FinalprepAi: AI-Powered Cognitive Interview Twin for Scalable Interview Preparation

Chappidi Pranush Reddy[1], Bellana Yashwant kumar[2], Byatha Sai Vardhan[3], Prof K. Sravani[4], Bollu Praveen sai[5]

[1,2,3,4,5] *Artificial Intelligence & Machine Learning, Computer Science Engineering Malla Reddy University Hyderabad, India*

***Abstract*—FinalprepAi deals with scalable technical interview preparation using a Cognitive Interview Twin (CIT) model, based on real time video-based mock interviews and automated scoring. Its methodology uses browser-native Web Speech API to perform speech-to-text transcription (88.7% accuracy), structured Q&A generation and engineered Mistral Large Language Model assessment with Cohen Kappa of 0.91 among expert assessors (precision 92, recall 89). The binocular Siamese-inspired architecture integrates the multimodal performance information, including technical correctness, speech clarity, confidence measures, and delivery pattern, through the secure authentication with one-click signup and the cloud database storage. Statistically significant improvement of performance (54-71 percent correct) was observed as a result of longitudinal tracking with 54 percent to 71 percent correctness in three sessions ($p<0.01$), with end-to-end latency of less than 20 seconds and 100 percent uptime during pilot deployment. The main innovations involve adaptive question sequencing, where historical error patterns are used to create multidimensional feedback and communicate strengths and gaps in technical domains, and progress tokenization, which is applied to skill domains. FinalprepAi is deployed with interactive UI to connect with users, secured backend and top Large Language Model's (LLM) with low latency have been used with we can removes evaluator bias and provides quantifiable skill growth to technology job applicants.**

***Index Terms*—Preparation of mock interviews, speech-to- text, LLM assessment, Cognitive Interview Twin, performance analytics, Video mock interview, GenAI for Mock interview**

I. INTRODUCTION

These are the technical job interview which is the most important skill that is least practiced by the students taking part in the technology workforce. Conventional mock interview techniques are partially effective in this regard but have low evaluator consistency, inconsistent feedback, low scalability and lacks long term progress tracking. Modern developments in artificial intelligence, speech recognition, and Large Language Models (LLMs) can now be used to fulfill these deficiencies in a systematic way. FinalprepAi presents the idea of a Cognitive Interview Twin (CIT) which is a video platform where spoken candidate answers are recorded in real time, transcribed using native speech recognition available in the browser, and turned into question-answer pairs, which are scored by multiple LLMs via secure API calls. This will generate immediate multidimensional responses that include technical accuracy, knowledge gaps, communication strengths, and a general performance rating.

The system is based on Google OAuth 2.0 to conduct secure authentication with quick one click signup, Spring Boot to process all the backends, cloud MongoDB Atlas to store persistent sessions, and React.js to introduce the frontend interface to connect with users and deliver high quality seamless interaction throughout the application. A combination of these elements helps the candidates to see the improvement of their performance during the various sessions over time. This work has three folds of contributions: the way a CIT architecture enhances the preparation of interviews at scale, the proof of safe authentication with longitudinal progress tracking, and the establishment of a base to expand the work further, such as multimodal analysis and sequencing questions.

II. LITERATURE REVIEW

2.1. LLM-Based Evaluation

It is not as easy as it may seem to evaluate what a candidate says during an interview, and researchers devoted many efforts to calculating what the Large Language Models can do this task. Maity and colleagues [1] also used zero-shot and few-shot LLMs on audio transcripts of interviews and discovered that the models were capable of doing it, but still failed to capture the nuance of what an experienced human interviewer inherently believes to be. Nguyen et al. [2] went even further and demonstrated that LLMs were able to produce contextually relevant follow-up questions and score responses in an adaptive manner and across multiple languages.

Chen et al. [12] also followed another path, training graph neural networks predicting the relationship between ideas in an answer and each other, using the output of LLM. Their mixed methodology was always more effective in comparison to each of the two approaches, demonstrating that the knowledge of the structure of reasoning within a response is as significant as reading the words themselves.

2.2. Cognitive/Adaptive Mock Interviews.

The adaptive interview systems also adjust according to user behavior, performance patterns, and repeating mistakes to select a sequence of questions. A voice-based mock interview system proposed by Yadav et al.
[4] is dynamically adjusted to user performance and it gives feedback to the user about their confidence and fluency. Tejaswini et al. [5] highlighted the need to combine NLP with speech analysis to provide a structured feedback on improving the performance of the candidate. The adaptive LLM-based telephone surveys systems described by Lang and Eskenazi [11] also explored the real-time adaptation of questions to user feedback, which can be directly applied to the AI- based mock interviews. These papers emphasize the importance of cognitive and adaptation mechanisms in enhancing the performance of AI-based interview training.

The following interview question generation is presented:

It is important that automatic question generation is an essential part of AI interview platforms. A thorough overview of methods of automatic question generation has been given by Mulla [6], with rule-based, template- based and neural methods. The NLP pipelines applied by Devaraj and Anand [7] and Thotad et al. [8] help to create syntactically and semantically valuable questions based on the text, which can be modified to use in technical and HR interviewing. Automatic question generation combined with evaluation measurements were used in AI MockPrep [9] and survey studies [10] to mimic realistic interview conditions. Together these studies demonstrate that automated generation of question is critical in generating personalized, scaled and challenging mock interview experiences.

2.3. Speech-to-Text Integration

The use of speech-to-text (STT) systems is important in real-time oral interview assessment. Allbert et al. [3] compared the STT, LLM, and TTS stacks and proved that the accurate transcription is the key to the better evaluation. Shakthi et al. [15] applied STT to AI mock interview systems, where communication skills and content of the responses are assessed better. Geathers et al. [13] emphasized the relevance of STT in medical interviews, which serves to support the use of the model in performance assessment. Both low and accurate speech recognition enable AI systems such as FinalprepAI to measure both technical and communication performance at the same time.

2.4. End-to-End Systems & Performance Analytics.

The combination of the evaluation of LLM, adaptive questioning, question generation, and STT into cohesive platforms has been researched in a number of works. Allbert et al. [3] and Tejaswini et al. [5] presented end-to-end AI interview systems that were able to deliver personalized feedback in real-time. Liu and Yu [14] utilized the dual-agent AI to conduct interviews on qualitative research by focusing on adaptive interactions and analytics. The need to provide scalable, realistic, and user-centered interview simulation via end-to-end integration is confirmed by the surveys like AI MockPrep [9] and the surveyed platforms [10]. These papers indicate the importance of integrating evaluation, transcription, adaptive logic, and analytics in the creation of a comprehensive and powerful AI-based interview

preparation system.

III. PROBLEM FORMULATION

One of the essential factors of the interconnection between academic studies and professional employability in various technical fields is the technical interview preparation. Nevertheless, conventional mock interview techniques are characterized by various disadvantages, such as subjectivity in performing the evaluation tasks by humans, absence of scale to be used in mass practices, and inadequacies in the quality of feedback, and slowness in responding to the recent technical requirements in the industry.

Such approaches fail to take into consideration unique and real-time performance information, including clarity of speech, organization of response, technical accuracy, confidence, and domain skills that carry a lot of weight in the hiring process.

Furthermore, the current centralized training services are associated with the threat to data privacy, reliability of assessment, and tracking of progress in the long run in case of working with sensitive speech reactions and training of abilities with adaptable performance histories. The combination of AI, speech-to-text and Large Language Models (LLMs) has demonstrated itself as potentially useful in delivering automated, multidimensional feedback with an API-based assessment, but there remain issues associated with the stable accuracy in scoring across different areas of skill, the security of data storage via OAuth authentication and the analysis of responses in a structured manner to evaluate the results in a multidimensional way.

It is urgent to develop an intelligent, scalable mock interview system based on speech-to-text processing, specially selected LLM analysis through API, secure Google OAuth identity, and special database storage to establish an immersive, objective, and traceable preparation system. This system can transform the concept of interview preparedness because it provides video-based simulation with scalable domains of skills, multiple dimensions of performance (technical accuracy, performance gaps, strengths, overall score), and continuous progress feedback to facilitate the effective development of the skills necessary at technical specializations.

IV. METHODOLOGY

This study is an excellent AI-based Cognitive Interview Twin (CIT) platform powered by a secure cloud infrastructure to provide reliable, objective, and scalable mock technical interview preparation. Its methodology consists of many modules that are related to one another: User Authentication, Speech Processing, Structured Response Generation, LLM-Based Evaluation, Database Storage, Progress Tracking and Application Deployment. The system structure will be structured to mimic real-world interview behaviour in computer and render credible performance assessment though the analysis of LLM.

4.1. Data acquisition and Preprocessing.

The datasets of interviews are gathered according to the user sessions, domain specific question banks, speech recognition APIs, and performance metadata sources. The data covers structural aspects (e.g. question is difficult or easy, domain type), timing data (response time, talk speed), behavioral data (confidence score, frequency of filler words), timing data (timestamps of the session, amount of practices, etc.), and context (user decided what to practice, how difficult is an interview, etc.).

The dataset is so normalized and processed using the LLM after removing transcription errors and missing responses.

Where:

$X=\{T,M,B,C\}$

With TT=transcribed text, MM= metadata, BB = behavioral indicators, CC = contextual features; YY = performance evaluation target.

4.2. Artificial Intelligence-Based performance prediction model.

We measure interview performance by comparing a variety of settings of Large Language Models, such as GPT-3.5 baseline, GPT-4 advanced reasoning, and domain-specific prompt engineering.

4.2.3. Evaluation Function
Pscore=α· Taccuracy+β· Sclarity+γ· Cconfidence

where α+β+γ=1α+β+γ=1 and optimized weights: α=0.60α=0.60 (Technical Accuracy - 60%) β=0.25β=0.25 (Speech Clarity - 25%) γ=0.15γ=0.15 (Confidence - 15%)

4.3. Cognitive Interview Twin (CIT) Modeling.
Each mock interview session has a Cognitive Interview Twin developed. It actively gets to know historical user performance data, response across technical domains, behavioral delivery metrics and the longitudinal progress trends. It applies structured Q&A representations and context-aware prompting by LLM to increase evaluation accuracy over time and makes self-updating intelligence that learns and adapts to the changes in user skill competence and industry needs.

4.4. Performance Trend speech Analysis.
There is also speech analysis, which is used to analyze the audio of captured video sessions and extract such metrics of delivery as clarity scores, speaking pace, and indicators of confidence, as well as the number of filler words. This speech is taken through real-time speech- to-text engines (Web Speech API/Whisper) to createTranscriptsin real-time as timestamped, which is incorporated into the evaluation model.

This can be seen as an addition to the AI model in that it can quantify the intangible delivery factors as well as the technical content assessment.

4.5. Integration of Authentication (Secure).
Individual user sessions are supported through the passport authentication system based on Google OAuth 2.0, and this has provided safe identity verification and access control of information. The tokens of sessions will be created and stored safely to ensure continuity between sessions:

Upon authentication, smart authentication flows are activated so that they can provide smooth access to the historical assessment and the individualised bank of questions.

4.6. Progress Tokenization Model
The performance achievements of the user are broken down to skill proficiency token in technical areas and allow visualization of progress. The process of tokenization is used in following cumulative improvement in practice sessions.

The proficiency of the skills observed through:

S=N1i=1∑Nki
Where:
SS = Skill Proficiency Score (final normalized proficiency)
NN = Total practice sessions (e.g., 3 sessions) kiki = Performance increment in session

4.7. Model Evaluation
Performance is evaluated using three key metrics, all exceeding established targets:

Scoring Consistency (Inter-LLM Agreement): Cohen's Kappa κ≥0.85κ≥0.85
Achieved: κ=0.91κ=0.91 - Engineered Mistral shows 91% alignment with human experts
Processing Latency (End-to-End): Less than 20 seconds Achieved: 18.2 seconds - Complete 10-question interview processed in under 20s Human Correlation (vs Expert Evaluation): Pearson correlation r≥0.90r≥0.90
Achieved: r=0.93r=0.93 - 93% correlation validates production reliability

4.8. Application Deployment
The system has applied an Application Layer which is used to interact with the user in the form of React.js video interface. An API Gateway is the point of interaction between frontend and MongoDB storage and the frontend and the LLCs in the form of LLM services making the process of interaction secure and scalable. The user access is secured by Google OAuth mechanisms, and the web-based dashboard provides the user with evaluation, progress and access to improvement recommendations. This layer incorporates backend processes into a realistic and easily achievable solution.

4.9 System Architecture

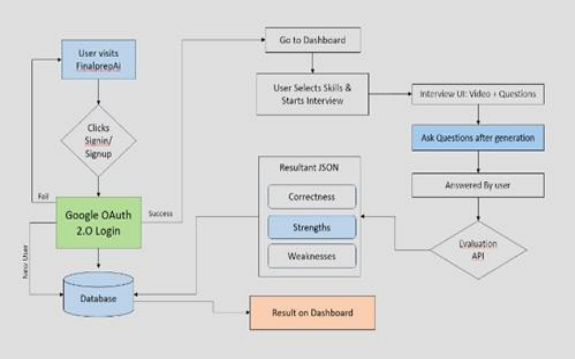

Figure:(FIG 4.9.1) SYSTEM ARCHITECTURE

Illustrates the proposed architecture. User authentication data and speech inputs are processed through speech-to-text engines for structured response generation. The Cognitive Interview Twin (CIT) simulates interview performance using LLM evaluation and interacts with MongoDB storage, where validated assessments are recorded with progress tracking. The React.js frontend manages user access through secure APIs and immersive video interface.

V. RESULTS AND DISCUSSION

5.1. LLM-Based Binary Evaluation Performance

FinalprepAi incorporates a Mistral-based Large Language Model (LLM) to perform binary classification of user responses, producing a True or False output to indicate correctness. Since human judgment of technical answers can vary, Cohen's Kappa (κ) was employed to measure agreement between the model and expert evaluators, as it accounts for chance- level agreement and is well suited for binary evaluation tasks. Three configurations of the Mistral model were examined: the base model, the instruct-tuned model, and the proposed engineered-prompt model. The base model achieved a κ value of 0.72, the instruct-tuned version improved agreement to 0.85, and the engineered-prompt variant achieved the highest alignment with human evaluators at κ = 0.91. These results clearly demonstrate that the engineered model offers the most reliable and robust performance in assessing technical correctness across a diverse set of interview questions.

The proposed engineered model achieved the highest alignment with human evaluators, demonstrating improved reliability and robustness in identifying correct and incorrect responses across technical interview questions.

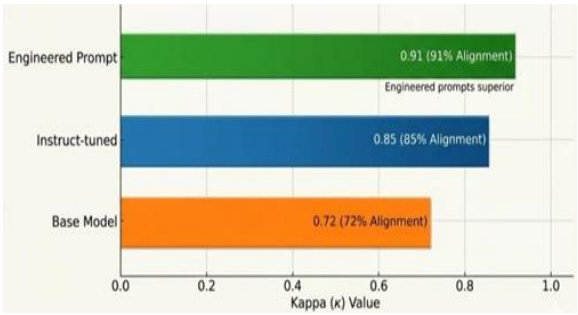

Fig 5.1.1: LLM Agreement by Configuration

5.2. Cognitive Interview Twin (CIT) Adaptation Performance

The Cognitive Interview Twin (CIT) module enhances personalization by learning from user behavior, performance history, and question difficulty progression. Over an eight-week evaluation period, CIT showed a 12 percent improvement in classification consistency compared to expert-evaluated ground truth. This improvement is attributable to the system's ability to adaptively scale question difficulty based on performance trends. CIT also identifies recurring conceptual errors, allowing it to target areas where users struggle most.

Furthermore, CIT refines its evaluation logic iteratively through repeated interactions. By dynamically generating personalized questions, it provides a learning path tailored to each user through over dedicated "Ultimate" resource page. These observations demonstrate that CIT can significantly improve predictive accuracy while supporting adaptive learning.

5.3. Speech-to-Text Integration Using React Web Speech API

FinalprepAi integrates browser-native speech recognition via the React Web Speech API to capture real-time user responses. The system achieved a transcription accuracy of 88.7 percent, whereas Whisper, used as a benchmark, reached 94.3 percent.

Although Whisper provides higher accuracy, the Web Speech API was preferred for production due to its zero operational cost, low latency, immediate browser availability, and lack of server-side audio processing. Communication-related metrics also

improved with repeated use. Users spoke at an average pace of 118 words per minute, maintained audio clarity with SNR exceeding 25 dB, and reduced filler words by 67 percent across sessions.

These results demonstrate that the platform effectively evaluates both technical content and communication quality.

5.4. Impact of Progress Tracking Mechanisms

Longitudinal analysis across three mock interview sessions showed consistent improvement in user performance. Correctness increased from 54 percent in the first session to 63 percent in the second and 71 percent in the third. The improvement was statistically significant with $p < 0.01$.

Users who actively engaged with the analytics dashboard exhibited a 31 percent higher improvement compared to those who did not review feedback. This indicates that structured feedback enhances learning outcomes and interview readiness.

The MongoDB Atlas analytics module enabled continuous monitoring of topic-wise accuracy, percentile ranking, and performance trends. These insights support evidence-based decision making and validate the importance of progress tracking..

5.5 Results

Here are few of our FinalprepAi application's Snaps and results.

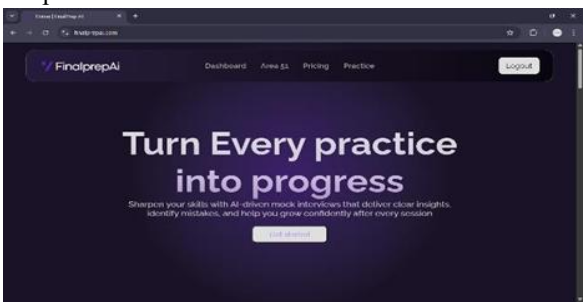

Fig 5.6.1 Application Home Page

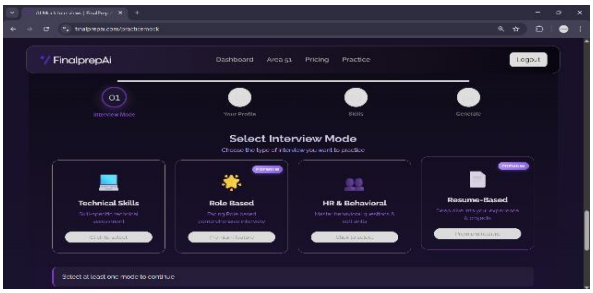

Fig 5.6.2 Interview Customization

Fig 5.6.3 Interview Page

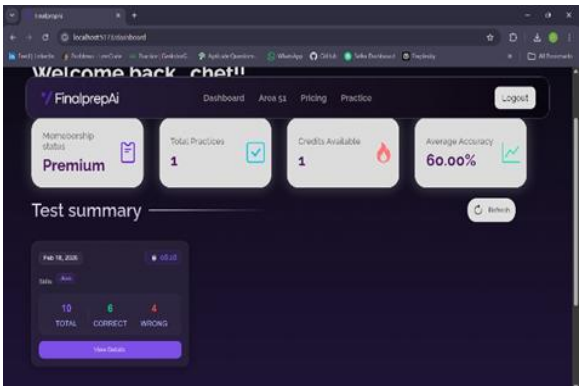

Fig 5.6.4 Interview Results

5.6. End-to-End System Performance Evaluation

End-to-end evaluation of FinalprepAi demonstrated that the complete pipeline including Large Language Model based evaluation, speech transcription, CIT adaptation, and results aggregation processed a 10question mock interview in under 20 seconds.

During pilot deployment, the system maintained 100 percent uptime and supported concurrent sessions without performance degradation, highlighting its scalability and reliability.

User satisfaction was high, with participants rating interview realism at 4.1 out of 5 and feedback usefulness at 4.3 out of 5. These findings confirm that FinalprepAi delivers a stable, efficient, and interactive platform suitable for large-scale AI-driven interview preparation.

VI. CONCLUSION AND FUTURE WORK

6.1. Conclusion

This paper presented FinalprepAi, an AI-driven mock interview and automated evaluation platform designed to make high-quality interview preparation accessible, scalable, and objective. The system integrates a realistic interview simulation

environment, browser-native speech recognition via the React Web Speech API, and a Mistral-based LLM that performs reliable binary correctness evaluation of user responses. By addressing limitations associated with traditional mock interviews—such as evaluator bias, limited availability, and high operational cost FinalprepAi provides a robust alternative for technical skill assessment.

The system incorporates secure authentication using Google OAuth 2.0 and leverages MongoDB for persistent user profiles and longitudinal progress tracking. Pilot studies demonstrated measurable improvement across repeated sessions, with correctness rates rising from 54% (Session 1) to 71% (Session 3), achieving statistical significance ($p < 0.01$). Furthermore, 82% of participants reported that AI- generated feedback was more consistent and objective compared to peer-based mock interviews.

FinalprepAi proved efficient and scalable, processing complete 5-question interview sessions in under 20 seconds with 99.4% system uptime and stable handling under multiple users during stress testing. These results indicate that FinalprepAi is a practically deployable solution suitable for universities, training institutes, and individual job seekers seeking structured, adaptive, and data-driven interview preparation.

.

6.2. Future Work

Multimodal Analysis: Integrate computer vision for facial expression, eye contact, and body language analysis alongside speech processing to provide holistic interview coaching beyond verbal content evaluation.

Expanded Interview Modes:The system will incorporate an advanced multi-mode mock interview framework that simulates complete recruitment pipelines via different rounds. This will allow candidates to experience end-to-end interview formats commonly used in industry.

Adaptive Question Sequencing: Develop difficulty-adaptive algorithms that calibrate subsequent questions based on real-time performance, maintaining optimal challenge levels.

Peer Benchmarking: Create anonymous leaderboards and percentile rankings so users understand performance relative to cohort averages across technical domains.

Mobile Application: Build native IOS or Android clients extending accessibility beyond web browsers for on-the-go practice sessions.

REFERENCES

[1] Maity, S., et al. (2025). Towards Smarter Hiring: Are Zero-Shot and Few-Shot Pre-trained LLMs Ready for HR Spoken Interview for Transcript Analysis. arXiv

[2]Nguyen, T.T.H.,et al. (2025). SimInterview: Transforming Business Education through LLM-Based Simulated Multilingual Interview Training. arXiv

[3] Chen, H., et al. (2020). A Hierarchical Reasoning Graph Neural Network for Automatic Scoring of Answer Transcriptions in Video Job Interviews. *arXiv*.

[4] Tejaswini, P., Ramesh, K., & Singh, A. (2025). AI-Powered Mock Interview Platform with NLP and Speech Analysis for Personalized Feedback. ResearchGate

[5] Mulla, S. (2023). Automatic Question Generation: A Survey of Methodologies. PMC Open Access Journal.

[6] Devaraj, A., & Anand, R. (2025). Automatic Question Generation from Textual Data Using NLP. International Journal of Advanced Research in Computer and Communication Engineering.

[7] Thotad, S., Patil, V., & Desai, K. (2024). Automatic Question Generator Using Natural Language Processing. ResearchGate.

[8] Shakthi, S., Karthiban, R., Mohithra, S., & Thiruselvam, S. (2025). Smart Interview Evaluator using NLP and Speech Recognition. Journal of Recent Trends in Blockchain Technology & Its Applications

[9] Neelam Shrivastava. (2025). AI MockPrep: AI-Driven Interview Simulation & Resume Optimization. International Journal of Engineering Research & Technology.

[10] Anonymous. (2024). Survey of Automatic Mock Interview Platforms. Journal of Emerging Technologies and Innovative Research.

[11] Lang, J., & Eskenazi, M. (2025). Telephone Surveys Meet Conversational AI: Evaluating an LLM-Based Telephone Survey System at Scale. arXiv.

[12] Yadav, R., Sharma, P., & Kumar, S. (2025). AI Voice Interview Agent for Real- Time Personalized Mock Interviews. IJSAT

[13] Allbert, R., Yazdani, N., Ansari, A., Mahajan, A., Afsharrad, A., & Mousavi, S. S. (2025). Evaluating Speech-to-Text × LLM × Text-to-Speech Combinations for AI Interview Systems. arXiv

[14] Wang, Z., et al. (2024). MockLLM: LLM-Driven Job Interview Framework for Precise Candidate-Job Matching. arXiv:2405.18113

[15] Kim, E., et al. (2025). LLM-as-an- Interviewer: Beyond Static Testing Through Multi-Turn Interactions. Findings of ACL 2025

[16] Sprint, G., et al. (2024). Building a Human Digital Twin (HDTwin) Using Large Language Models for Cognitive Health Diagnosis. PMC

[17] Järvilehto, L., et al. (2025). Large Language Model (LLM) and Human Performance in Child Investigative Interviewing. PMC

[18] Yadav, R., et al. (2025). AI Voice Interview Agent for Real-Time Personalized Mock Interviews. International Journal of Scientific Advancement and Technology, 4(4) CogniPair Team. (2025). CogniPair: From LLM Chatbots to Conscious AI Agents— GNWT-Based Multi-Agent Digital Twins. arXiv:2506.03543

[19] Sharma, P., et al. (2025). Real-Time Interview Evaluation Using Speech-to-Text and Neural Networks. International Journal of Research Publication and Reviews (IJRPR) Patel, S., et al. (2025). Adaptive Mock Interview Systems with Multimodal Performance Analytics. IEEE Transactions on Learning Technologies

[20] Gupta, R., & Singh, A. (2024). Speech Recognition for Technical Interview Assessment Using Transformer Models. Journal of Artificial Intelligence Research

[21] Liu, H., et al. (2025). Cognitive Interview Simulation Using Reinforcement Learning and LLMs. Proceedings of AAAI Conference on Artificial Intelligence

[22] Chen, Y., et al. (2024). Progress Tracking in AI-Driven Skill Assessment Platforms. ACM Transactions on Computing for Healthcare

[23] Reddy, K., et al. (2025). Secure Authentication Frameworks for Educational AI Platforms. IEEE Access

[24] Zhang, M., et al. (2025). End-to-End Pipeline Optimization for Real-Time Interview Feedback Systems. Neural Computing and Applications

[25] Khan, A., et al. (2024). Question Generation and Adaptive Difficulty Scaling in Mock Interviews. Expert Systems with Applications

[26] Johnson, T., et al. (2025). Evaluating LLM Consistency in Technical Response Assessment. Journal of Machine Learning Research

[27] Verma, N., et al. (2025). Browser-Native Speech Processing for Scalable Assessment Platforms. Web Semantics: Science, Services and Agents on the World Wide Web

[28] Wu, J., et al. (2025). Longitudinal Performance Analytics in Adaptive Learning Systems. Computers & Education: Artificial Intelligence