

InSightHire

Revolutionizing Candidate Assessment: A Comprehensive AI-Driven System for Video Interviews, Emotion Analysis, and Content Evaluation

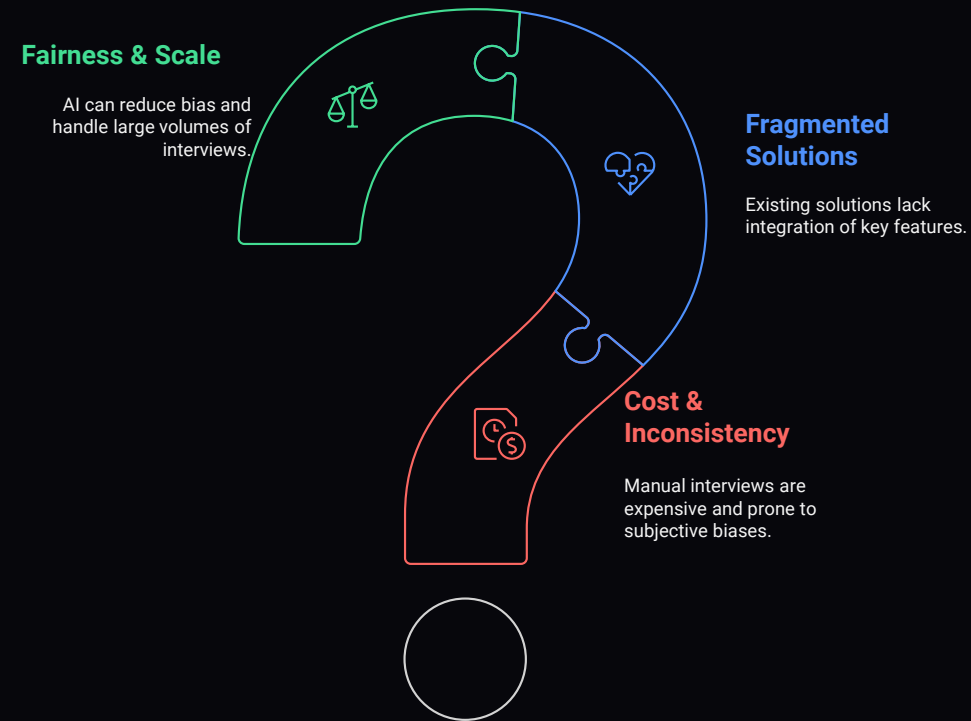
Name: Haneen Ahmed Qandil

ID: 220456

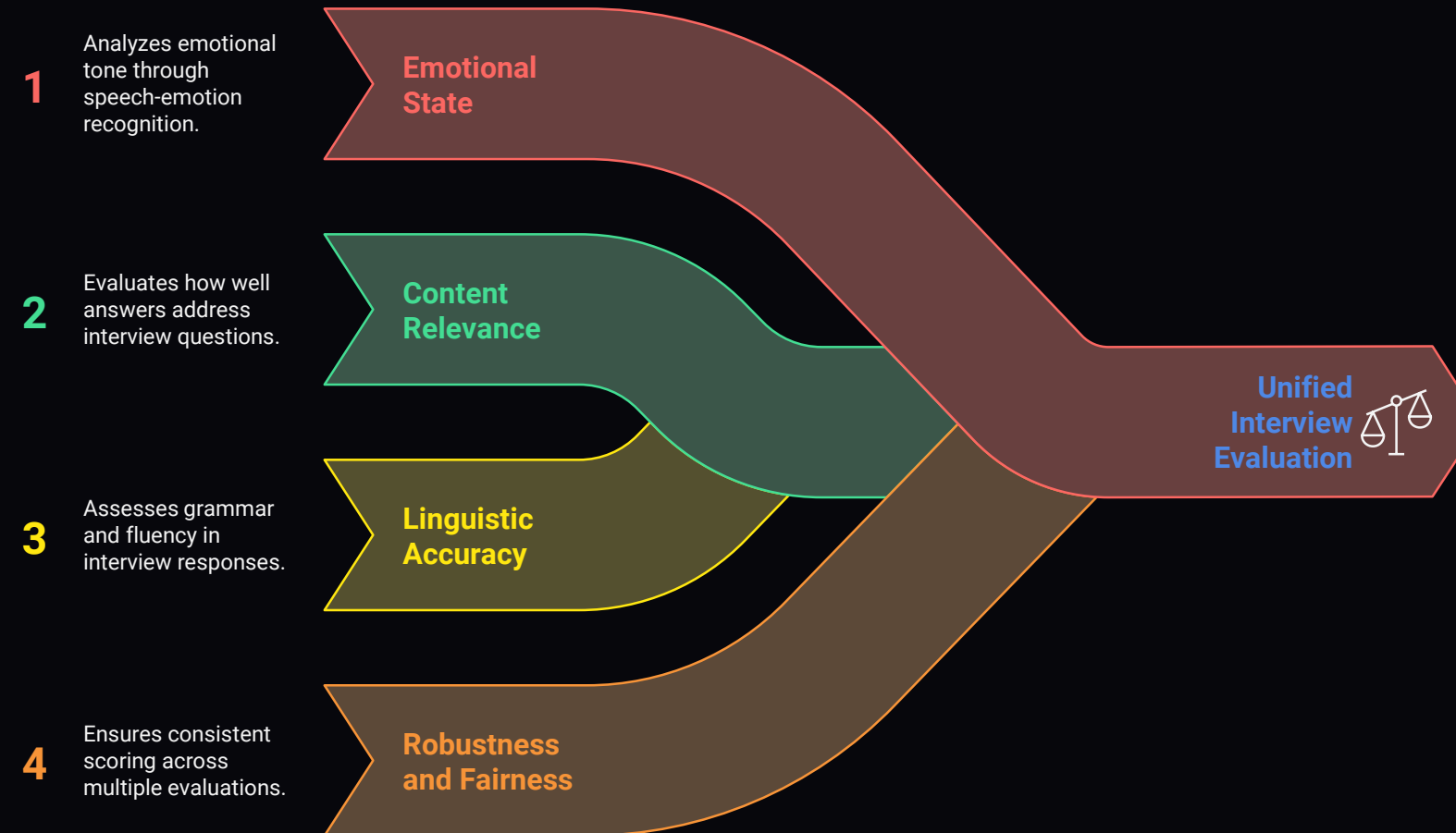
Supervisor: Prof. Andreas Pester

Major: AI

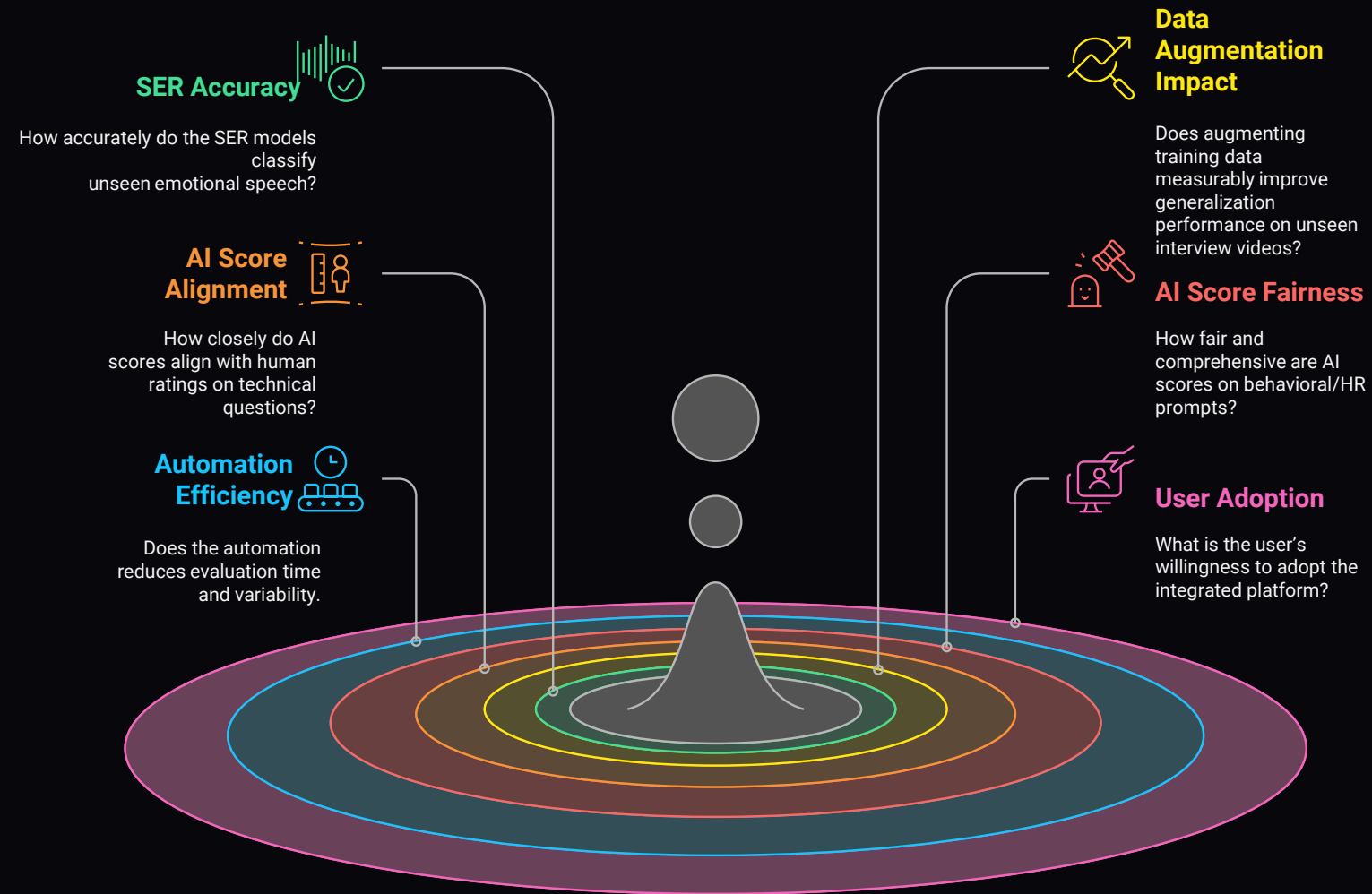
Why build an end-to-end AI interview tool?



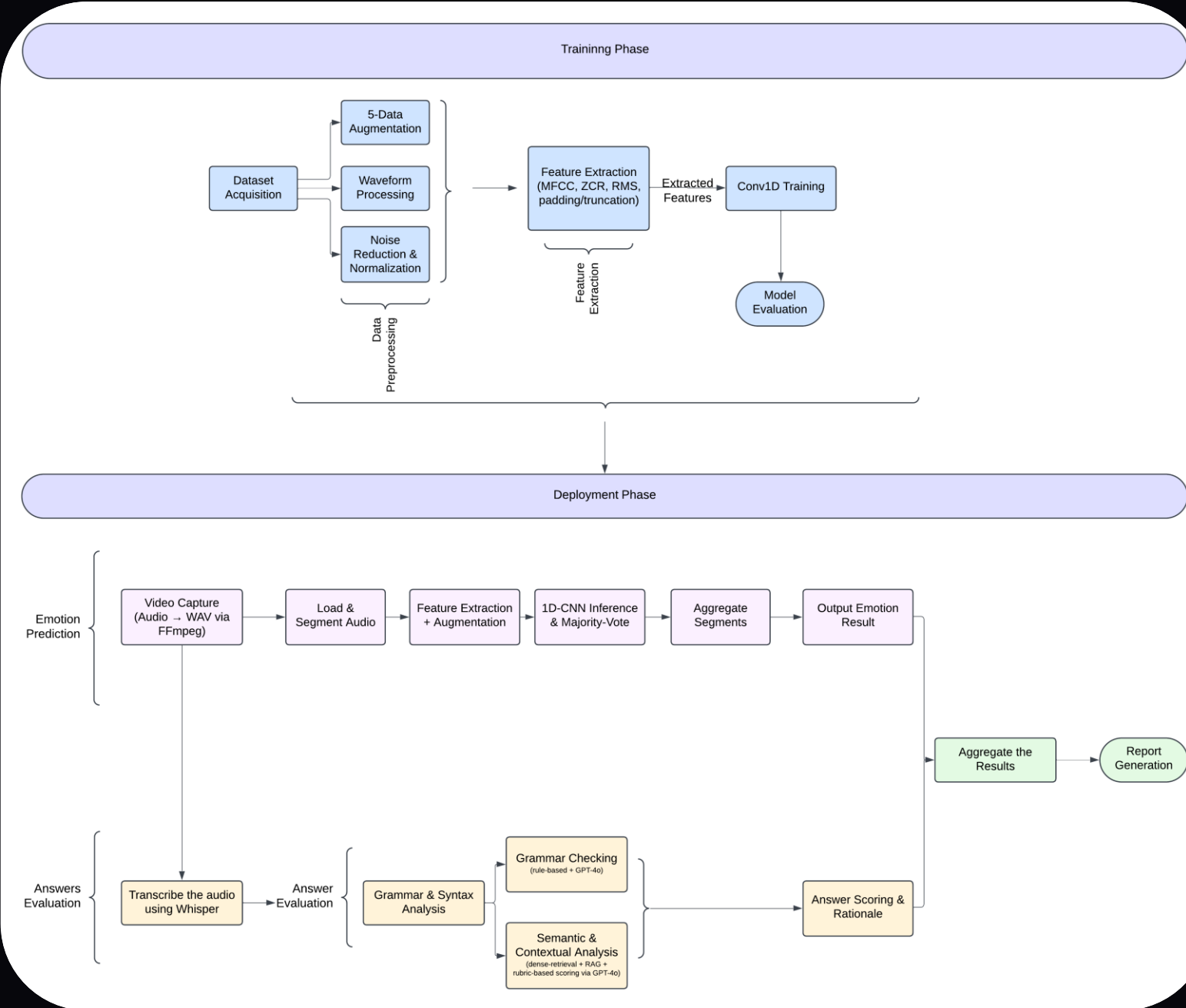
Problem Statement



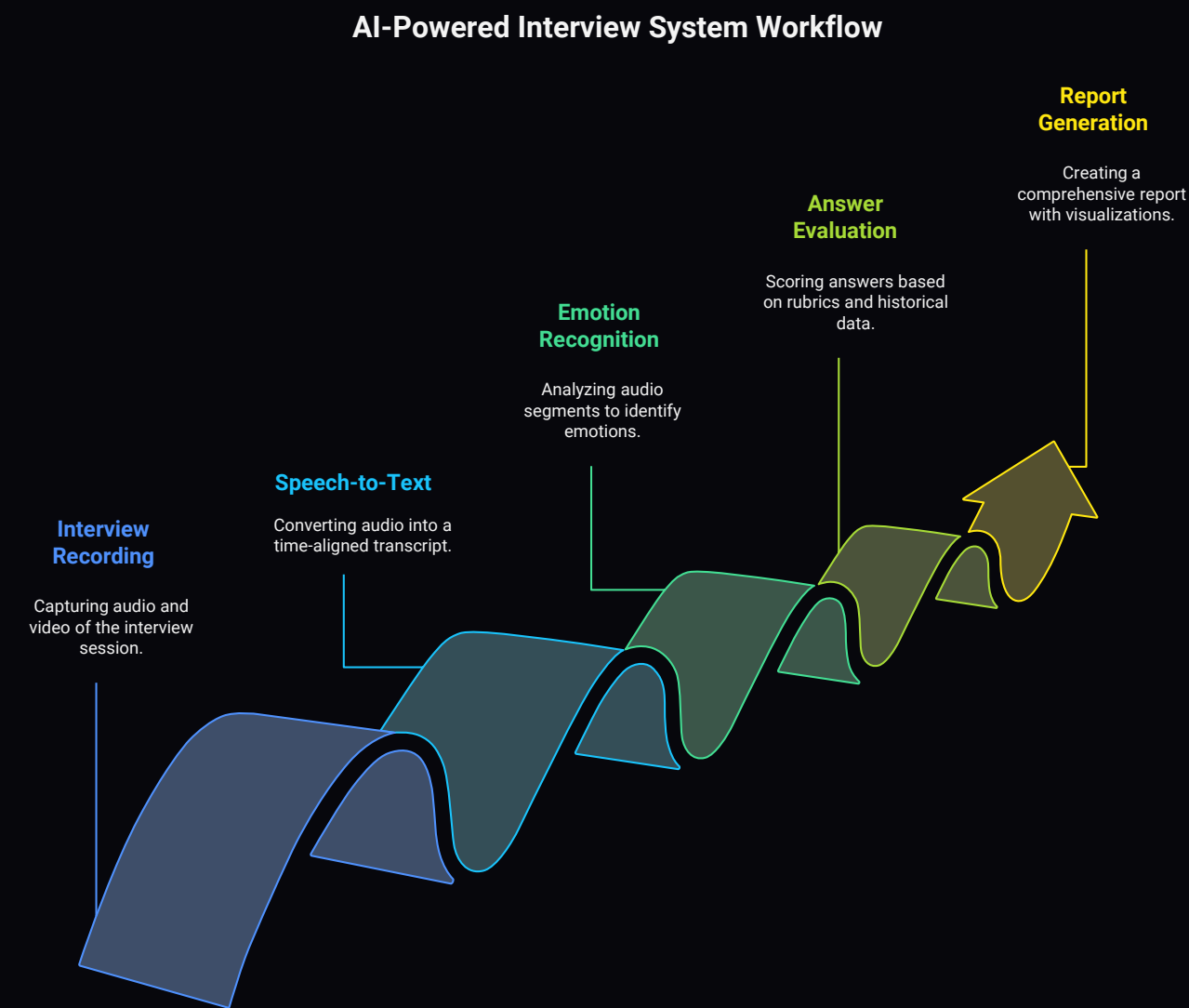
Research Questions



Methodology: End-to-End Pipeline

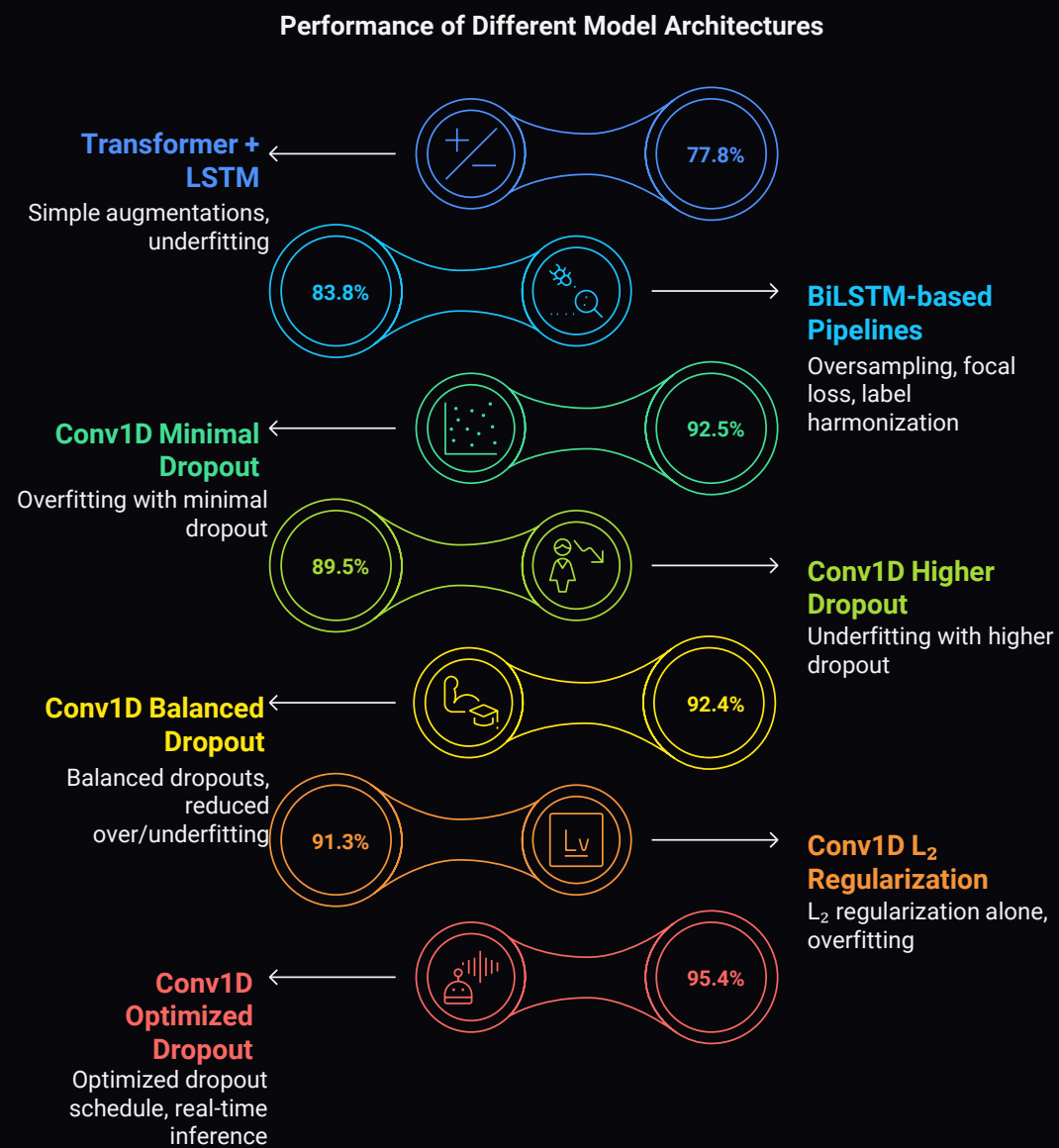


Methodology: End-to-End Pipeline

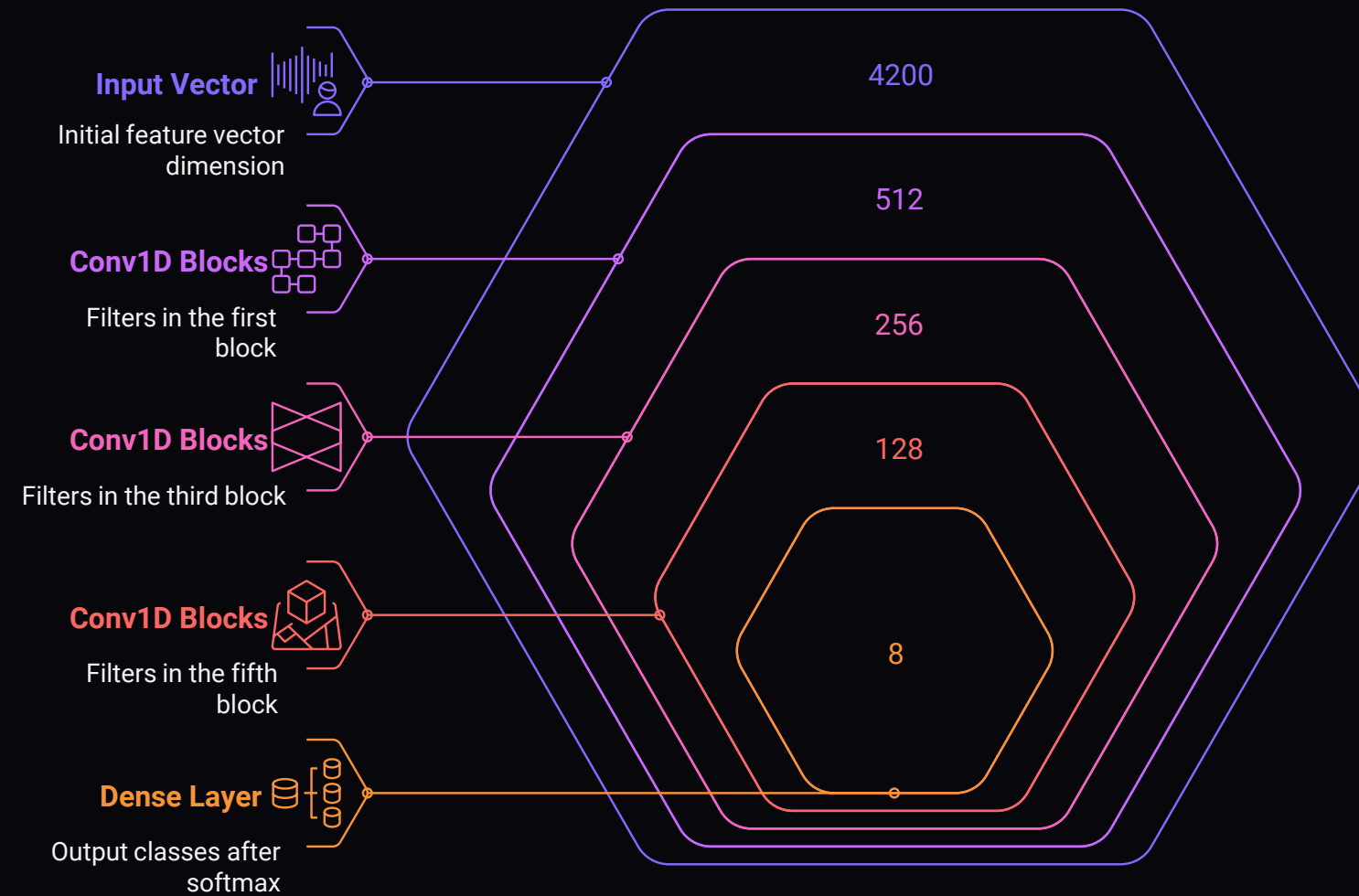


Speech Emotion Recognition

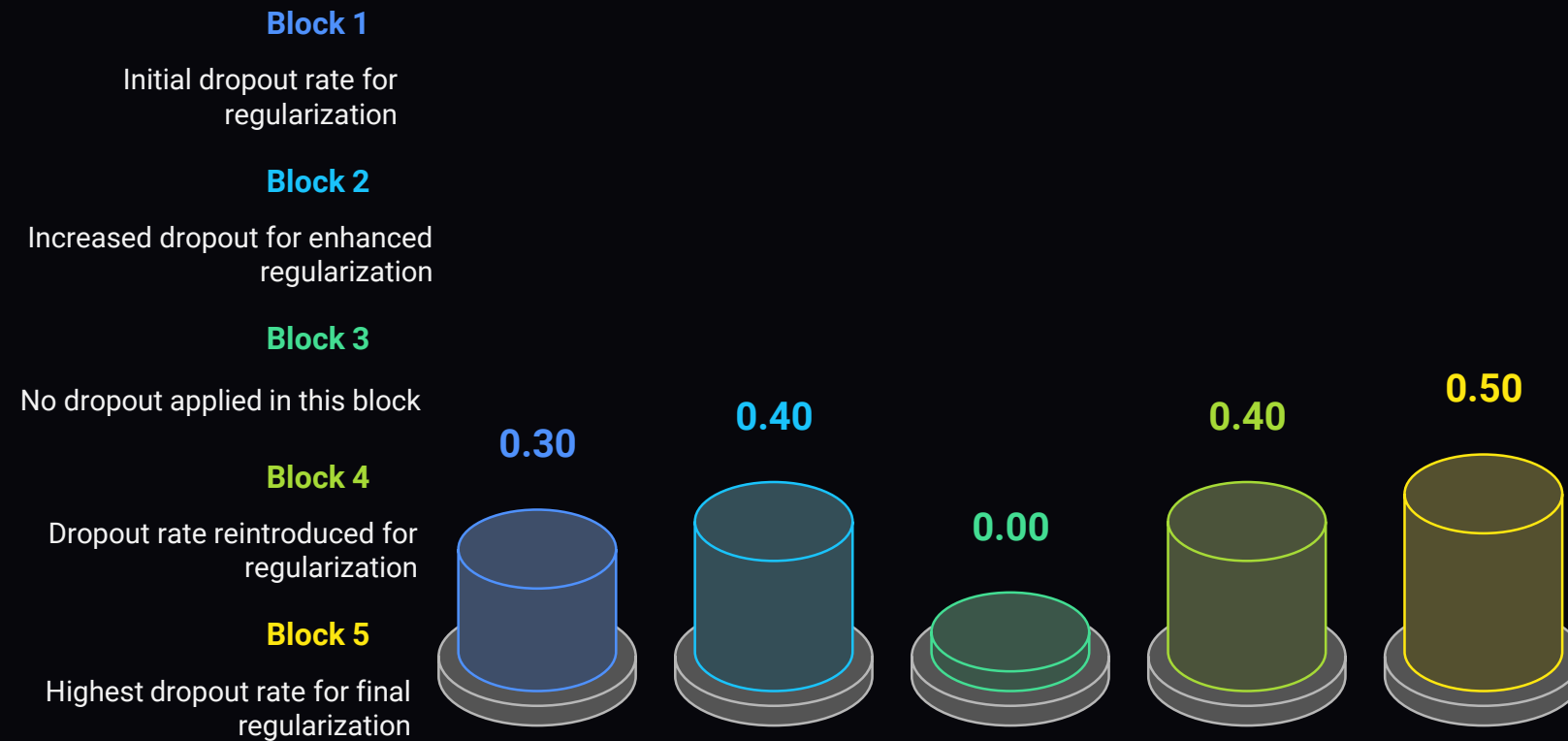
SER Model Selection Trials



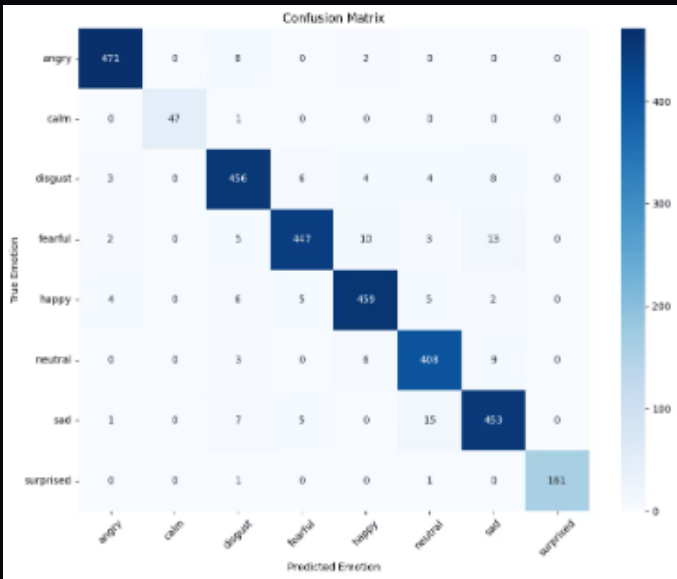
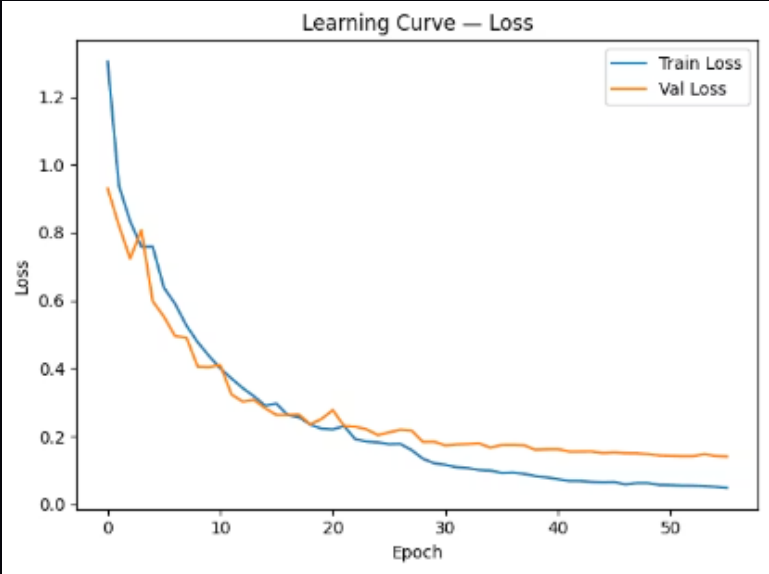
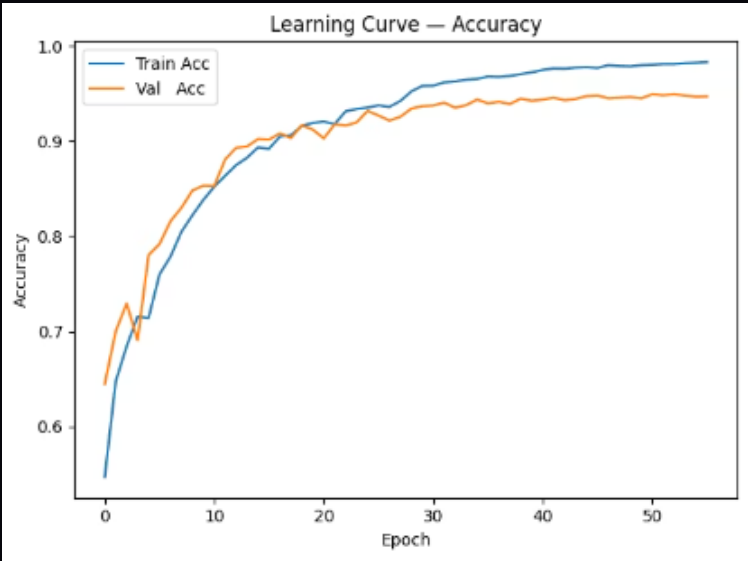
Audio Feature Vector Transformation



Dropout Rates in Conv1 D Blocks



SER Performance: Accuracy & Confusion Matrix



Train Accuracy:	99.98 % → near-perfect fit
Validation Accuracy:	94.93 %
Test Accuracy:	95.43 %
Key Finding:	Minimal gap between val/test confirms that overfitting has been effectively solved through progressive dropout, batch normalization, and delayed early-stopping

Overall Metrics

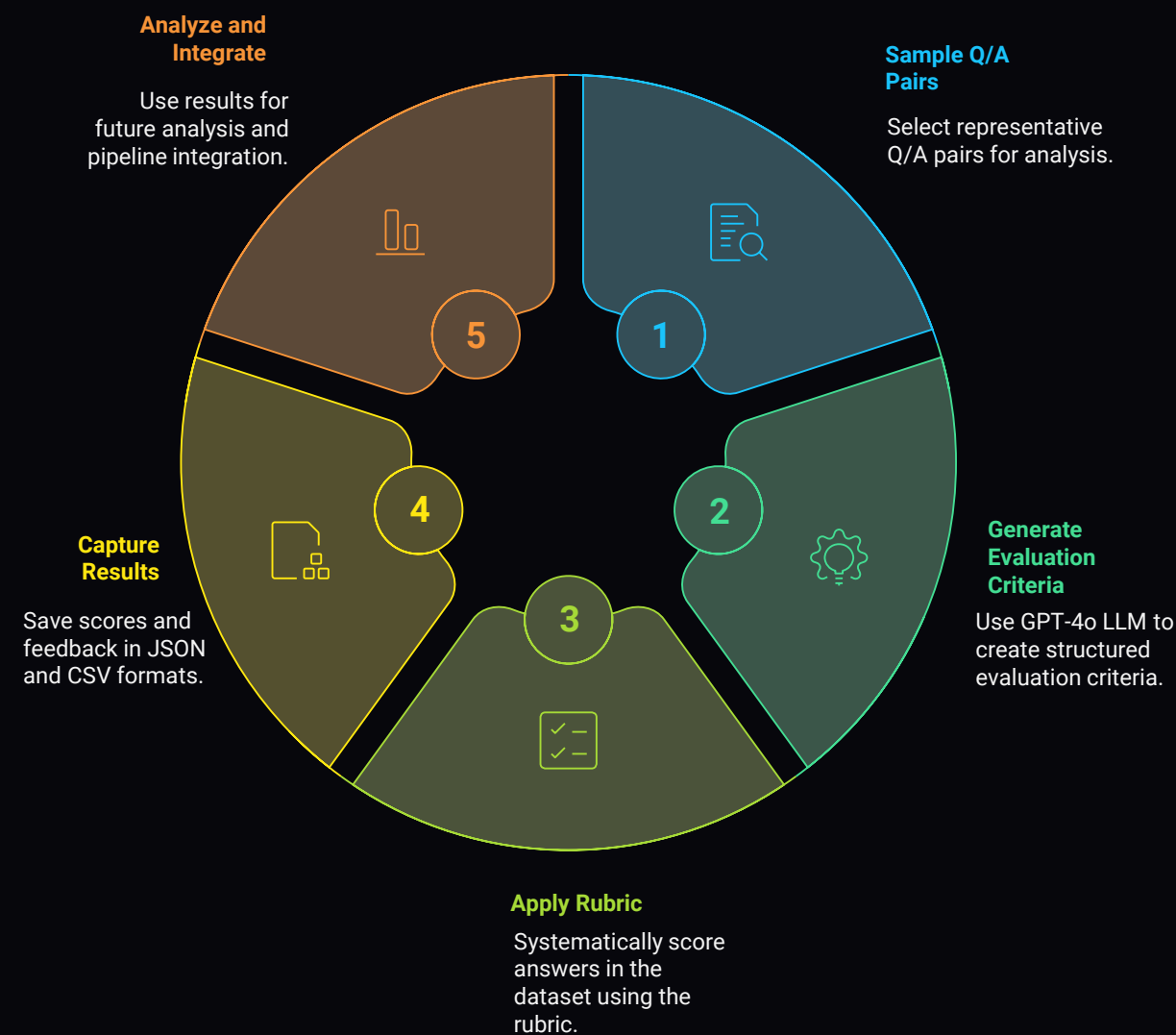
Macro-averaged F ₁ :	0.96
Overfitting addressed:	Regularization and callbacks ensured robust generalization
Supports	reliable emotion classification in AI-driven interview evaluation pipelines

Candidate Answer Evaluation

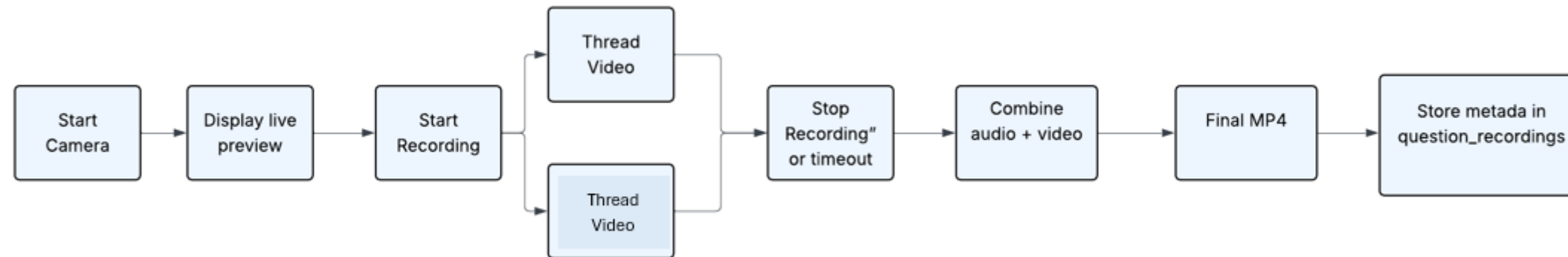
HR & Technical Datasets Preprocessing

Characteristic	Technical	HR
 Initial Steps	ASCII normalization, whitespace collapsing	ASCII normalization, whitespace collapsing
 Additional Step	None	Instruction-to-concrete- answer transformation using GPT-4o
 Empty Answer Handling	Filtered out empty/placeholder answers	Flagged instructional answers for transformation
 Markup Removal	HTML tags, Markdown artifacts, list bullets excised	HTML tags, Markdown artifacts, list bullets excised
 Final Output	qa_preprocessed.csv	interview_best_answers_cleaned.csv

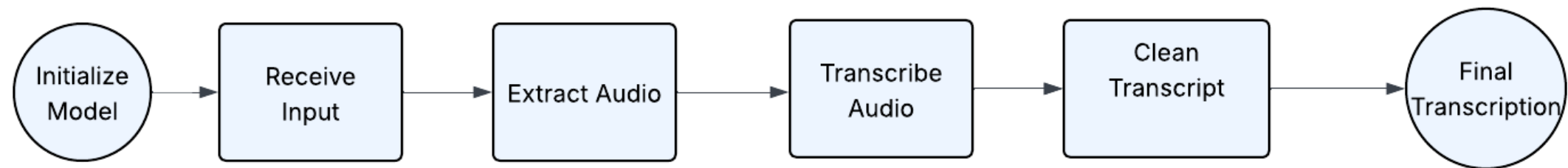
Rubric-Based Generation Cycle





Audio/Video Recording Workflow







Transcription with Whisper



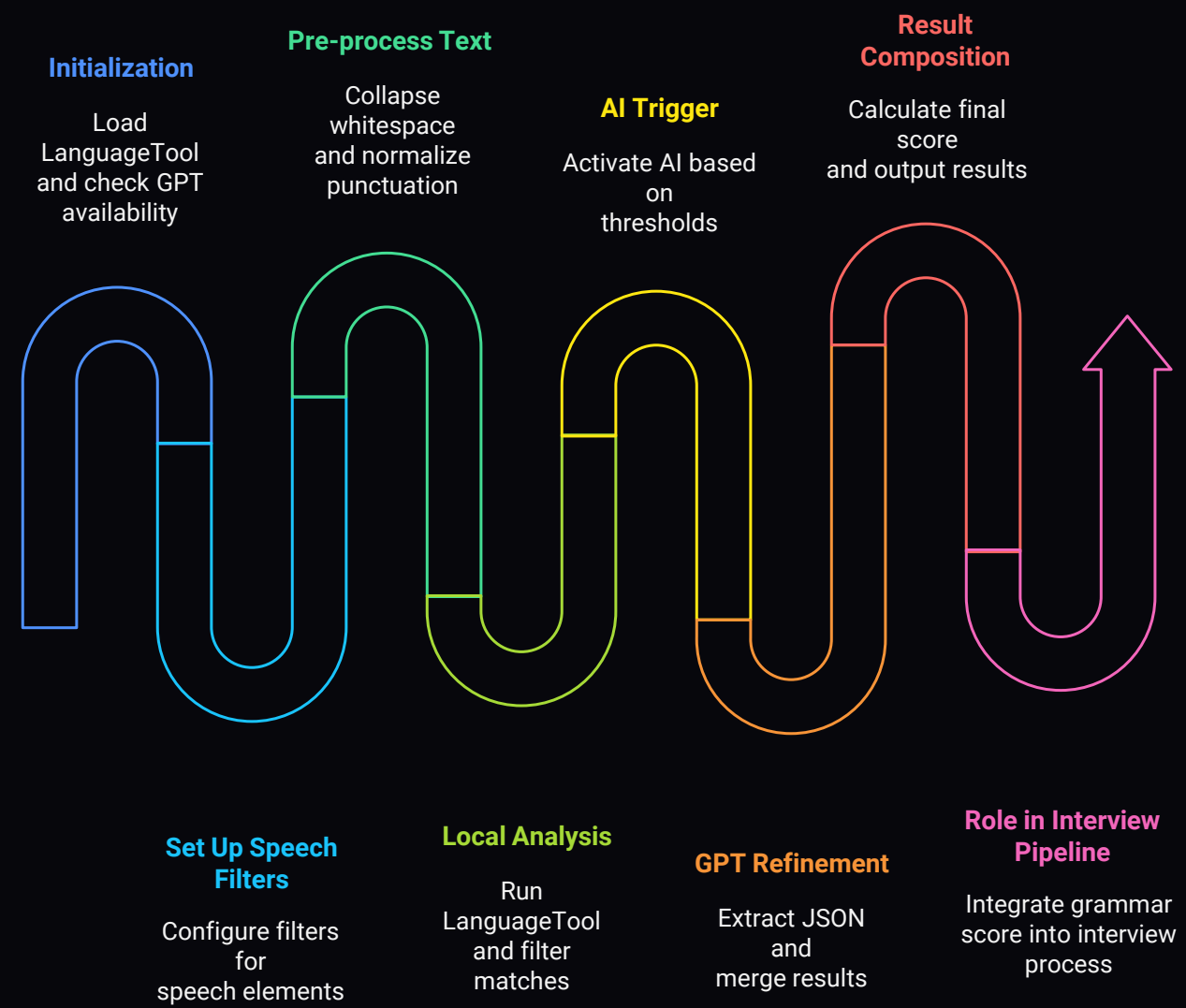
Score Integration Based on Retrieval Outcome

Outcome	Exact Match	Relevant Neighbors	No Match/Relevance
 Old Score	Retrieve matched question's old score	Compute average old score	N/A
 Combination Rule	If old score > 70: 70% old + 30% fresh; else: 100% fresh	30% average old + 70% fresh	100% fresh

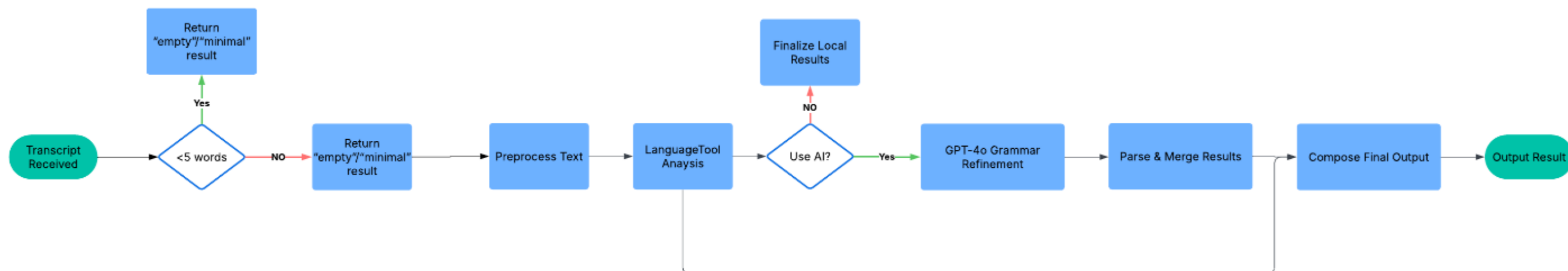
Enhancing GPT-4o Evaluation by RAG

Characteristic	RAG Module	GPT-4o Evaluation
 Data Source	Real, high-quality exemplar answers	Model's latent space
 Hallucination Risk	Sharply reduces	Higher risk
 Judgment Alignment	Aligns with domain-specific gold standards	May deviate from standards
 Variance in Pilot Audits	Up to 15-point lower	Higher

Grammar Evaluation Process

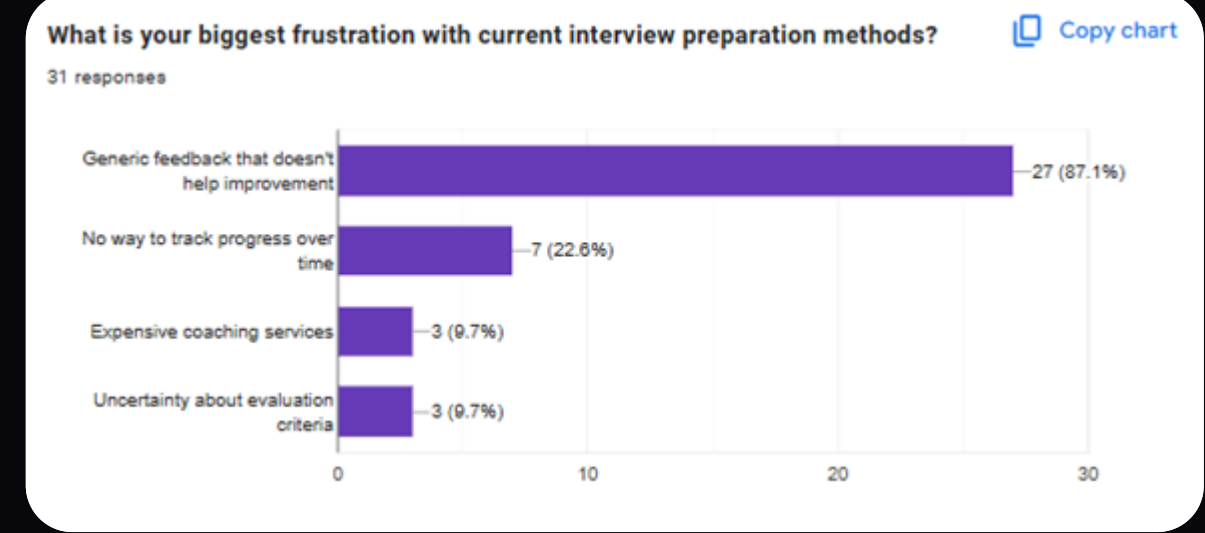
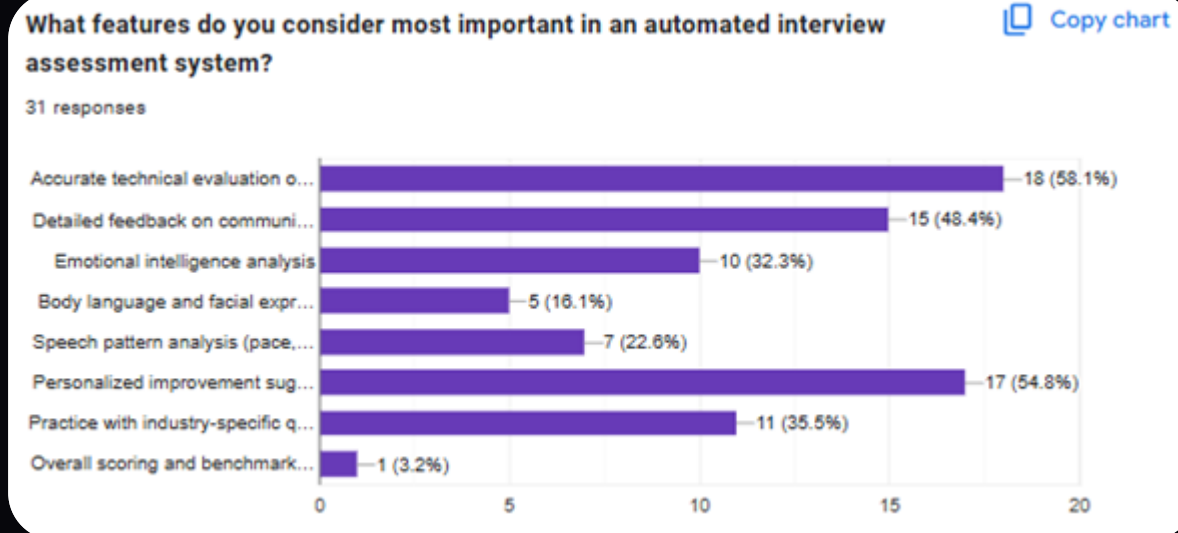


Grammar Evaluation Process



Human Evaluation Results

Participant Expectations for an Automated Interview System

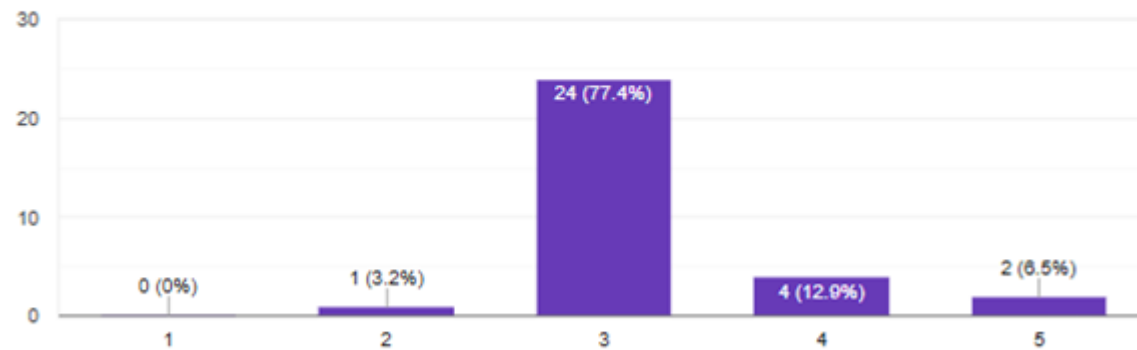


Technical Interview Question Evaluation

How fair is the AI's Technical scoring?

 [Copy chart](#)

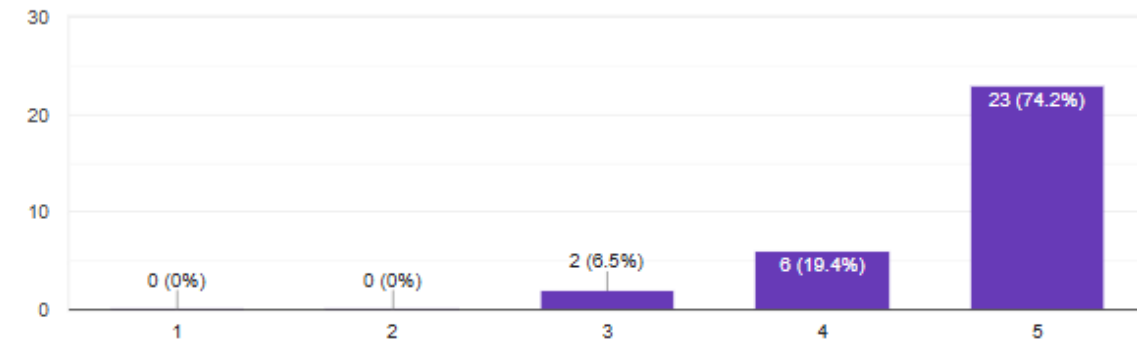
31 responses



How accurate is the AI's evaluation of the technical correctness?

 [Copy chart](#)

31 responses

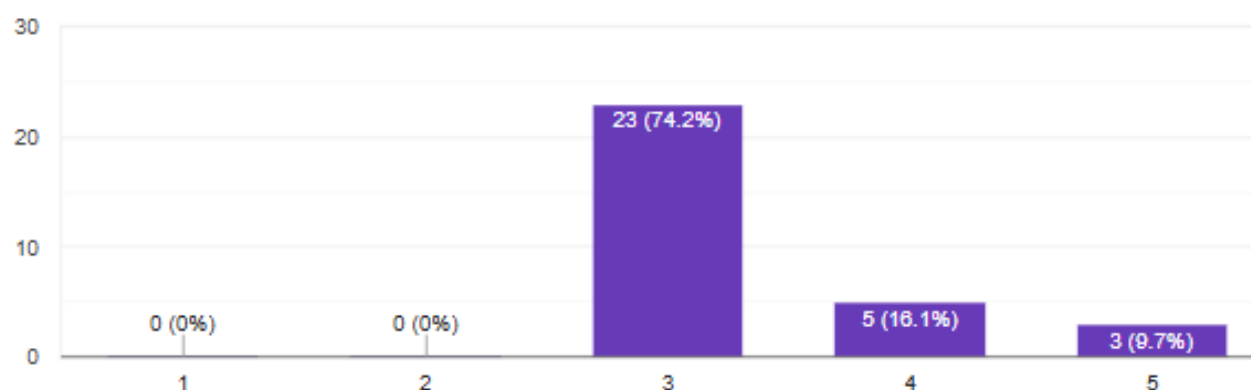


HR Interview Question Evaluation

How fair is the AI's HR scoring?

 [Copy chart](#)

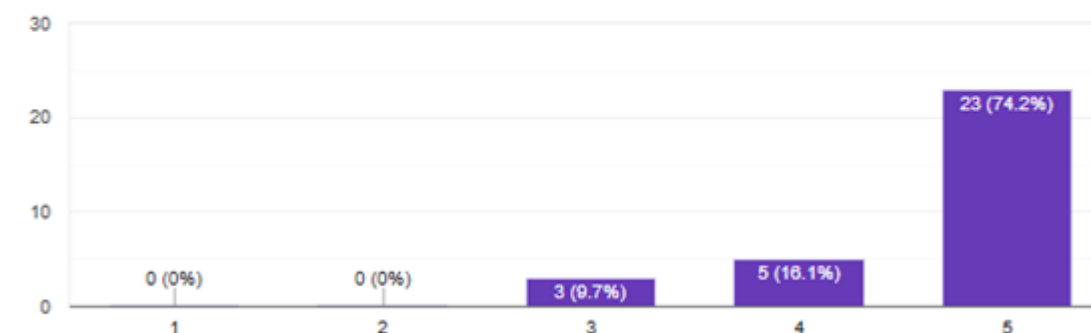
31 responses



How accurate is the AI's evaluation of the HR correctness?

 [Copy chart](#)

31 responses

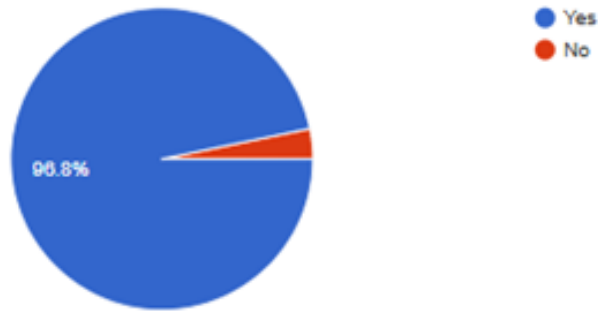


Overall Assessment of LLM Evaluation Quality

Would you use such systems to practice for interview?

31 responses

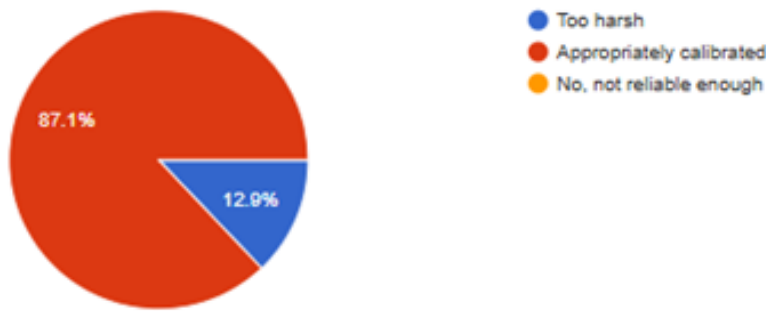
 Copy chart



Were the evaluations...

31 responses

 Copy chart



Ethical Concerns

Ethical Concerns

Key Concerns

- **Privacy & Consent**
 - Audio/transcripts = personal data → explicit opt-in & clear usage disclosures
- **Bias & Fairness**
 - Accent/demographic gaps → risk of unequal performance
- **Transparency**
 - Opaque LLM decisions → need for explainable scores

Planned Mitigations

- **Privacy by Design**
 - End-to-end encryption, strict ACLs, auto-delete policies
- **Bias Monitoring**
 - Regular subgroup audits, diverse data augmentation, debiasing methods
- **Explainable Scoring**
 - Publish rubrics & weights, per-criterion rationales, confidence intervals, human-in-the-loop overrides
- **Governance & Compliance**
 - GDPR/CCPA alignment, Ethics Board oversight, continuous policy updates

Conclusion

Critical Analysis & Literature Comparison

SER Model Performance

- Our Conv1D Network

95.4 % accuracy (macro-F1 \approx 0.96) on four heterogeneous English corpora

Light augmentations + optimized dropout (0.3 \rightarrow 0.5) solved overfitting and boosted “calm” recall to 98 %

~5 M parameters, < 10 ms inference per 2.5 s window

Key Advantages

- Joint training on RAVDESS, CREMA-D, TESS, SAVEE
- MFCC + ZCR + RMS feature stack with deep 1D convolutions
- Demonstrated robustness across diverse accents, recording conditions

Compared to Prior Work

Liu et al. [1] & Wani et al. [2]: reported 77–86 % on similar English datasets

Issa et al. [3]: 71.6 % (RAVDESS), 86.1 % (EMO-DB), 95.7 % (trimmed EMO-DB)

Limitations: small, acted German speech \rightarrow poor real-world generalization

Answer-Evaluation Pipeline

Our Retrieval-Augmented Approach

Custom rubrics + FAISS anchors + fresh GPT-4o scoring

93 % technical & 90 % HR score agreement with human raters

96.8 % user adoption willingness

Compared to Static Rubric Methods

Traditional fixed rubrics lack adaptability to varied Q&A types

Our dynamic retrieval blending historical & fresh scores yields higher alignment and user trust



Conclusion: Research Questions Revisitation

1. **SER Effectiveness**The final 1-D Conv consistently classifies unseen emotional speech with high accuracy, and optimized dropouts are shown to significantly improve generalization, especially on under-represented classes.
2. **LLM-Driven Answer Evaluation**AI-generated rubric scores closely mirror expert human ratings on both technical and behavioral prompts, confirming that the retrieval-augmented, multi-pass evaluation reliably captures content correctness and nuance.
3. **Overall System Acceptance**Integrating SER and LLM-based scoring into a unified pipeline substantially streamlines the interview review process—reducing evaluation effort and inter-rater variability—and most users report they would adopt the platform for both practice and formal screening.

References

- [1] Z.-T. Liu, M.-T. Han, B.-H. Wu, and A. Rehman, “Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning,” *Applied Acoustics*, vol. 202, pp. 109178, Dec. 2022. [Online]. Available: <https://doi.org/10.1016/j.apacoust.2022.109178>.
- [2] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, “A comprehensive review of speech emotion recognition systems,” *IEEE Access*, vol. 9, pp. 47795–47813, 2021. [Online]. Available: <https://doi.org/10.1109/ACCESS.2021.3068045>.
- [3] D. Issa, M. F. Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomedical Signal Processing and Control*, vol. 59, 2020, pp. 101894. [Online]. Available: <https://doi.org/10.1016/j.bspc.2020.101894>.

Thank You !