



# Can Quantized Audio Language Models Perform Zero-Shot Spoofing Detection?

Bikash Dutta\*, Rishabh Ranjan\*, Shyam Sathvik, Mayank Vatsa, Richa Singh

<sup>1</sup>Indian Institute of Technology Jodhpur, India

{d22cs051, ranjan.4, b22ee036, mvatsa, richa}@iitj.ac.in

## Abstract

Quantization is essential for deploying large audio language models (LALMs) efficiently in resource-constrained environments. However, its impact on complex tasks, such as zero-shot audio spoofing detection, remains underexplored. This study evaluates the zero-shot capabilities of five LALMs, GAMA, LTU-AS, MERaLiON, Qwen-Audio, and SALMONN, across three distinct datasets: ASVspoof2019, In-the-Wild, and Wave-Fake, and investigates their robustness to quantization (FP32, FP16, INT8). Despite high initial spoof detection accuracy, our analysis demonstrates severe predictive biases toward spoof classification across all models, rendering their practical performance equivalent to random classification. Interestingly, quantization to FP16 precision resulted in negligible performance degradation compared to FP32, effectively halving memory and computational requirements without materially impacting accuracy. However, INT8 quantization intensified model biases, significantly degrading balanced accuracy. These findings highlight critical architectural limitations and emphasize FP16 quantization as an optimal trade-off, providing guidelines for practical deployment and future model refinement.

**Index Terms:** large audio language models, quantization, audio spoofing.

## 1. Introduction

Large Audio Language Models (LALMs) [1] have significantly advanced audio processing tasks such as speech recognition, sound event detection, and audio captioning [2, 3]. These models, trained on extensive multimodal datasets, frequently achieve near-human proficiency, demonstrating impressive capabilities in complex audio understanding and generation. A key feature of LALMs is their strong zero-shot learning ability [4, 5, 6], enabling them to effectively generalize to new and unseen tasks without explicit task-specific training. This capability is particularly valuable in domains with limited or rapidly evolving labeled data, such as audio spoofing detection.

Despite their powerful capabilities, the practical deployment of LALMs is challenging due to substantial computational resource requirements, especially on devices with limited memory and processing power. To address these challenges, quantization has emerged as the essential optimization strategy. Quantization involves reducing the numerical precision of model parameters from higher precision formats, such as 32-bit floating-point, to lower precision formats, such as 16-bit floating-point (FP16) or 8-bit integers (INT8) [7]. This precision reduction significantly decreases memory usage and boosts inference speeds, making it feasible to deploy sophisticated

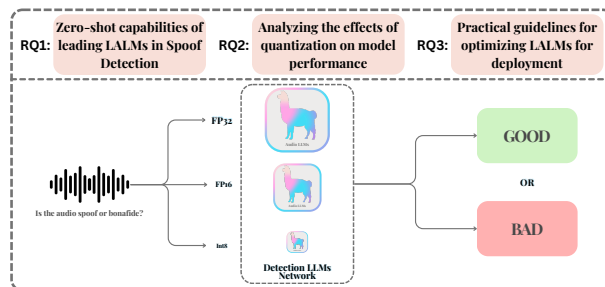


Figure 1: Overview of our evaluation framework addressing three research questions (RQs). RQ1 evaluates the zero-shot spoof detection capabilities of prominent LALMs. RQ2 investigates the impact of quantization precision (FP32, FP16, INT8) on model performance and computational efficiency. RQ3 offers practical guidelines for optimizing LALMs for real-world deployment, categorizing model performance as either 'GOOD' or 'BAD' based on robustness and efficiency.

audio-language models in real-time and resource-constrained environments. However, quantization is not without trade-offs, and its impact on the performance of audio tasks such as zero-shot spoof detection, remains largely unexplored.

Audio spoofing detection is an important domain where the potential of LALMs and the challenges of their practical deployment intersect. Rapid advancements in audio synthesis technologies, such as text-to-speech (TTS) and voice conversion (VC), have enabled the creation of highly realistic synthetic audio. These spoofed audio samples pose substantial threats to speaker verification systems, voice-controlled devices, and forensic audio analysis [8, 9, 10]. Traditional spoof detection methods primarily rely on labeled datasets that rapidly become obsolete due to the continuous evolution of spoofing techniques. Consequently, these conventional approaches struggle to maintain effectiveness against emerging and sophisticated attacks.

In this context, zero-shot learning offers a promising alternative by leveraging pretrained audio-language models to detect spoofed audio samples without requiring labeled examples of new or emerging spoofing techniques. However, the effectiveness and robustness of zero-shot spoof detection capabilities under different quantization conditions remain inadequately studied. Understanding the interplay between zero-shot learning and quantization is crucial for deploying robust audio models in resource-constrained real-world scenarios.

This study addresses this critical research gap by evaluating the zero-shot spoof detection performance of five state-of-the-art LALMs, GAMA, LTU-AS, MERaLiON, Qwen-Audio, and SALMONN, across different quantization precisions (16-bit and 8-bit). Through detailed experiments and analysis, we

\*These authors contributed equally to this work

explore how these models perform in zero-shot detection scenarios and explicitly quantify the effects of quantization on their accuracy and reliability. Our key contributions are:

- Conduct the first comprehensive analysis of zero-shot spoof detection capabilities of quantized LALMs, providing foundational insights into inherent strengths and limitations.
- Deliver empirical insights into trade-offs among quantization precision, model accuracy, and computational efficiency, enabling informed decision-making for practical deployments.
- Present practical guidelines and recommendations for optimizing LALMs, ensuring robust spoof detection performance within realistic computational constraints.

## 2. Formulation and Experimental Setup

Consider an audio sample represented as a sequence of acoustic frames  $X = \{x_1, x_2, \dots, x_T\}$ , where each frame  $x_i \in \mathbb{R}^d$ ,  $T$  is the total number of frames, and  $d$  is the dimensionality of each frame. Let  $f_\theta$  denote a pretrained LALM parameterized by weights  $\theta$ . Given an audio sample  $X$  and a prompt  $p$ , the LALM outputs a probability distribution over the vocabulary  $\mathcal{V}$ :  $\mathbf{P} = f_\theta(X, p) \in \Delta^{|\mathcal{V}|-1}$ . The binary classification decision is based on comparing the probabilities of the target classification tokens, assigning the class  $y = \text{spoof}$  if  $P(t_{\text{spoof}} | X, p) > P(t_{\text{bonafide}} | X, p)$ , else bonafide, where  $t_{\text{spoof}}, t_{\text{bonafide}} \in \mathcal{V}$  are the vocabulary tokens corresponding to the class labels.

In the zero-shot spoof detection scenario,  $f_\theta$  leverages general audio representations learned during pretraining, without explicit fine-tuning on spoof-specific data. Model quantization reduces the precision of parameters, defined by  $\theta_{\text{quant}} = \mathcal{Q}_p(\theta)$ , where  $p \in \{32, 16, 8\}$  represents precision levels (FP32, FP16, INT8 respectively), and  $\mathcal{Q}_p$  is the quantization function converting high-precision parameters into lower-precision formats. This study systematically explores three key research questions: **(i) RQ1 - Zero-shot spoof detection**, which investigates the intrinsic capability of pretrained LALMs to identify spoofed audio without explicit fine-tuning; **(ii) RQ2 - Impact of quantization**, examining how varying quantization precision (FP32, FP16, INT8) affects the performance of LALMs; and **(iii) RQ3 - Deployment optimization**, aimed at providing practical guidelines for selecting optimal quantization precision that balances detection performance with computational efficiency.

**Experimental Framework and Baseline Models:** To rigorously assess the zero-shot spoof detection capabilities, robustness to quantization, and practical deployment considerations of LALMs, we developed a comprehensive experimental framework. We evaluated five representative state-of-the-art pretrained LALMs selected for their architectural diversity and proven effectiveness in audio-related tasks. Specifically, **GAMA [11]** integrates an Audio Q-Former, which generalizes audio semantics, with multi-layer Audio Spectrogram Transformer (AST) modules capturing detailed audio characteristics, leveraging the pretrained Llama model [12] for advanced audio understanding. **LTU-AS [13]** combines Whisper’s robust encoder-decoder architecture [14] with a specialized time- and layer-wise transformer (TLTR) to jointly process speech transcription and audio event detection, further refining these representations through integration with Llama. **MERaLiON [15]**, specifically tailored for multilingual environments, combines a localized version of Whisper-large-v2 with SEA-LION V3 [16], and incorporates an adaptor module explicitly designed to enhance audio-text alignment and multilingual capability. **Qwen-Audio [17]**, developed by Alibaba, fuses Whisper-base’s au-

Table 1: *Distribution of spoofed and bonafide samples across ASVspoof2019, In-the-Wild, and WaveFake datasets.*

Dataset	Spoof	Bonafide	Total Samples
ASVspoof2019	63882	7355	71237
In-the-Wild	11816	19963	31779
WaveFake	13100	13100	26200

dio processing capabilities with a transformer decoder, enabling versatile audio-text alignment applicable across various audio forms including speech, music, and environmental sounds. Finally, **SALMONN [18]** integrates Whisper and BEATs audio encoders through a window-level Q-Former combined with a Vicuna-based [19] LLM, providing unified and sophisticated multimodal processing across speech, audio events, and music.

Our experiments utilize multiple diverse datasets containing bonafide and synthetically spoofed audio samples that were created using various synthesis methods. We specifically selected datasets that reflect distinct class distributions, thereby allowing comprehensive evaluation under realistic scenarios: ASVspoof2019 [20] is significantly skewed toward spoof samples, making it challenging for models to correctly identify genuine audio; In-the-Wild [21] predominantly comprises bonafide samples, testing the models’ sensitivity and specificity in a practical context; and WaveFake [22] presents a balanced dataset equally composed of spoofed and bonafide samples, providing a controlled setting for evaluating model generalization and robustness. Table 1 summarizes key dataset characteristics. This carefully chosen suite of datasets enables insightful analysis and enhances the practical relevance of our findings for effectively deploying robust LALMs in real-world applications.

**Evaluation Metrics and Implementation Details:** To thoroughly evaluate model performance, we employ a suite of complementary metrics: F1-Score, Accuracy, Balanced Accuracy, Matthews Correlation Coefficient (MCC), and classwise accuracy. The F1-Score provides a balanced measure of precision and recall, particularly valuable in handling class imbalance. Accuracy captures the overall prediction correctness, while Balanced Accuracy averages recall across classes, making it highly effective for imbalanced scenarios. MCC offers a comprehensive assessment of classification quality by considering all confusion matrix categories. Classwise accuracy delivers detailed insights into model behavior across individual classes. All experiments were executed using an NVIDIA A100 GPU, leveraging mixed-precision inferencing to enhance computational efficiency and accelerate matrix computations.

## 3. Results and Analysis

This section presents an in-depth analysis addressing our three primary research questions using five state-of-the-art LALMs across multiple datasets and precision levels.

**RQ1 - Zero-Shot Detection Capabilities:** We evaluated the zero-shot detection capabilities of GAMA, LTU-AS, MERaLiON, Qwen-Audio, and SALMONN across three distinct datasets: ASVspoof2019, In-the-Wild, and WaveFake. Initial results on the ASVspoof2019 indicated promising spoof detection performance, with MERaLiON and LTU-AS achieving high F1-scores of 0.946 and 0.925, respectively, at FP32 precision. However, deeper examination of balanced accuracy and class-wise accuracy metrics highlights critical issues. We observed a severe bias toward classifying most or all inputs as spoofed. This bias artificially inflated spoof detection accuracy while drastically reducing the ability to detect bonafide sam-

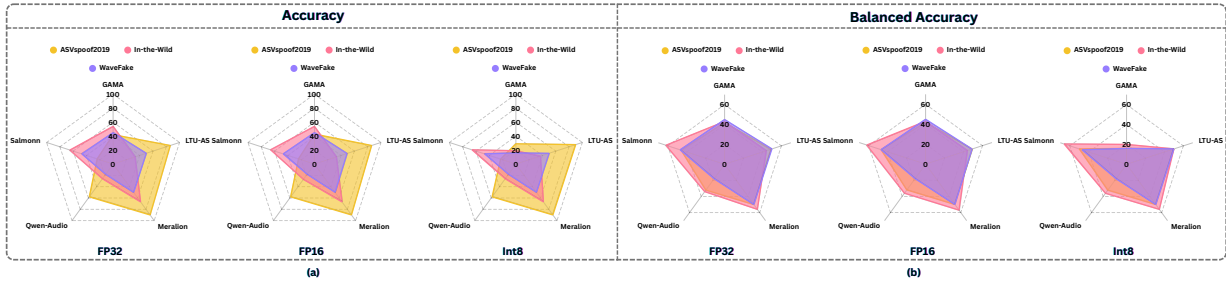


Figure 2: Illustrate **Accuracy (part a)** and **Balanced Accuracy (part b)** for five Large Audio Language Models (GAMA, LTU-AS, MERaLiON, Qwen-Audio, and SALMONN) across three datasets: ASVspoof2019 (yellow), In-the-Wild (pink), and WaveFake (purple). The results are presented for three quantization levels: FP32 (left), FP16 (center), and INT8 (right). Each plot highlights the models' ability to detect audio spoofing attacks under varying precision settings, showcasing differences in robustness and generalization capabilities across datasets and quantization strategies.

Table 2: Comparison of five Large Audio Language Models evaluated on the different datasets. Each row represents a specific metric: F1-Score, Accuracy, Balanced Accuracy, Matthews Correlation Coefficient (MCC), (Spoof or Bonafide) class-wise Accuracy. The columns list the performance of each model under three quantization levels (FP32, FP16, INT8), with values separated by slashes.

ASVspoof2019 (FP32 / FP16 / INT8)					
Metrics / Model	GAMA	LTU-AS	Meralion	Qwen	Salmonn
<b>F1-Score</b>	0.570 / 0.580 / 0.454	0.925 / 0.925 / 0.946	0.946 / 0.946 / 0.946	0.734 / 0.733 / 0.731	0.297 / 0.293 / 0.265
<b>Accuracy</b>	0.424 / 0.433 / 0.293	0.861 / 0.861 / 0.897	0.897 / 0.897 / 0.897	0.579 / 0.578 / 0.576	0.237 / 0.237 / 0.226
<b>Balanced Accuracy</b>	0.415 / 0.422 / 0.164	0.488 / 0.489 / 0.499	0.500 / 0.500 / 0.500	0.323 / 0.323 / 0.321	0.458 / 0.467 / 0.496
<b>MCC</b>	-0.104 / -0.096 / -0.419	-0.039 / -0.035 / 0.014	0.000 / 0.000 / 0.000	-0.231 / -0.232 / -0.233	-0.065 / -0.053 / -0.007
<b>Spoof Accuracy</b>	0.426 / 0.436 / 0.327	0.958 / 0.957 / 1.000	1.000 / 1.000 / 1.000	0.646 / 0.645 / 0.643	0.179 / 0.177 / 0.156
<b>Bonafide Accuracy</b>	0.405 / 0.407 / 0.000	0.017 / 0.020 / 0.001	0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000	0.737 / 0.756 / 0.836

In-the-Wild (FP32 / FP16 / INT8)					
Metrics / Model	GAMA	LTU-AS	Meralion	Qwen	Salmonn
<b>F1-Score</b>	0.014 / 0.014 / 0.177	0.489 / 0.490 / 0.542	0.247 / 0.254 / 0.238	0.408 / 0.425 / 0.418	0.510 / 0.507 / 0.592
<b>Accuracy</b>	0.536 / 0.542 / 0.193	0.329 / 0.330 / 0.372	0.670 / 0.672 / 0.668	0.256 / 0.270 / 0.264	0.648 / 0.657 / 0.662
<b>Balanced Accuracy</b>	0.429 / 0.433 / 0.201	0.438 / 0.490 / 0.500	0.563 / 0.565 / 0.560	0.344 / 0.363 / 0.355	0.616 / 0.620 / 0.661
<b>MCC</b>	-0.231 / -0.223 / -0.591	-0.254 / -0.250 / 0.000	0.246 / 0.250 / 0.238	-0.254 / -0.250 / 0.000	0.236 / 0.313 / 0.247
<b>Spoof Accuracy</b>	0.009 / 0.009 / 0.233	0.864 / 0.866 / 1.000	0.146 / 0.150 / 0.140	0.864 / 0.866 / 1.000	0.493 / 0.475 / 0.659
<b>Bonafide Accuracy</b>	0.849 / 0.170 / 0.858	0.012 / 0.013 / 0.000	0.981 / 0.980 / 0.981	0.012 / 0.013 / 0.000	0.740 / 0.764 / 0.664

WaveFake (FP32 / FP16 / INT8)					
Metrics / Model	GAMA	LTU-AS	Meralion	Qwen	Salmonn
<b>F1-Score</b>	0.511 / 0.512 / 0.280	0.661 / 0.660 / 0.667	0.667 / 0.667 / 0.667	0.300 / 0.301 / 0.300	0.456 / 0.456 / 0.438
<b>Accuracy</b>	0.446 / 0.445 / 0.163	0.495 / 0.494 / 0.500	0.500 / 0.500 / 0.500	0.177 / 0.177 / 0.176	0.469 / 0.470 / 0.470
<b>Balanced Accuracy</b>	0.446 / 0.445 / 0.163	0.495 / 0.494 / 0.500	0.500 / 0.500 / 0.500	0.177 / 0.177 / 0.176	0.469 / 0.470 / 0.470
<b>MCC</b>	-0.113 / -0.114 / -0.713	-0.039 / -0.048 / 0.010	0.000 / 0.000 / 0.000	-0.691 / -0.691 / -0.692	-0.061 / -0.061 / -0.060
<b>Spoof Accuracy</b>	0.580 / 0.582 / 0.325	0.982 / 0.981 / 1.000	1.000 / 1.000 / 1.000	0.354 / 0.354 / 0.353	0.445 / 0.445 / 0.414
<b>Bonafide Accuracy</b>	0.311 / 0.309 / 0.000	0.009 / 0.008 / 0.000	0.000 / 0.000 / 0.000	0.000 / 0.000 / 0.000	0.494 / 0.494 / 0.527

ples. As a result, the overall accuracy metrics were misleading and resembled random predictions.

Further dataset-wise analysis provides deeper insights into the distinct behaviors of each model. The spoof-heavy ASVspoof2019 dataset exacerbated model biases, causing robust models such as MERaLiON and LTU-AS to frequently misclassify genuine audio samples as spoofed. The balanced WaveFake dataset offered a clearer assessment of generalization capabilities, but here too, models exhibited substantial biases. The predominantly bonafide In-the-Wild dataset highlighted LALMs' difficulties in accurately detecting genuine audio, emphasizing fundamental limitations in these architectures.

Quantitative evaluations further emphasized these concerns. MERaLiON demonstrated exceptional stability across quantization levels, consistently achieving an identical F1-score (0.946) and overall accuracy (0.897). However, its perfect spoof detection (accuracy of 1.000) came at the complete expense of bonafide detection (accuracy of 0.000). LTU-AS displayed similarly robust spoof detection, with an anomalous improvement

in spoof accuracy under INT8 precision (from 0.958 to 1.000). This anomaly was not indicative of improved performance but rather a reinforcement of the spoof-classification bias. GAMA showed heightened sensitivity to quantization, with its balanced accuracy deteriorating sharply from 0.415 at FP32 precision to 0.164 under INT8 precision, indicating potential vulnerabilities.

Qwen-Audio maintained moderate yet consistent spoof detection performance (F1-scores around 0.732) across precision levels but entirely failed to detect bonafide samples. SALMONN uniquely presented relatively strong bonafide detection capabilities (0.737–0.836 accuracy), despite generally poor overall metrics (F1-score ranging from 0.265 to 0.297). Across all models, MCC values remained consistently low or negative, emphasizing severe prediction bias and inadequate balanced performance. These findings highlight a troubling trend: current LALMs are strongly biased towards spoof detection at the expense of accurate bonafide identification, reflecting critical limitations that must be addressed through substantial architectural refinements to achieve reliable, balanced spoof

Table 3: Comparison of memory usage (GB) and inference time (seconds per 100 samples) at different quantization precisions.

Model / Factor	FP32		FP16		INT8	
	Memory	Time	Memory	Time	Memory	Time
GAMA	26.19	129.06	13.12	126.20	6.94	177.76
LTU-AS	25.35	98.67	12.71	91.43	6.74	150.49
Meralion	37.05	90.77	18.49	57.55	10.24	107.31
Qwen	31.35	52.51	15.64	50.06	9.12	112.87
Salmomn	51.47	1020.38	27.34	960.38	16.08	1140.80

detection suitable for real-world applications.

**RQ2 - Effects of Quantization on Model Performance:** Evaluating quantization at FP32, FP16, and INT8 precision levels showed performance trade-offs, highlighting the varying sensitivity of LALMs to reduced numerical precision. Generally, lowering precision negatively impacted model accuracy and reliability. However, we observed instances where INT8 quantization seemingly improved certain metrics, notably spoof detection accuracy. Upon deeper inspection, we identified these apparent improvements as statistical artifacts resulting from increased classification bias toward the spoof class. Specifically, INT8-quantized models excessively favored spoof predictions, artificially inflating overall accuracy while severely degrading balanced accuracy and genuine audio detection capability

A practically significant finding was the minimal performance difference between FP32 and FP16 across most models. Models such as MERaLiON and LTU-AS demonstrated almost identical performance at both FP32 and FP16 precision levels, suggesting FP16 quantization as an optimal practical choice. Specifically, FP16 precision can substantially reduce memory usage, approximately halving the required computational resources, without negatively impacting genuine detection performance. This finding positions FP16 quantization as an efficient strategy for deploying LALMs in real-world, resource-constrained scenarios.

GAMA, on the other hand, exhibited pronounced sensitivity to quantization, with a dramatic decline in balanced accuracy from FP32 to INT8 precision, indicating that certain architectural elements may significantly influence model resilience to precision reduction. In contrast, MERaLiON maintained consistency across all precision levels, potentially attributable to its adaptive audio-text alignment modules and multilingual training strategy. These modules likely enhance the robustness of internal representations, helping mitigate performance degradation under reduced precision conditions. Similarly, LTU-AS’s performance stability at FP16 precision suggests that integrating multiple transformer layers and multimodal embedding strategies can provide quantization resilience.

As precision decreases from FP32 to INT8, memory consumption consistently drops across all models. Moving from FP32 to FP16 reduces both memory consumption and inference time for all models as shown in Table 3, demonstrating clear efficiency gains. However, while INT8 quantization achieves the greatest memory reduction, it paradoxically increases inference time compared to FP16. This suggests that INT8 quantization, while highly effective for memory optimization, introduces computational overhead due to data type conversions and associated processing, negating its execution-speed benefits.

**RQ3 - Practical Guidelines for Deployment:** In our experiments, none of the tested LALMs demonstrated sufficient balanced performance or reliability suitable for immediate real-world deployment in spoof detection scenarios. The prevalent issue of pronounced predictive bias toward spoof classification significantly undermines their practical applicability, highlighting an urgent need for architectural and methodolog-

ical refinements. Nevertheless, our findings offer clear and actionable insights for practical deployment strategies. FP16 quantization consistently emerged as the most favorable precision level across evaluated models, delivering a near-optimal balance between computational efficiency and detection accuracy. The minimal performance difference between FP32 and FP16 precision levels observed for models such as MERaLiON and LTU-AS reinforces the practicality of deploying FP16 quantized LALMs. Specifically, FP16 precision substantially reduces computational and memory requirements, effectively halving resource usage, making it particularly suitable for edge computing and mobile application scenarios.

INT8 quantization presents a more complex and nuanced trade-off. While theoretically ideal for extremely resource-constrained environments, our analysis shows significant risks, as INT8 quantization notably exacerbates predictive biases, severely compromising balanced classification performance. Therefore, INT8 deployment thus requires cautious consideration and rigorous pre-deployment validation processes to ensure reliability and performance integrity in practical scenarios.

Finally, our results suggest clear directions for future research and practical deployment enhancements. Architectural features such as adaptive alignment modules, robust multimodal embeddings, and multilingual training appear critical for enhancing quantization resilience. Incorporating these elements could mitigate quantization-induced biases, substantially improving the reliability and performance of LALMs in resource-constrained environments. These insights provide a foundational roadmap for future efforts aimed at refining LALMs into truly effective, balanced, and robust models suitable for real-world audio spoof detection.

## 4. Conclusion

This study presents the first comprehensive evaluation of zero-shot spoof detection capabilities of quantized LALMs, uncovering significant biases toward spoof classification that undermine practical applicability. Despite promising initial spoof detection results, deeper analysis of balanced metrics revealed fundamental limitations across all evaluated models, emphasizing the urgent need for substantial architectural refinements. Our analysis elucidates clear performance-precision trade-offs, highlighting FP16 quantization as an optimal choice that effectively halves memory and computational requirements with negligible accuracy degradation compared to FP32. Conversely, INT8 quantization, despite theoretical benefits in extreme resource-constrained scenarios, significantly intensifies predictive biases, requiring cautious consideration supported by rigorous pre-deployment validation. Additionally, our results showcased an interesting trade-off: while INT8 quantization significantly reduces memory consumption, it unexpectedly increases inference time compared to FP16. This highlights the necessity of carefully evaluating quantization strategies not only in terms of memory but also computational speed. Future architectures should incorporate adaptive alignment modules, robust multimodal embeddings, and multilingual training for improved quantization resilience. These strategies provide foundational guidance for deploying reliable, efficient LALMs in spoof detection.

## 5. Acknowledgment

This research is supported by a grant from the NSM, MeitY. The authors also gratefully acknowledge the support of IndiaAI and Meta through Srijan: Centre of Excellence for Generative AI.

## 6. References

- [1] J. Peng, Y. Wang, Y. Xi, X. Li, X. Zhang, and K. Yu, "A survey on speech large language models," in *arXiv-2410.18908, CoRR*, 2025.
- [2] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*, 2023.
- [3] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, "Listen, think, and understand," in *The Twelfth International Conference on Learning Representations*, 2024.
- [4] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, and et. al., "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2025.
- [5] S. Ji, Z. Jiang, H. Wang, J. Zuo, and Z. Zhao, "Mobilespeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 13 588–13 600.
- [6] B. Dutta, R. Ranjan, A. Jain, R. Singh, and M. Vatsa, "Can rag-driven enhancements amplify audio llms for low-resource languages?" in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.
- [7] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," in *arXiv-2106.08295, CoRR*, 2021.
- [8] R. Ranjan, M. Vatsa, and R. Singh, "Sv-deit: Speaker verification with deitcap spoofing detection," in *IEEE International Joint Conference on Biometrics*, 2023, pp. 1–10.
- [9] R. Ranjan, B. Dutta, M. Vatsa, and R. Singh, "Faking fluent: Unveiling the achilles' heel of multilingual deepfake detection," in *IEEE International Joint Conference on Biometrics*, 2024, pp. 1–10.
- [10] R. Ranjan, M. Vatsa, and R. Singh, "Context encoded multi-modal attention network for detecting audio spoofing," in *IEEE International Joint Conference on Biometrics*, 2024, pp. 1–11.
- [11] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 6288–6313.
- [12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," in *arXiv-2302.13971, CoRR*, 2023.
- [13] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. R. Glass, "Joint audio and speech understanding," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8.
- [14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 28 492–28 518.
- [15] Y. He, Z. Liu, S. Sun, B. Wang, W. Zhang, X. Zou, N. F. Chen, and A. T. Aw, "Meralion-audiollm: Bridging audio and language with large language models," in *arXiv-2412.09818, CoRR*, 2024.
- [16] AI-Singapore, "Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia," in <https://github.com/aisingapore/sealion>, 2024.
- [17] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," in *arXiv-2311.07919, CoRR*, 2023.
- [18] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "SALMONN: towards generic hearing abilities for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.
- [19] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality," in <https://lmsys.org/blog/2023-03-30-vicuna>, 2023.
- [20] A. Nautsch, X. Wang, N. W. D. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.
- [21] N. M. Müller, P. Czempin, F. Dieckmann, A. Froggyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *International Speech Communication Association*, 2022, pp. 2783–2787.
- [22] J. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.