International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# Benchmarking LLMs and AI-Driven Speech Processing for Interviews: An End-to-End Pipeline

Gopalsingh Saraf
*Department of Information Technology*
*Pune Institute of Computer Technology*
Pune, India gopalsaraf02@gmail.com

Rishikesh Revandikar
*Department of Information Technology*
*Pune Institute of Computer Technology*
Pune, India
rishikeshrevandikar3110@gmail.com

Prathamesh Shriramwar
*Department of Information Technology*
*Pune Institute of Computer Technology*
Pune, India
prathameshshriramwar100@gmail.com

Shivanjali Thorat
*Department of Information Technology*
*Pune Institute of Computer Technology*
Pune, India shivanjalithorat28@gmail.com

Mrs. S. A. Jakhete
*Department of Information Technology*
*Pune Institute of Computer Technology*
Pune, India sumeetra.kasat@gmail.com

***Abstract*—The competitive nature of today's job market has amplified the challenges faced by both freshers and professionals, particularly those with limited practical experience, leading to increased anxiety, lack of confidence, and inadequate performance. This paper addresses the urgent need for personalized and effective interview preparation tools by introducing an AI- powered mock interview platform designed to simulate real- world scenarios and offer comprehensive, actionable feedback. Leveraging advanced AI models, including Llama 3.1 for dynamic, context-aware simulations, Coqui for high-quality text-to- speech output, and Whisper for accurate speech-to-text process- ing, the platform delivers a full-featured solution for technical, behavioral, and HR interviews. Key features include real-time transcription, intelligent feedback, and in-depth performance assessments to improve communication skills, boost technical readiness, and build overall confidence. Our findings demonstrate that AI-driven solutions can significantly enhance the interview preparation process, offering scalable, unbiased, and highly effective experiences that empower candidates to succeed in a rapidly evolving job market.**

***Index Terms*—Mock Interviews,Large Language Models, Intel- ligent Feedback, Performance Assessment, Llama 3.1, Whisper, Coqui**

## I. INTRODUCTION

In today's competitive job market, freshers and working professionals face challenges in interview preparation, includ- ing lack of practical experience, inadequate communication skills, and limited exposure to professional formats. Both groups struggle to update their skills and adapt to evolving industry standards, resulting in a gap between their capabilities and job market demands.Numerous studies have explored the potential of AI-driven solutions in providing individualized support for job preparation. Research conducted by Patil et al. [14] emphasized the potential of AI(Artificial Intelligence) to cater to specific user needs, helping bridge learning and skills gaps. Likewise, Babashahi et al. [5] showcased the role of AI in offering real-time, data-driven feedback, highlighting

how this technology could be applied to mock interviews. These studies reveal the untapped potential of AI technologies in interview preparation, especially for those seeking more targeted, tailored approaches to improve their performance.

The current state of interview preparation tools exhibits sig- nificant limitations. Traditional resources available to freshers, such as online tutorials and college placement services, often lack real-world applicability and fail to provide the personal- ized guidance necessary for building confidence and technical proficiency. Working professionals face similar challenges, struggling to find platforms that adequately address their spe- cific needs for

career transitions or skill enhancement. While platforms like Pramp offer peer-to-peer practice and HireVue provides video-based solutions, they have inherent limitations. Pramp's effectiveness depends heavily on peer quality, while HireVue lacks interactive elements. Although Interviewing.io connects users with industry professionals, its availability is restricted by human schedules. Recent studies have shown that AI-powered interview systems can reduce hiring time by up to 40% while maintaining high assessment accuracy [15]. Furthermore, research indicates that such systems can effectively combine interview preparation with skill assess- ment [17], making them particularly valuable for educational and professional development contexts. The adoption of AI interview systems has been shown to follow specific decision- making frameworks that balance technical capabilities with human factors [16].

We propose an AI-powered mock interview platform for freshers and professionals, offering comprehensive interview simulations with real-time, adaptive feedback. Users can track their progress, receive personalized recommendations, and re- fine their strategies. This solution provides consistent, scalable, and interactive interview preparation, bridging the gap between traditional methods and industry requirements.

In the subsequent sections, we first present a Literature Review of existing interview preparation methods and related

work. The Proposed Methodology section details our system's architecture, implementation technologies, and experimental results. Finally, the Conclusion summarizes our findings and suggests future research directions.

## II. Literature Survey

This literature survey examines four key technological components essential to modern AI-powered interview sys- tems. We analyze existing interview preparation platforms like Pramp and HireVue [27], [29], discussing their capabilities and limitations in scalability and real-time interaction [15]. We explore Speech-to-Text (STT) technologies, comparing lead- ing models like Whisper and AudioPaLM [6], [7], focusing on their accuracy and multilingual capabilities. The survey examines Text-to-Speech (TTS) systems, evaluating modern approaches in speech synthesis based on naturalness and real- time performance. Finally, we investigate Large Language Models (LLMs), particularly their applications in generating contextually relevant interview questions and responses [21], emphasizing their ability to maintain conversation coherence.

### A. *Interview Preparation Platforms*

The landscape of automated interview preparation platforms has evolved significantly, with several key players offering dis- tinct approaches. Our analysis examines the primary platforms and their limitations compared to our proposed system.

*1) Peer-to-Peer Platforms:* Pramp represents the peer-to- peer interview preparation model, connecting users for mutual practice sessions. While this approach offers real human interaction, its effectiveness heavily depends on peer quality and availability. Users often encounter scheduling conflicts and inconsistent interview quality [14]. In contrast, our AI- powered system provides consistent, on-demand interview simulations without scheduling constraints.

*2) Professional Interview Platforms:* Interviewing.io con- nects candidates with industry professionals for mock in- terviews. While offering authentic feedback, the platform's scalability is limited by human interviewer availability and high costs [14]. Our solution provides unlimited practice opportunities with consistent quality at a fraction of the cost.

### B. *Speech to Text*

Speech-to-text (STT) technologies have become fundamen- tal to automated systems. Several models have emerged, each offering distinct advantages.

Whisper supports 99 languages, making it one of the most versatile STT models for multilingual transcription tasks. Au- dioPaLM excels in multimodal tasks, handling high- and low- resource languages through its integration of text and audio, offering strong performance in automatic speech translation (AST) and cross-lingual tasks [7]. GigaSpeech 2 focuses on low-resource Southeast Asian languages, providing superior performance in these regions but lacking the broad multilin- gual adaptability of Whisper or AudioPaLM [1].

Whisper delivers high accuracy, achieving competitive rates across a wide variety of languages, including high- resourceand some low-resource languages [6]. AudioPaLM demon- strates slightly lower WER( Word Error Rate) in cross-lingual AST tasks, benefiting from its multimodal architecture [7]. GigaSpeech 2 outperforms both models in low-resource lan- guages, with a WER of 12.83% for

Vietnamese and 14.92% for Indonesian [1].

Whisper stands out in real-time applications, offering fast transcription with low latency, making it ideal for use cases such as live interviews and real-time transcription in web applications [6]. AudioPaLM suffers from high latency due to its larger model size, making it less suitable for real-time STT applications [7]. GigaSpeech 2 is better suited for offline tasks and does not perform well in live settings [1].

In terms of model size, Whisper is relatively large, with up to 1.5 billion parameters, striking a balance between accuracy and computational efficiency [6]. AudioPaLM is resource- intensive and less efficient for real-time use, making it better suited for research or specialized tasks [7]. GigaSpeech 2 is the most efficient in terms of size, with only 151.9 million parameters [1].

The comparative analysis shows that Whisper is the most versatile model for real-time, multilingual transcription tasks in diverse domains, providing high accuracy, adaptability, and low latency. AudioPaLM excels in multimodal and cross- lingual tasks but is less efficient in real-time STT applications due to its large size. GigaSpeech 2 is the superior model for low-resource languages, especially in Southeast Asia, but its specialization makes it less flexible for general-purpose or real-time use cases.A visual comparison of these models is presented in Figure 1.

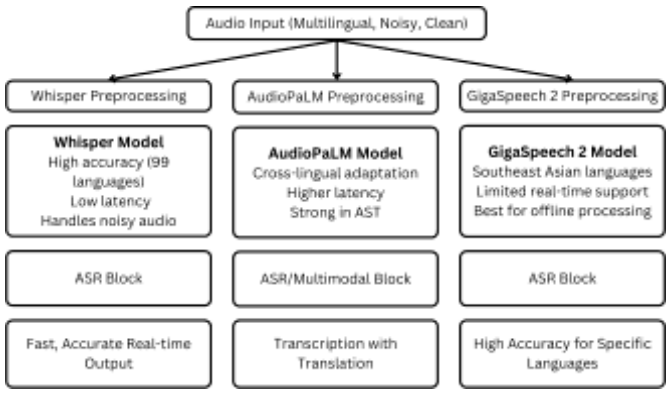


Fig. 1. *Comparison between different STT models*

### C. *Text to Speech*

Text-to-Speech (TTS) systems have made significant advancements in recent years, with numerous research contribu- tions pushing the boundaries of speech synthesis quality, effi- ciency, and multilingual support (see Table I). This literature survey provides a comprehensive analysis of prominent open- source TTS models, highlighting their respective strengths and challenges.

TABLE I
COMPARATIVE EVALUATION OF OPEN-SOURCE TEXT-TO-SPEECH
MODELS

| TTS Model | Speech Quality (MOS) | Inference Speed (RTF) | Multilingual Support | References |
|---|---|---|---|---|
| Coqui TTS | 4.1 | 0.06 | 10+ languages | [2], [11], [12], [8] |
| Mozilla TTS | 4.0 | 0.07 | 20+ languages | [2], [12], [8], [19] |
| ESPnet-TTS | 4.2 | 0.05 | 30+ languages | [12], [20], [19], [13] |
| OpenTTS | 3.8 | 0.08 | Limited | [11], [18], [9] |
| Fairseq S2T | 4.2 | 0.05 | 50+ languages | [18], [8], [13] |

Recent advancements in TTS research include the development of models capable of handling multiple languages with minimal supervision. XTTS(Extensible Text To Speech), a massively multilingual zero-shot TTS model, addresses the challenge of synthesizing speech in low-resource languages by leveraging a shared phonetic representation and training on a large-scale multilingual dataset [2]. Fairseq S2T has also integrated multilingual support, providing robust zero-shot TTS capabilities [2].

Long-form speech synthesis presents unique challenges, particularly in maintaining coherence and preventing misalign- ment over extended texts. The paper on Location-Relative Attention Mechanisms proposes a novel approach for robust alignment during long-form synthesis, effectively mitigating the issue of attention drift [12].

Glow-TTS introduces a flow-based generative model that simplifies the process of monotonic alignment search and im- proves the efficiency and quality of TTS models [11]. ESPnet- TTS(End-to-End Speech Processing Network) has

incorpo- rated such flow-based models to enhance its TTS capabilities [13].

Deep Voice 2 extends the capabilities of TTS systems to support multi-speaker synthesis using a single model [20]. This model leverages speaker embeddings to generate diverse voices, allowing for the synthesis of speech from multiple speakers with varying styles.

FastSpeech 2 builds on its predecessor by improving both the speed and quality of end-to-end TTS systems [8]. It intro- duces additional variance information, such as pitch, energy, and duration, to control the prosody of synthesized speech, making it both fast and high-quality.

High Fidelity Speech Synthesis with Adversarial Networks explores the use of GANs(Generative Adversarial Network) to improve the naturalness and quality of synthesized speech [19]. This approach has been adopted by several TTS systems, including ESPnet-TTS and Coqui TTS.

End-to-End Adversarial Text-to-Speech models combine the benefits of adversarial training with end-to-end learning frame- works, resulting in robust and high-quality TTS systems [9].This methodology allows for the direct training of models from text to waveform, reducing the complexity of intermediate representations.

### *D. Large Language Models*

Recent advancements in machine learning have significantly enhanced the capabilities of language models to generate con- textually relevant and adaptive questions, thereby improving the quality of interviews across various contexts. This literature survey delves into several critical aspects of Llama 3.1, Gemma 2, and GPT-4(Generative Pre-trained Transformer), emphasizing their architectures, contextual understanding, per- formance metrics, resource efficiency, and safety mechanisms. Llama 3.1 features a robust architecture with 70 billion parameters, enabling it to generate nuanced and contextu- ally relevant questions. Its large parameter count allows it to integrate information from previous responses effectively, facilitating a more engaging interview experience [21]. In contrast, Gemma 2, with 27 billion parameters, is designed for efficiency, which impacts its ability to adaptively generate questions during dynamic conversations [22]. GPT-4, boasting approximately 180 billion parameters, excels in generating complex and emotionally intelligent questions, but its larger size comes with increased computational demands [21].

The ability to understand context and generate adaptive questions is critical in interview simulations. Llama 3.1 excels in this area by effectively retaining context across long interac- tions, making interactions feel more natural and less scripted [21]. GPT-4 demonstrates superior contextual understanding, providing the ability to generate nuanced questions that require advanced reasoning [21]. Gemma 2 is primarily optimized for instruction-following tasks, limiting its effectiveness in generating adaptive questions [22].

In high-stakes interview situations, the choice of model can significantly impact the quality of the interaction. Llama 3.1 strikes a balance by offering adaptive questioning that meets various interview formats' needs without excessive computa- tional requirements [21]. GPT-4 shines in high-stakes scenarios due to its advanced capabilities, but its computational intensity can hinder scalability [21]. Gemma 2 performs adequately in simpler contexts but lacks the depth necessary for high-stakes interviews [22].

When evaluating resource efficiency, Llama 3.1 demon-strates a commendable balance between performance and computational requirements, allowing effective deployment in real-time applications [21]. Gemma 2 is the most resource- efficient model, but its efficiency comes at the cost of con- textual nuance and adaptability [22]. GPT-4 is significantly more resource-intensive, limiting its scalability in web-based interview applications [21].

Performance metrics across standard benchmarks are crit-ical for understanding a model's effectiveness in interview simulations. Llama 3.1 has shown strong results on various benchmarks such as MMLU(Mean Message Length Unit), achieving a score of 79.2, and GSM8K(Generalized Speech Model 8K), with a score of 76.9, highlighting its capability

in generating coherent and contextually relevant responses [21]. Gemma 2, while achieving respectable scores of 75.2 on MMLU and 74.0 on GSM8K, tends to fall short in more com- plex scenarios that require deeper contextual understanding [22]. In contrast, GPT-4 excels in performance metrics, scoring

86.4 on MMLU and 88.3 on GSM8K, demonstrating its ability to handle intricate questioning and reasoning tasks effectively. This superior performance makes GPT-4 particularly well- suited for high-stakes interview applications where nuanced understanding is paramount [21].

TABLE II
COMPARATIVE EVALUATION OF LARGE LANGUAGE MODELS FOR INTERVIEW SIMULATIONS

| Benchmark | Llama 3.1 | Gemma 2 | GPT-4 |
|---|---|---|---|
| Model Si (Parame- ters) | 70B | 27B | 180B |
| Context Length | 128K tokens | 8K tokens | 32K tokens |
| Adaptive Questioni | 8.5/10 | 7/10 | 9.5/10 |
| MMLU (5-shot) | 79.2 | 75.2 | 86.4 |
| GSM8K (8-shot) | 76.9 | 74.0 | 88.3 |
| Winogrande (0-shot | 85.3 | 83.7 | 90.2 |
| Human Preferenc (Chatbot) | 1206 Elo | 1218 Elo | 1286 Elo |
| Inference Speed | Moderate | Fast | Slow |
| Deployment Suitability | High | Medium | Low |

Safety and bias mitigation are essential factors in the deployment of AI models for interview simulations, particularly in sensitive contexts. Llama 3.1 incorporates advanced safety mechanisms, including the Llama Guard model, which enhances input and output safety, making it a reliable choice for professional interview settings [22]. These safety fea- tures ensure that generated responses remain respectful and appropriate, which is critical in interviews. Gemma 2 also employs rigorous data filtering techniques and pre-training data curation to minimize the risk of producing unsafe or biased outputs [22]. However, it may not match the nuanced safety capabilities of Llama 3.1 or GPT-4. GPT-4 leads in this area, incorporating sophisticated mechanisms for filtering and mitigating biased responses, enhancing its reliability for high- stakes interviews [21].

In summary, as illustrated in Table II, Llama 3.1 offers a balanced approach, combining robust contextual understanding with lower resource consumption, making it suitable for both dynamic and structured interview formats [21]. Gemma 2 is efficient but lacks depth for complex interactions, while GPT-
4 excels in reasoning and contextual generation but faces scalability limitations due to high computational demands [22]. The selection of the appropriate model will depend on specific requirements regarding complexity, resource availability, and the necessity for nuanced, adaptive questioning.

## PROPOSED METHODOLOGY

The proposed AI-powered mock interview system consists of two main components: Conducting Mock Interviews and Generating Scores and Feedback. Leveraging AI and server technologies, the system streamlines real-time interaction be- tween a candidate and the system, providing detailed, data- driven feedback based on multiple dimensions of performance analysis. Figures 2, 3, and **??** illustrate the overall architec- ture and data flow, facilitating understanding of the system's methodology

### *A. Conducting Mock Interviews*

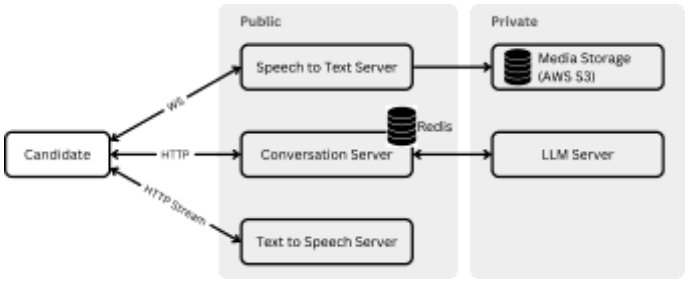


Fig. 2. *Proposed System Architecture for Conducting Mock Interviews*

As illustrated in the figure 2, the interview is facilitated using various AI components that handle real-time audio, video, and conversational data. This stage encompasses the following elements:

*1) Speech-to-Text (STT):* The candidate's voice is converted into text using Whisper model. The frontend communicates with the Speech-to-Text (STT) server via WebSocket (WS). This connection continuously streams the candidate's audio and video from the client-side to the server. The media storage system (e.g., AWS S3) simultaneously records the audio and video in real time for future analysis.

*2) Real-Time Transcription:* Once the audio is received, the Whisper model generates real-time transcriptions, which are essential for the AI system to understand and respond. These transcriptions are sent back to the frontend and passed on to the Conversation Server for generating relevant responses.

*3) Conversation Server with Redis Memory:* The transcrip- tions are sent via HTTP( Hypertext Transfer Protocol) requests to the Conversation Server, which is powered by LLaMA and Langchain. This server is designed to produce intelligent responses based on the context of the conversation.

To maintain continuity and coherence during the conversa- tion, Redis serves as the conversation memory. This ensures that the AI retains a memory of previous exchanges, providing more context-aware and fluid responses.

The Conversation Server then sends the generated response back to the frontend over an HTTP REST(Representational State Transfer) call.

*4) Text-to-Speech (TTS):* The system uses Coqui TTS to convert the AI's text-based response into spoken words. The Text-to-Speech (TTS) Server is connected through HTTP Stream, providing real-time audio output to the candidate, thus completing the conversational loop.

Through this pipeline, the candidate interacts with the system as if conversing with a human interviewer. The con- versation, along with the accompanying video and audio data, is stored on the backend for further analysis, as highlighted in the figure 2.

### *B. Generating Scores and Feedback*

Once the interview has concluded, the system transitions to the second stage: score generation and feedback. This part is explained with the help of the figures 3.

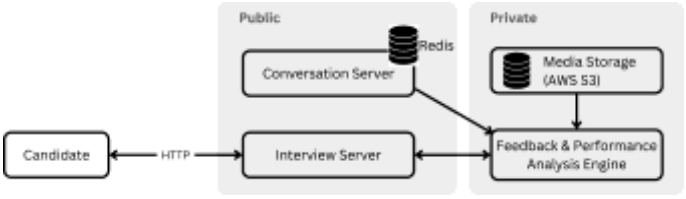


Fig. 3. *Proposed System Architecture for Score and Feedback Generation*

*1) Event-Driven Notification:* At the end of the interview, the system triggers the next phase by sending an event notifi- cation via RabbitMQ. This event informs the Score Generation Server that the interview data, including video, audio, and conversation transcripts, is now ready to be processed. The process begins as shown in the figure 3.

*2) Data Aggregation:* The Score Generation Server collects data from the following sources

- Media Storage (AWS S3): The server retrieves the recorded video and audio data stored during the interview.
- Redis (Conversation Memory): The conversation tran- scripts and any associated context retained by Redis are accessed.
- Sentiment and Emotional Analysis: Additional layers of analysis are performed using AI models that process both the video (frame-by-frame facial emotion recognition) and audio (intonation and emotional frequency analysis) to extract deeper insights into the candidate's emotional responses and communication patterns during the inter- view.

*3) Processing and Score Generation:* The system evaluates the candidate's performance across multiple dimensions.

- Speech Clarity and Conversational Coherence: The qual- ity of the candidate's responses is measured based on the fluency, clarity, and relevance of the conversation. This involves natural language processing (NLP) and coherence checks powered by LLaMA.

Once the system has processed all data, it generates a performance score that reflects the candidate's overall aptitude. This score is calculated by a custom scoring algorithm thattakes into account multiple parameters such as conversational engagement and accuracy of responses.

*4) Feedback Generation and Display:* After the scoring process, the system compiles detailed feedback based on the candidate's performance. This feedback includes:

- Strengths: Key areas where the candidate performed well, such as clarity of speech or confidence.
- Improvement Areas: Suggestions for how the candidate could improve, based on the conversational data.
- Overall Score: A numerical value that summarizes the candidate's performance, providing an easy-to-understand measure of their skills.

Finally, the feedback and score are sent back to the frontend, where the candidate can view the results. The system ensures that the feedback is constructive and

comprehensive, helping the candidate understand their performance in depth.

### C. System Computational Efficiency

Our system's real-time performance relies on the efficient operation of its core components, with particular focus on speech processing and conversation management. The system achieves high efficiency in speech processing through Azure- hosted Whisper for Speech-to-Text processing, which has an inference speed of 1 second, enables real-time transcription via WebSocket connections, processes audio in small chunks for immediate feedback, and optimizes network routing through Azure data centers. Additionally, Azure-hosted Coqui is used for Text-to-Speech synthesis, which also achieves an inference speed of 1 second, utilizes streaming synthesis for minimized latency, and employs HTTP Stream-based audio delivery. The conversation processing is powered by LLaMA 3.1, which achieves a processing speed of 250 tokens per second (T/s) to 1,660 T/s, representing a 6x improvement over standard implementations, and enables real-time response generation. The system also optimizes memory management through Redis, achieving an average storage per interview of 273 KB, a total storage for 15 interviews of 4 MB, and high- speed data retrieval. Overall, the system maintains a total processing pipeline of approximately 2–2.5 seconds, consist- ing of Speech-to-Text Processing (1 second), Text-to-Speech Synthesis (1 second), and additional processing overhead (0.5 seconds), ensuring natural conversation flow while maintaining high-quality speech processing, making the system suitable for real-time interview scenarios.

## III. CONCLUSION

In conclusion, this paper presents a comprehensive survey of existing AI-based mock interview platforms and identifies key challenges candidates face during interview preparation, such as limited personalized feedback, inadequate real-world simulation, and interview anxiety. To address these gaps, we propose an advanced AI-driven mock interview system that integrates Speech-to-Text, Text-to-Speech, and conversational AI models. Our methodology enables the platform to conduct realistic, interactive mock interviews that cover both technical and soft skills, providing users with immediate, detailed feed-back based on real-time analysis of their responses. Unlike traditional methods, our system offers adaptive, context-aware assessments, allowing candidates to identify specific areas for improvement and track their progress over time. By leveraging AI technologies, the platform delivers a scalable, unbiased, and personalized experience, making effective interview prepara- tion accessible to a diverse range of users. The proposed solu- tion effectively bridges the gap between conventional tools and modern interview demands, empowering candidates to build confidence, improve their skills, and succeed in competitive job markets.

## REFERENCES

[1] Y. Yang, Z. Song, J. Zhuo, M. Cui, J. Li, B. Yang, Y. Du, Z. Ma, X. Liu,
Z. Wang, K. Li, S. Fan, K. Yu, W.-Q. Zhang, G. Chen, and X. Chen, *GigaSpeech 2: An Evolving, Large-Scale and Multi-domain ASR Corpus for Low-Resource Languages with Automated Crawling, Transcription and Refinement*, arXiv, 2024.

[2] E. Casanova, K. Davis, E. Go¨lge, G. Go¨knar, I. Gulea, L. Hart, A. Aljafari,
J. Meyer, R. Morais, S. Olayemi, and J. Weber, *XTTS: A Massively Multilingual Zero-Shot Text-to-Speech Model*, arXiv, 2024.

[3] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman,
A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, *The Llama 3 Herd of Models*, arXiv preprint arXiv:2407.21783, 2024.

[4] Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupati- raju, L. Hussenot, T. Mesnard, B. Shahriari, A. Rame´, J. Ferret, P. Liu,
P. Tafti, A. Friesen, M. Casbon, S. Ramos, *Gemma 2: Improving Open Language Models at a Practical Size*, arXiv preprint arXiv:2408.00118, 2024.

[5] Babashahi, Leili and Barbosa, Carlos Eduardo and Lima, Yuri and Lyra, Alan and Salazar, Herbert and Argoˆlo, Matheus and Almeida, Marcos Antonio de and Souza, Jano Moreira de, *AI in the Workplace: A Systematic Review of Skill Transformation in the Industry*, Administrative Sciences, 2024.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, *Robust Speech Recognition via Large-Scale Weak Supervision*, Proceedings of the

40th International Conference on Machine Learning, vol. 202, pp. 28492–28518, July 2023.

[7] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. El Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. Tadmor Ramanovich, M. Tagliasacchi,
A. Tudor, M. Velimirovic´, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, *AudioPaLM: A Large Language Model That Can Speak and Listen*, arXiv, 2023.

[8] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*, arXiv, 2022.

[9] J. Donahue, S. Dieleman, M. Bin´kowski, E. Elsen, and K. Simonyan,
*End-to-End Adversarial Text-to-Speech*, arXiv, 2021.

[10] Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, Shinji Watanabe, *ESPnet2-TTS: Extending the Edge of TTS Research*, arXiv, 2021.

[11] J. Kim, S. Kim, J. Kong, and S. Yoon, *Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search*, Advances in Neural Information Processing Systems, vol. 33, pp. 8067–8077, 2020.

[12] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao,
M. Shannon, and T. Bagby, *Location-Relative Attention Mechanisms for Robust Long-Form Speech Synthesis*, ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 6194–6198.

Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, Shinji Watanabe, *ESPnet2-TTS: Extending the Edge of TTS Research*, arXiv, 2021.

[13] Pankaj Rambhau Patil, Shinde Rushikesh Rajendra, Gosavi Vishakha Mahendra, Bhamare Bhagyashri Jijabrao, Patil Paresh Dilip, *Elevating Performance Through AI-Driven Mock Interviews*, IJRASET, 2020.

[14] B. C. Lee, B. Y. Kim, *Development of an AI-Based Interview System for Remote Hiring*, International Journal of Advanced Re- search in Engineering and Technology 12 (3) (2021) 654–663. doi:10.34218/IJARET.12.3.2021.060.

[15] B. C. Lee, B. Y. Kim, *A Decision-Making Model for Adopting an AI-Generated Recruitment Interview System*, International Journal of Management 12 (4) (2021) 548–560. doi:10.34218/IJM.12.4.2021.046.

[16] H. K. Fulk, H. L. Dent, W. A. Kapakos, B. J. White, *Doing More with Less: Using AI-based Big Interview to Combine Exam Preparation and Interview Practice*, Issues in Information Systems 23 (4) (2022) 204–217. doi:10.48009/4 iis–2022 118.

[17] Y. Liu and J. Zheng, *Es-Tacotron2: Multi-Task Tacotron 2 with Pre- Trained Estimated Network for Reducing the Over-Smoothness Problem*, Information, vol. 10, no. 4, article 131, 2019.

[18] M. Bin´kowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, *High Fidelity Speech Synthe- sis with Adversarial Networks*, arXiv, 2019.

[19] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, *Deep Voice 2: Multi-Speaker Neural Text-to- Speech*, arXiv, 2017.

[20] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman,
A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, *The Llama 3 Herd of Models*, arXiv preprint arXiv:2407.21783, 2024.

[21] Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupati- raju, L. Hussenot, T. Mesnard, B. Shahriari, A. Rame´, J. Ferret, P. Liu,
P. Tafti, A. Friesen, M. Casbon, S. Ramos, *Gemma 2: Improving Open Language Models at a Practical Size*, arXiv preprint arXiv:2408.00118, 2024.

[22] S. R. Livingstone, F. A. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English*, PLOS ONE 13 (5) (2018) 1–35. doi:10.1371/journal.pone.0196391.

[23] F. Eyben, M. Wo¨llmer, B. Schuller, *OpenSmile - The Munich Versatile and Fast Open-Source Audio Feature Extractor*, MM'10 - Proceedings of the ACM Multimedia 2010 International Conference (2010) 1459–1462. doi:10.1145/1873951.1874246.

[24] P. Boersma, D. Weenink, *Praat, A System for Doing Phonetics by Computer*, Glot International 5 (2001) 341–345.

[25] A. Bhavan, P. Chauhan, Hitkul, R. R. Shah, *Bagged Support Vector Ma- chines for Emotion Recognition from Speech*, Knowledge-Based Systems 184 (2019) 104886. doi:https://doi.org/10.1016/j.knosys.2019.104886.

[26] Pramp, *Practice Makes Perfect - Technical Interview Practice*, https: //www.pramp.com/

[27] interviewing.io, *Practice interviews with engineers from Google, Face- book, and more*, https://interviewing.io/

[28] HireVue, *Video Interviewing & Assessment Software*, https://www. hirevue.com/