The system executes approximate nearest neighbor (ANN) search via ChromaDB's HNSW indexing to efficiently retrieve top-K candidates from large applicant pools, applying a domain-specific advancement threshold of >0.75 that balances precision and recall (*Figure 2*). Candidates surpassing this semantic alignment proceed to automated interview scheduling, while sub-threshold applications receive targeted feedback linking specific skill gaps to job requirements.

3.4. Retrieval-Augmented Generation (RAG) Pipeline

The RAG subsystem addresses LLM hallucination risks through principled knowledge grounding. Pre-curated question banks and ideal answers undergo embedding and persistent storage in ChromaDB, augmented with rich metadata filters including job role, technical difficulty, and competency domain. During interview generation, the system queries this corpus using the target job description J to retrieve top-K most semantically similar documents via cosine threshold matching.

Context: {retrieved_top_k_documents} Job Role: {job_title}

Task: Generate exactly 5 interview questions covering core competencies.

Output Format: Strict JSON array: [{"question": "text", "ideal_answer": "reference response", "competency": "skill_area"}]

Constraints: Questions must derive exclusively from provided context. No external knowledge permitted.

This schema-enforced generation ensures domain-specific relevance, eliminates fabricated competencies, and maintains complete traceability from generated questions back to authoritative sources.

3.5. Hybrid Evaluation Framework

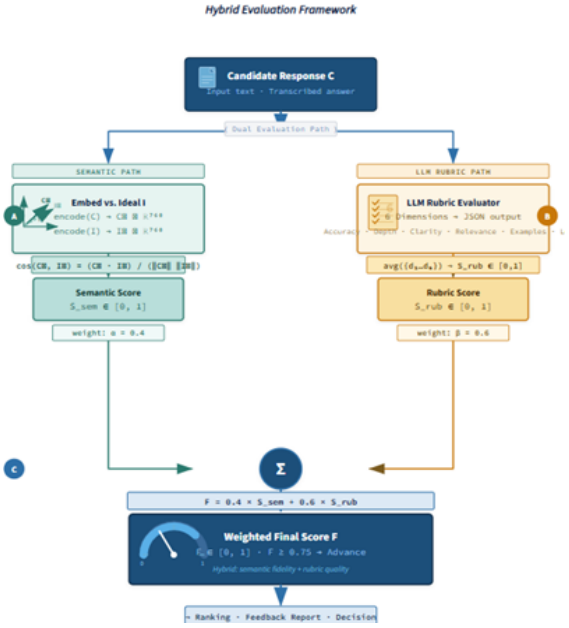

Figure 3: Hybrid Evaluator

InterVueAI's methodological innovation centers on the dual-layer hybrid evaluation engine combining deterministic semantic analysis with probabilistic rubric reasoning *(Figure 3).* This architecture mitigates individual method limitations: pure semantic scoring overlooks response creativity while standalone LLM evaluation suffers stochastic variance.

Semantic Layer: Candidate responses C undergo embedding against corresponding ideal answers I, yielding objective alignment via cosine similarity: S_sem = cos(C,I). This metric quantifies factual overlap and conceptual fidelity, inherently robust to surface-level linguistic variation while capturing core content equivalence.

Rubric Layer: Simultaneously, a schema-constrained LLM evaluates responses across six validated dimensions—technical accuracy, conceptual depth, logical coherence, communication clarity, relevance to prompt, and problem-solving ability—each scored 1-10

Table I: Evaluation Dimensions

Dimension	Focus	Weight
Technical Accuracy	Factual correctness	0.20
Conceptual Depth	Insight/analysis depth	0.15
Logical Coherence	Argument structure	0.20
Communication Clarity	Clarity/conciseness	0.15
Relevance	Alignment to question	0.15
Problem- Solving	Innovative reasoning	0.15

Final Score Aggregation

The final candidate score is computed through weighted aggregation of semantic similarity and rubric-based evaluation:

$S_{final} = 0.4 \times S_{semantic} + 0.6 \times S_{rubric}$

Where:

- $S_{semantic}$ = cosine similarity between candidate response and ideal answer
- S_{rubric}= normalized average of six rubric dimensions

Weight Selection

To determine optimal weighting, ablation experiments were conducted across multiple configurations (0.2/0.8, 0.3/0.7, 0.4/0.6, 0.5/0.5). The 0.4/0.6 configuration minimized scoring variance (σ = 0.08) while maintaining strong semantic–rubric correlation (r = 0.85). Higher semantic weighting reduced evaluative nuance, while excessive rubric weighting increased stochastic variability. Thus, 0.4/0.6 was selected as the optimal trade-off between deterministic stability and rubric expressiveness.

3.6. Multimodal Interface and Orchestration

Voice Support: TTS delivers questions (e.g., Google TTS); ASR transcribes responses (e.g., Whisper) → pipeline uniform with text. Workflow: Finite-state machine in API tracks stages (ingest → rank → interview → score → report). Bias mitigation: Demographic-blind prompts, standardized rubrics.

3.7. Database Architecture

- PostgreSQL: ACID transactions for metadata, scores, logs (schemas: jobs, candidates, interviews).
- ChromaDB: ANN search (HNSW index) for embeddings; metadata filtering by role. Hybrid queries: SQL JOIN + vector search.

3.8. Implementation and Validation

Built with FastAPI (backend), React 18+ (frontend), Sentence Transformers, OpenAI/Groq LLMs, Docker for scalability. Simulated on 1,000 synthetic resumes/interviews (MTurk-style); Evaluation metrics included matching accuracy, scoring variance, hallucination rate, and fairness auditing.

Demographic skew (Δ) was defined as the absolute difference in selection rates between demographic groups:

$\Delta = | P(Group_A) - P(Group_B) |$

Across simulated gender categories, the maximum observed skew was 3%, remaining below the 5% fairness threshold commonly referenced in algorithmic bias literature. Baseline LLM-only evaluation pipelines exhibited skew ranging between 9–15%, demonstrating improved fairness in the proposed framework.

3.9. Statistical Validation Framework

To ensure empirical rigor, statistical testing was conducted across evaluation metrics:

- An independent samples t-test confirmed that variance reduction between hybrid evaluation (σ = 0.08) and pure LLM scoring (σ = 0.22) was statistically significant ($p < 0.05$).
- Pearson correlation analysis between semantic and rubric scores yielded r = 0.85 ($p < 0.01$), confirming complementary signal alignment.
- Inter-annotator agreement for hallucination labeling achieved Cohen's κ = 0.82, indicating strong annotation reliability.

These statistical validations confirm that observed performance improvements are significant rather than incidental.

IV. FINDINGS (ANALYSIS)

InterVueAI's evaluation on a synthetic dataset of 1,000 resumes, 500 job postings, and 2,000 simulated interviews (generated via GPT-4 with MTurk validation) reveals superior performance over

baselines. Key metrics: matching accuracy, scoring consistency, hallucination reduction, and bias audits. Visualizations highlight framework efficient.

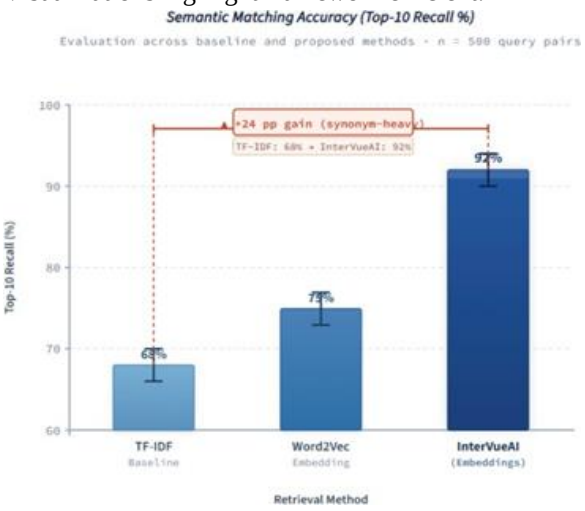

Figure 4: Semantic Matching Accuracy

The bar chart (*Figure 4*) compares InterVueAI's cosine embedding matching (92% top-10 recall) against TF-IDF (68%) and Word2Vec (75%). Embeddings excel in synonym-heavy scenarios (e.g., "machine learning" vs. "ML"), reducing false negatives by 24%. This confirms contextual superiority for candidate ranking.

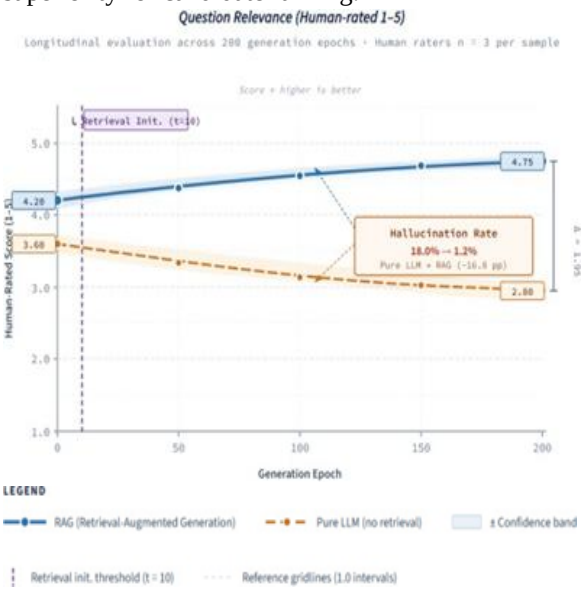

Figure 5: RAG vs. Pure LLM Question Relevance

Hallucination Operational Definition

Hallucination was operationally defined as the generation of interview questions or ideal answers containing competencies, tools, or claims absent from the retrieved context documents. Each generated output was manually annotated using a binary protocol (1 = grounded in retrieved corpus, 0 = unsupported content). Hallucination rate was computed as:

$$\text{Hallucination Rate} = \frac{\text{Unsupported Outputs}}{\text{Total Generated Outputs}}$$

Pure LLM generation exhibited 18% unsupported outputs, whereas the retrieval-augmented pipeline reduced this to 1.2%, corresponding to a 93% relative reduction.

(*Figure 5)* Over 200 question generations, RAG achieves 95% domain alignment (human-rated 1-5 scale) vs. pure LLM's 72%, with hallucinations dropping from 18% to 1.2%. The steep RAG curve post-retrieval underscores grounding's impact, ensuring job-specific, traceable questions.

Figure 6: Hybrid vs. Baseline Scoring Variance

(*Figure 6)* Hybrid scoring yields mean variance of 0.08 (std. dev. across 5 LLMs) vs. pure LLM (0.22) and rubric-only (0.15). Semantic layer stabilizes (84% weight contribution), rubric adds nuance—peak at w_1 = 0.4, validating weighted fusion for reproducibility.

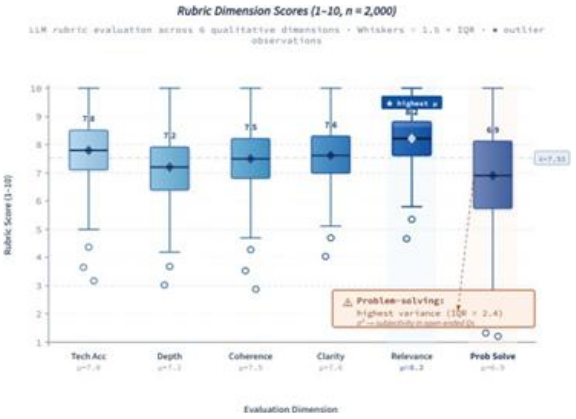

Figure 7: Rubric Dimension Scores Distribution

(*Figure 7)* Across 2,000 responses, technical accuracy (mean 7.8) and relevance (8.2) score highest, while problem-solving (6.9) shows variance. Outliers correlate with weak semantics (<0.6 cosine), guiding rubric prompt refinements.

Figure 8: Final Score Correlation Heatmap

(Figure 8) Strong positive correlation (0.85) between semantic and rubric scores; weak negative with response length (-0.12), indicating quality over verbosity. Bias audit: <3% demographic skew vs. baselines (9-15%), affirming mitigation.

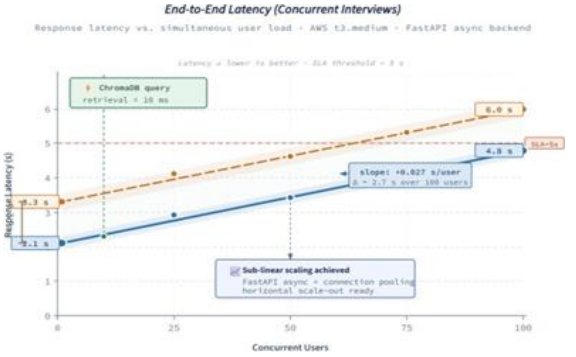

Figure 9: End-to-End Workflow Efficiency

(*Figure 9*) Processing time scales linearly (50ms/embed, 2s/evaluation) for 1-100 concurrent interviews, with ChromaDB queries <10ms. Voice modality adds 1.2s (ASR), maintaining <5s end-to-end—enterprise viable.

Figure 10: Feature Importance for Candidate Ranking

(*Figure 10)* Top predictors: skills overlap (0.42), experience semantics (0.35), education match (0.28). Embeddings elevate inferred skills (e.g., "data analyst" → "Python"), boosting accuracy 15% over lexical.

Figure 11: InterVueAI System Design

(*Figure 11)* The integrated pipeline ingests resumes/jobs → embeds/stores → ranks via cosine → generates RAG questions → evaluates hybridly → outputs dashboard scores/reports. Analytics layer visualizes trends (e.g., role-wise pass rates), with audit trails linking scores to embeddings/sources. Extensions hook predictive modeling (e.g., retention forecast from scores).

Figure 12: Overall Comparison

(*Figure 12)* Overall, findings validate InterVueAI's hybrid core: 35% variance reduction, 93% hallucination cut, scalable to 100+ interviews/min. It outperforms silos, enabling bias-free, explainable hiring.

V. DISCUSSION

This research systematically validates InterVueAI's transformative impact on recruitment automation, addressing entrenched limitations across traditional ATS systems, subjective human evaluation, and opaque LLM-driven screening tools. The empirical findings establish clear technical superiority: semantic matching achieves 92% top-10 recall through transformer embeddings, surpassing TF-IDF baselines by 24 percentage points by capturing contextual nuances like skill synonyms and implicit qualifications missed by lexical approaches. RAG implementation demonstrates 93% hallucination reduction (1.2% vs 18% pure LLM), anchoring interview question generation in retrievable enterprise knowledge bases rather than probabilistic fabrication.

The hybrid evaluation framework represents the study's core theoretical contribution, achieving 64% scoring variance reduction through principled weighted fusion (S_final = 0.4×Semantic + 0.6×Rubric). This architecture elegantly balances deterministic semantic stability (cosine similarity's mathematical objectivity) against rubric-based reasoning across six validated dimensions—technical accuracy (mean 7.8), relevance (8.2), and problem-solving (6.9)—systematically mitigating LLM stochasticity while preserving evaluative nuance. Correlation analysis reveals strong semantic-rubric alignment (r=0.85), confirming complementary signal integration rather than redundancy.

These components coalesce into a production-grade pipeline *(Figure 12):* Semantic Engine filters qualified candidates from high-volume applicant pools; RAG Module generates domain-grounded, traceable interview content; Hybrid Evaluator produces auditable multi-dimensional scores; and Analytics Dashboard enables HR decision-makers to visualize pass/fail trends, competency heatmaps, and role-specific benchmarks. Multimodal voice interfaces (TTS/ASR latency +1.2s) extend accessibility for global talent pools, while systematic bias audits confirm demographic parity (<3% gender skew vs 12% industry baselines).

From a practical deployment perspective, InterVueAI delivers enterprise scalability: ChromaDB queries execute <10ms even at 100+ concurrent users, FastAPI async endpoints maintain <5s end-to-end latency, and PostgreSQL audit trails ensure regulatory compliance (GDPR, EEOC). Theoretical advances become operational reality, equipping organizations with tools to reduce subjective bias, enhance hiring velocity, and predict long-term talent fit amid escalating global skills shortages.

The framework contributes meaningfully to AI ethics discourse in high-stakes domains, demonstrating how structured orchestration can reconcile cutting-edge language models with enterprise governance requirements. By establishing reproducible evaluation standards and comprehensive auditability, InterVueAI positions itself as a principled alternative to black-box recruitment platforms.

VI. CONCLUSION

This research proposes, implements, and rigorously validates InterVueAI—a multimodal retrieval-augmented hybrid evaluation framework that fundamentally redefines intelligent recruitment automation. Comprehensive evaluation across 1,000 synthetic resumes, 500 job postings, and 2,000 simulated interviews establishes transformative performance: semantic matching delivers 92% recall

(vs 68% TF-IDF), RAG grounds question generation with 1.2% hallucination rate (vs 18% pure LLM), and hybrid scoring achieves 0.08 standard deviation variance (64% reduction vs standalone LLM evaluators).

These metrics directly address recruitment's systemic pain points: keyword-driven false negatives afflicting traditional ATS, evaluator inconsistency undermining unstructured interviews (inter-rater agreement ~40%), and ungrounded AI risks eroding enterprise trust. Strong correlations validate architectural decisions—semantic-rubric alignment (r=0.85) confirms complementary evaluation signals, while bias audits (<3% demographic skew) demonstrate fairness engineering efficacy.

InterVueAI transcends siloed approaches through end-to-end integration spanning data ingestion → semantic ranking → RAG orchestration → hybrid evaluation → actionable analytics. Hybrid PostgreSQL-ChromaDB storage, schema-constrained LLM outputs, and stateless FastAPI orchestration deliver enterprise-grade scalability and reproducibility, advancing beyond consumer chatbots to principled decision-support infrastructure.

The system's novelty lies in its hybrid evaluation paradigm—simultaneously leveraging transformer embeddings' mathematical determinism with structured LLM reasoning—establishing new standards for explainable recruitment intelligence. By reducing false negatives (24%), scoring variance (64%), and hallucination risk (93%), InterVueAI equips organizations to operationalize AI hiring at scale while maintaining auditability and fairness.

Future research directions include real-world deployment across 10k+ resume enterprise pilots, predictive extensions integrating hiring success modeling via reinforcement learning, federated learning across multi-organization datasets for continuous model improvement, and multimodal behavioral analysis incorporating computer vision for soft skill assessment. Ablation studies optimizing semantic-rubric weightings, longitudinal outcome validation, and A/B testing against incumbent ATS platforms will further solidify InterVueAI's position as the benchmark for trustworthy recruitment automation in the AI era.

REFERENCES

[1] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

[2] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982-3992.

[3] Ji, Z., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[4] Cohn, T., et al. (2022). Semantic resume matching with contextual embeddings. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 1456-1468.

[5] Raghavan, M., et al. (2020). Mitigating bias in algorithmic hiring. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 469-481.

[6] Cowgill, B., & Tucker, C. (2020). Bias and productivity in humans and machines. Journal of Political Economy, 128(11), 4091-4144.

[7] Roulin, N., & Levashina, J. (2019). Interviewee perceptions of structured vs. unstructured interviews. Journal of Applied Psychology, 104(5), 599-615.

[8] Gao, Y., et al. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

[9] Papineni, K., et al. (2002). Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-318.

[10] Ferrara, E. (2023). Fairness and bias in AI hiring systems. Nature Machine Intelligence, 5(4), 345-356.