AU-Harness: An Open-Source Toolkit for Efficient and Unified Evaluation of AUdio-LLMs

Anonymous ACL submission

Abstract

Large Audio Language Models (LALMs) are rapidly advancing, but evaluating them remains challenging due to inefficient toolkits that limit fair comparison and systematic assessment. Existing evaluation frameworks exhibit two critical limitations: (1) slow and inefficient processing pipeline that bottlenecks large-scale studies, (2) the absence of unified and scalable evaluation framework capable of keeping pace with the rapid growth of both LALMs and audio benchmarks. To address these issues, we introduce **AU-Harness**, an efficient and comprehensive evaluation framework for LALMs. Our system achieves a speedup of up to 151% over existing toolkits through optimized batch processing and parallel execution, enabling large-scale evaluations previously impractical. We provide standardized prompting protocols and flexible configurations for fair model comparison across diverse scenarios. Through evaluation across diverse sets of tasks, we reveal significant gaps in current LALMs. Our findings also highlight a lack of standardization in modalities of user-provided instructions existent across audio benchmarks, which can lead to performance differences of up to 7.2 absolute points on challenging complex instruction following downstream tasks. AU-Harness provides both practical evaluation tools and insights into model limitations, advancing systematic LALM development. [1]

1 Introduction

The emergence of Large Audio Language Models (LALMs) has opened new frontiers, extending capabilities beyond textual inputs to speech, sounds, and multimodal inputs (Tang et al., 2023; Cui et al., 2024). This progress has accelerated the development of frontier LALMs and audio-focused benchmarks. Recent multimodal LALMs like Gemini 2.5 (Comanici et al., 2025), Qwen2.5-Omni (Xu et al., 2025) have demonstrated substantial audio understanding capabilities well beyond traditional Automatic Speech Recognition (ASR) tasks. However, despite these advances, audio evaluation toolkits have comparatively received little attention. Thus, there is a need for efficient, customizable, and consistent evaluation framework for fair model comparisons which can evolve as audio tasks and model complexities grow.

Existing efforts including AudioBench (Wang et al., 2025a), Kimi-Eval (Ding et al., 2025), VoiceBench (Chen et al., 2024) and LMMS-Eval (Zhang et al., 2025b) have provided extensive task coverage from ASR to spoken question answering and scene understanding. However, prevailing toolkits still face three persistent limitations. First, **throughput**: many pipelines underutilize batching and parallelism, creating bottlenecks that preclude large-scale, systematic comparisons. Second, **reproducibility**: ad-hoc prompting and non-standardized evaluation settings lead to incomparable performance across setups. Third, **task scope**: evaluations remain largely restricted to the static single-turn interactions, failing to account for LALM assessment over extended interactions in multi-round conversational settings which are more frequent in real world conversations.

Most current evaluation frameworks depend on simplistic yet inefficient input processing pipelines that struggle to scale with the increasing volume and complexity of audio benchmarks and LALMs. These limitations not only constrain the throughput of large-scale evaluations but also hinder fair and reproducible comparisons across models of different sizes and architectures. As the field progresses toward more diverse and challenging audio tasks, the shortcomings of current evaluation infrastructure may pose a critical bottleneck, ultimately hampering the potential progress of LALMs. Unlike previous evaluation frameworks, we introduce an efficient token request orchestration together

[1] https://anonymous.4open.science/r/AU-Harness-5C15

with effective data sharding to scale the evaluations across multiple nodes and hardware architectures, leading to improved efficiency for audio benchmark evaluations.

Beyond computational efficiency, existing toolkits also suffer from a notable lack of customizable configurations for different audio task configurations, severely limiting their utility for diverse research and application needs. Insufficient attention to task-specific customizations remains a significant challenge for LALMs' evaluation and comparison across different benchmarks. Prompt sensitivity further compounds customizability concerns, since LALMs' outcomes can be sensitive to prompt phrasing (Cui et al., 2024). The challenges of the existing toolkits are summarized in Table 1.

Our contributions are as follows:

- We propose an **efficient evaluation engine** that leverages vLLM batching and dataset sharding to scale evaluations to multi-node infrastructures without sacrificing fidelity.
- A **unified, configurable framework** that standardizes prompting and metrics across benchmarks, enabling fair, reproducible comparisons and easy task integration.
- To the best of our knowledge, **AU-Harness** is the first evaluation frameworks to explicitly support **multi-turn** conversational capabilities assessment in LALMs, enabling systematic in-depth analyses of dialogue behaviors over extended interactions.

2 Related Work

Audio Benchmarks. Benchmarks play a critical role in the development of LALMs. SUPERB (Yang et al., 2021) established core task axes (Content, Speaker, Semantics, Paralinguistics) for audio model evaluation. DynamicSUPERB (Huang et al., 2024b) and DynamicSUPERB-2.0 (Huang et al., 2024a) expanded coverage to instruction-tuned and sequence generation tasks across speech, music, and environmental audio. Instruction-following and agentic conversational behaviors have been further probed by AIR-Bench (Yang et al., 2024) and VoiceBench (Chen et al., 2024). More recently, AudioBench (Wang et al., 2025a) unified 8 task families over 26 datasets for AudioLLMs.

Complementary efforts in 2025 broaden the breadth and depth with audio reasoning capabilities: X-ARES (Zhang et al., 2025a) systematically assesses general audio encoders across domains, MECAT (Niu et al., 2025) targets fine-grained audio understanding with expert-guided captions and QA. MMAR (Ma et al., 2025), MMAU-PRO (Kumar et al., 2025), and MMSU (Wang et al., 2025b) focus on understanding and analyzing complex audio scenes, spatial relationships, and mixed-audio reasoning. CodecBench (Wang et al., 2025c) benchmarks codecs from acoustic and semantic perspectives. Despite the rapid growth of audio benchmarks, the development of audio evaluation frameworks allowing for fair and consistent comparisons between frontier models and benchmarks remains fairly understudied. This critical gap necessitates the development of a unified and efficient evaluation engine designed specifically for scalable audio evaluations under the rapid expansion of LALMs and audio benchmarks.

Audio Evaluation Kits. In contrast with Audio Benchmark development, Audio Evaluation Kits have received less attention. This can be primarily attributed to the straightforward nature and minimal setup requirements of the early audio tasks, as presented in Huang et al. (2024b) and Yang et al. (2024). However, the rapid growth of LALMs and the increasing complexity of newly curated audio benchmarks have underscored the critical need for comprehensive evaluation kits, as exemplified through the recent development of extensive evaluation kits (Ding et al., 2025; Wang et al., 2025a; Zhang et al., 2025b). For instance, AudioBench (Wang et al., 2025a) offers versatile evaluation support for up to 8 tasks across 26 datasets. VERSA (Shi et al., 2025) introduces a comprehensive framework to evaluate the quality of various speech, audio and music signals, with the focus on text-to-audio applications. Despite these advancements, most current evaluation kits operate on the simplified assumption that *a single model is evaluated against a single benchmark per run*. Addressing this limitation, we introduce an efficient, customizable evaluation kit to support the massive scale of the current LALMs and audio benchmarks as summarized in Table 1.

3 LALM Evaluation Challenges

3.1 Inference Efficiency

Most existing LALM evaluation kits have been designed based on the assumption that *a single model should be evaluated against a single benchmark per run*. However, this constrains researchers

EvalKit	vLLM support	HF Support	Multi-turn	Multi-task Parallel	Configurable Customizations
AudioBench	✗	✓	✗	✗	✗
Kimi-Eval	✗	✓	✗	✗	✗
VoiceBench	✗	✓	✗	✗	✗
LMMs-Eval	✓	✓	✗	✓	✗
AU-Harness	✓	✓	✓	✓	✓

Table 1: **Feature comparison of contemporary LALM evaluation toolkits.** We evaluate key technical capabilities across existing frameworks: vLLM integration for efficient batching, HuggingFace(HF)-model support, multi-turn dialogue support for conversational scenarios, parallel processing support for multi-task concurrent evaluation, and configurable customizations for flexible evaluation design. Our framework is the first to provide comprehensive support across all dimensions.

EvalKit	RTF (↓)	Samples Processed per Second (↑)
AudioBench	19.9	0.66
Kimi-Eval	7.1	1.87
VoiceBench	87.9	0.15
LMMS-Eval	4.3	3.04
AU-Harness	3.6 (↓16.28%)	3.65 (↑ 20.07%)

Table 2: **Throughput efficiency comparison across LALM evaluation frameworks.** Results averaged over 500 samples from LibriSpeech-test-clean (1.05 hours total audio) across 3 LALMs (Qwen2.5-Omni, Voxtral-Mini and Phi-4-Multimodal). Real-time Factor (RTF, ↓ better) measures processing time relative to audio duration. Samples Processed per Second (↑ better) quantifies raw throughput. Our framework achieves 16.28% RTF reduction and 20.07% throughput increase over the best competing baseline, demonstrating substantial efficiency gains through vLLM integration and request orchestration.

from conducting systematic, large-scale comparisons across LALMs and audio benchmarks efficiently, slowing the iterative process of model development and refinement. The current evaluation kits also under-utilize parallel processing capabilities available in the high-performance computing clusters, resulting in failures in incorporating benefits of available hardware infrastructures.

Two essential task-agnostic metrics for evaluating the efficiency of LALM evaluation frameworks are *Real-time Factor (RTF)* and *Samples Processed per Second (SPS)*. RTF measures the processing time of an evaluation framework relative to the duration of the processed audio (Arriaga et al., 2024). Lower RTF is more desirable, indicating a more efficient audio evaluation framework. On the other hand, SPS directly quantifies the model's processing speed by measuring the average number of audio samples processed per second. It serves as a complementary measure to RTF, providing a more granular view of the model's throughput and computational efficiency. Both metrics are formularized as follows:

$$RTF = \frac{\sum_{i=1}^{N} T_{i,\text{proc}}}{\sum_{i=1}^{N} D_{i,\text{audio}}} \quad (1)$$

$$SPS = \frac{N}{\sum_{i=1}^{N} T_{i,\text{proc}}} \quad (2)$$

where T_{proc} is the total time (in seconds) taken by the framework to process the evaluation of the given audio. D_{audio} is the total duration of the input audio signal (in seconds) under 16kHz sampling rate, N is the total number of audio samples processed.

To quantify the efficiency of existing evaluation frameworks, we conduct a study on $N = 500$ audio samples (approximately 1.05 hours) of Librispeech-test-clean (Panayotov et al., 2015). As observed in Table 2, existing audio evaluation kits exhibit high RTF and slow sample processing speed. As the number of samples continues to increase with more diverse datasets, this challenge can significantly slow down the inference progress at scale.

3.2 Customizable Evaluation Configurations

Despite the strong support for various tasks and LALMs, existing evaluation frameworks exhibit significantly rigid architecture design. This fixed configuration fails to accommodate the modular requirements of audio benchmarks and LALMs.

Multi-turn Dialogue Support Previous audio evaluation toolkits have largely been constrained to tasks centered on single-turn user interactions. However, as the field moves toward building interactive and context-aware voice assistants, the ability to evaluate multi-turn tasks becomes increasingly critical. Multi-turn evaluation enables a more realistic assessment of dialogue continuity, contextual reasoning, and the model's capacity to adapt dynamically across extended conversations. Without such support, current evaluation approaches

risk overlooking key aspects of usability and robustness that are essential for next-generation LALMs in realistic agentic voice systems.

Evaluation Customization. The lack of customizable filtering poses a significant barrier for researchers aiming to conduct in-depth analyses of current LALM limitations. Without the ability to refine evaluation datasets based on specific criteria, it is challenging to gain granular understanding of model performance across diverse audio conditions. For instance, while certain LALMs might perform reliably on 10-second audio chunks, they might be unable to handle short-form audio typically encountered in dialogue-state tracking systems.

4 AU-Harness

In response to the presented challenges of current audio understanding evaluation toolkits, we propose a standardized, efficient, highly customizable evaluation framework, **AU-Harness**, as detailed in Figure 1.

AU-Harness is composed of 3 primary components: **Config**, **Request Controller** and **Concurrent Engines**. The Config module defines a structured and hierarchical representation of customizable configurations, enabling flexible and transparent evaluation settings. The Request Controller is responsible for managing token requests and coordinating execution across the framework. Finally, the Concurrent Engines module carries out task-specific evaluations in parallel, where each engine can support multi-model evaluations tailored to particular tasks. In the following sections, we introduce our architecture design in detail to address the presented challenges in Section 3.

4.1 Inference Efficiency

As illustrated in Figure 1, AU-Harness maximizes inference efficiency through a token-based request scheduling architecture. More specifically, we introduce a Central Request Controller that maintains and regulates a pool of available tokens which are accessible to all models across all evaluation engines. Here, a *token* refers to a concurrency slot representing permission to issue one inference request (not a model input token), which is acquired before dispatch and released upon completion. Each concurrent engine-specific requester periodically draws from the global pool. Within each engine, multiple models are executed concurrently on a targeted dataset, with inference calls dispatched in parallel to fully exploit available computational resources. This architecture ensures that evaluation throughput is not bottlenecked by model or engine-specific constraints, but rather governed solely by user-defined request limits set globally, providing both scalability and predictable performance guarantees. Furthermore, we allow user-specified retry counts on request errors, enabling users to set higher request limits with the assurance that occasional failures will be re-tried and successfully completed, thereby offering a tunable balance between throughput and reliability.

Furthermore, AU-Harness implements a layered request synchronization strategy that adaptively staggers request wait times across concurrent models. This design increases the probability that all models processing a given dataset segment complete their inference in a temporally aligned manner. By reducing discrepancies in model response times, the strategy minimizes idle periods within each engine, thereby mitigating intra-engine waiting time and improving overall throughput efficiency.

Additionally, we implement dataset sharding, which partitions the evaluation dataset into disjoint subsets to enable parallel processing across multiple model endpoints. To maximize efficiency, sharding is performed proportionally to each endpoint's capacity for concurrent requests, ensuring balanced utilization of heterogeneous resources. This enables near-linear scaling of inference throughput, effectively distributing the computational workload and minimizing bottlenecks. Finally, our native integration with vLLM leverages a range of inference-level optimizations, further accelerating model execution and overall evaluation system.

4.2 Customizable Evaluation Configurations

AU-Harness is highly customizable, allowing for seamless integration support of diverse LALMs and audio benchmarks.

Decoupling Inference and Model Hosting. AU-Harness decouples predictive inference and metric computation from model hosting infrastructure. In this way, regardless of whether the model is served through vLLM, a third-party API, or a lightweight FastAPI (Ramirez, 2018) deployment, the request handler requires only a standardized model specification to integrate seamlessly with the inference pipeline. This separation not only promotes modularity and extensibility of the evaluation framework

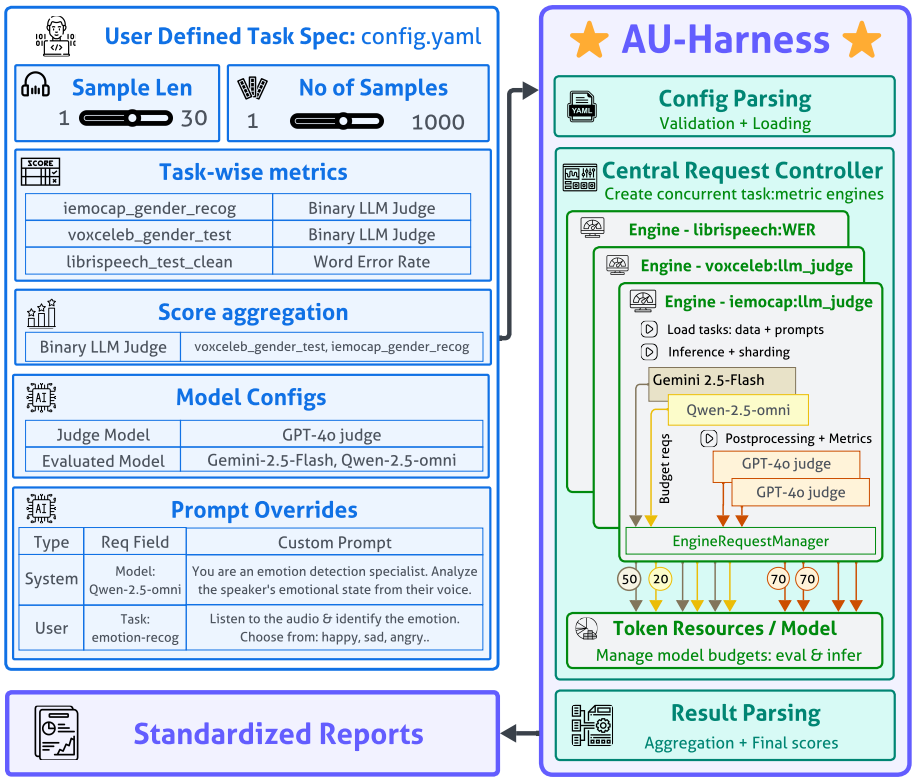


Figure 1: **Architecture overview of AU-Harness evaluation framework.** Our system comprises three core components: (1) *Config* module for hierarchical task configuration and standardized prompting, (2) *Central Request Controller* managing token-based concurrency limits across all engines with adaptive retry mechanisms, and (3) *Concurrent Engines* executing parallel model evaluation with dataset sharding. The Central Request Controller maintains a global Token Pool accessible to all engines, enabling efficient resource utilization and scalable throughput. Multiple concurrent connections between the controller and inference models illustrate parallel request dispatch, with each engine supporting the evaluation of multiple models on targeted datasets.

.

but also enables straightforward integration via simplified future integration of alternative inference strategies.

Wide Model Support. AU-Harness is designed for broad model compatibility, enabling out-of-the-box evaluation across diverse inference backends. It provides native support for vLLM compatible models, which deliver high-throughput and memory-efficient inference. Models not integrated with vLLM are also supported, as long as they expose a standard /v1/chat/completions endpoint. This flexibility maximizes model coverage by enabling seamless evaluation across both vLLM-compatible and non-compatible models. To facilitate the integration, we provide boilerplate FastAPI server implementations that make it easy to build lightweight inference endpoints. Alternatively, developers can also bring their own optimized inference stacks and wrap them with FastAPI to integrate smoothly with AU-Harness, ensuring minimal overhead and maximum compatibility.

Evaluation customization. AU-Harness is also designed for granular control over evaluation steps. First, AU-Harness supports both open-source and proprietary models, which might contain their individualistic settings. Second, we allow for customizable metric assignment on a per-dataset and/or per-task basis. For instance, LLM-as-judge supports configurable concurrency to maximize the throughput for evaluation stage. For a more comprehensive understanding of model performance, the framework offers configurable aggregation metrics. This capability allows for the multi-dimensional analysis of task and metric results, providing a comprehensive outlook that extends beyond simple, individual scores or sub-tasks.

As shown in Figure 1, users specify evaluation behavior entirely through a YAML configuration, including the task–wise metrics, model configurations, optional score aggregation across sets of related tasks, and optional prompt overrides. Importantly, no task-specific Python code or glue logic is required, even for complex benchmarks such as

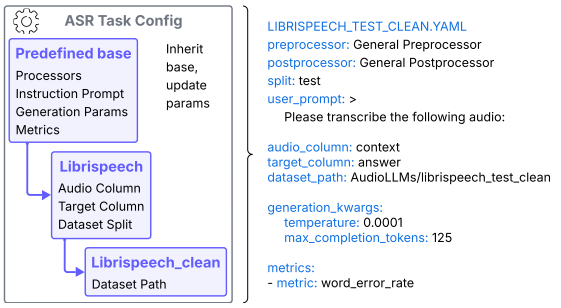


Figure 2: Sample task configuration for Librispeech clean in ASR, showcasing the ease of setup for new benchmarks which can utilize the base task configs and update required fields only for quick setup.

Emotion Recognition task suite that involves aggregation of multiple sub-datasets and LLM-based judging. This abstraction cleanly decouples tasks, metrics, models, and judges, enabling new evaluations to be launched by editing a single configuration file.

4.3 Support for features & tasks

Multi-turn Dialogue support. By using synchronous, turn-based evaluation chains that recursively append model outputs to the context, AU-Harness supports multi-turn evaluation of both audio and text datasets across LALMs. The simplicity and conciseness of our implementation establishes the robust foundations for future integration of highly complex and custom multi-turn benchmarks.

Continual engagement with emerging tasks. We aim to continue supporting up-and-coming benchmarks and models so AU-Harness can continue providing insights into LALM performance as they evolve. Figure 2 shows the ease of adding new datasets given the task category - the setup for a new task requires minimal updates with the given base configs to be supported in AU-Harness .

5 Results & Discussion

Without loss of generality, we adopt the task taxonomy proposed by Dynamic-SUPERB2.0 (Huang et al., 2024b) for the empirical evaluations with our proposed AU-Harness due to its exhaustive coverage. Table 3 characterizes the breadth of audio evaluation suite supported by our AU-Harness, demonstrating the flexibility of our AU-Harness in supporting diverse audio tasks. Following (Wang et al., 2025a), we adopt GPT-4o-mini as judge for LLM-judge metrics due to its advanced capability. Further details of datasets and metrics are provided in Appendix A.1.

5.1 Inference Efficiency

Evaluation Settings. We perform an empirical evaluation to compare AU-Harness against existing evaluation kits: AudioBench (Wang et al., 2025a), VoiceBench (Chen et al., 2024), Kimi-Eval (Ding et al., 2025), and LMMS-Eval (Zhang et al., 2025b). Our analysis focuses on the two key metrics RTF and Processed Samples per Second detailed in Section 3.1. We leverage 500 audio samples from 3 diverse datasets: librispeech-clean-test, ClothoAQA (Lipping et al., 2022), and MELD-Emotion (Poria et al., 2019) as detailed in Table 7. The evaluation is conducted on three different LALMs, including: Qwen2.5-Omni-7B (Xu et al., 2025), Phi-4-Multimodal (Abouelenin et al., 2025) and Voxtral-Mini-3B (Liu et al., 2025). For conciseness, we report the averaged metric across all 3 LALMs. Additional runtime setups, namely *Sequential* and *Parallel* [2], to assure a comprehensive and fair comparison among all existing evaluation kits are also examined as detailed in Appendix A.2.

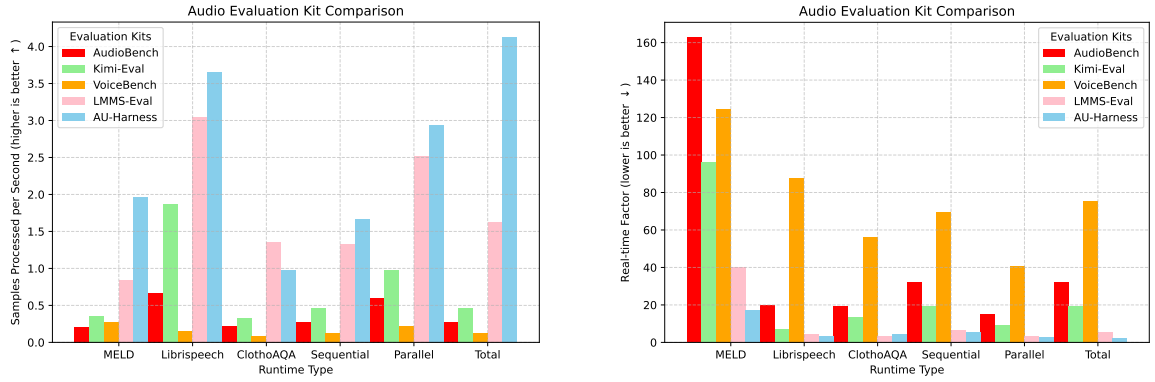
Evaluation Comparison As shown in Figure 3, AU-Harness consistently outperforms existing evaluation kits across all runtime scenarios in two key efficiency metrics. Specifically, AU-Harness achieves up to a 151% improvement in SPS and 61% reduction in RTFs compared to the next most competitive evaluation frameworks. More importantly, our *Total* evaluation execution, illustrated in Figure 4, is significantly more efficient than competing frameworks. These empirical results validate our framework as a highly efficient tool for LALM evaluation.

Orchestration pipeline beyond vLLM integration Although our AU-Harness builds upon vLLM, our proposed orchestration layer demonstrated in Figure 1 delivers effective efficiency gains compared to a naive vLLM integration. We conduct additional comparative studies against LMMS-Eval (Zhang et al., 2025b), which similarly integrates vLLM to address inference efficiency challenges of large-scale multi-modal evaluations. Unlike our proposed orchestration with dynamic concurrent requests, LMMS-Eval primarily relies on vLLM default processing pipeline with multi-threaded execution. As demonstrated in

[2]Parallel refers to Parallel(Optimal) where no actual overhead is accounted for unless specified otherwise.

Models	Speech											Audio & Music	
Task Category	**PR**	**ASR**	**Paralinguistics**	**Speaker & Language**	**SLU**	**Hearing Disorder**	**SE**	**Safety**	**Multi-turn**		**IF**	**AU**	**MU**
Dataset	voxangeles	Librispeech	IEMOCAP	Speaker Recognition	BigBench Audio	Stuttering Detection	NoiseDetection	Advbench	MT-Bench	SpokenWoz	IFEval	audiocaps_qa	mu_chomusic_test
Metrics	LB (↑)	WER (↓)	LB (↑)	LB (↑)	LBBA(↑)	LB (↑)	LB (↑)	SafetyJudge (↑)	MTJudge (↑)	JGA (↑)	IFScore (↑)	LB (↑)	LB (↑)
Small-sized Audio Language Models (<5B parameters)													
Voxtral-Mini-3B	0.1	2.1	54.9	45.8	43.5	12.9	14.5	78.5	65.88	24.92	40.02	14.96	45.4
Qwen2.5-Omni-3B	1.2	8.09*	81.5	55.9	44.8	58.4	58.4	97.3	59.81	31.30++	35.82	42.82	53.5
Medium Sized Large Audio Language Models (5B-20B parameters)													
Phi-4-Multi-modal	0	1.97	50.5	47.2	40.8	42.1	32.5	97.1	64.12	7.82	44.51	26.08	44.8
Qwen2.5-Omni-7B	21.6	1.74	85.8	62.3	50.9	68.2	59	98.3	64.56	37.3++	50.83	38.4	59.3
Kimi-Audio	1.3	1.41**	89	62.8	41.7	58.4	96	100	54.62	37.3	61.29	38.46	66.8
Large Sized Large Audio Language Models (>20B parameters)													
Voxtral-Small-24B	1.2	1.62	42.8	47.7	66.5	51.9	15.5	75.4	70.81	29.93	66.83	19.24	57.9
Qwen3-Omni-30B -A3B-Thinking	50.9	1.64**	82.5	65.3***	96.8	68.7	78	95	76.19	11.16	80.39	37.96	74.3
Proprietary Audio Language Models													
GPT-4o-mini-audio	0	6.25	–	40.3	63.7	3.8	53	88.1	65	28.39	70.47	15.08	50.2
Gemini2.5-Flash	35.8	2.17	92.7	60.2	90.3	64.6	87.5	98.5	74.5	52.09	84.63	36.16	72.9
Cascaded Systems													
Whisper-Large-v3 + GPT-oSS-20B	48	9.82	1.9	48.1	78.2	49.8	49	98.5	71.5	19.02	73.72	12.14	50.3
GPT-4o-transcribe + GPT-4.1-mini	30.7	4.71	30.2	45.8	74.3	49.7	47	97.3	67.62	19.44	66.69	17.44	52.7

Table 3: **LALM performance on audio tasks.** We evaluate representative LALMs from different spectra: Open-source LALMs(small-sized, medium-sized, large-sized), Proprietary LALMs and Cascaded System LALMs across representative task categories. Metrics include LLM-as-judge evaluations using GPT-4o-mini and task-specific automatic metrics. **PR**: Phoneme Recognition, **ASR**: Automatic Speech Recognition, **SLU**: Spoken Language Understanding, **SE**: Speech Enhancement, **IF**: Instruction Following, **AU**: Audio Understanding, **MU**: Music Understanding. **Bold**: highest; underline: second highest. Refer to Appendix A.1.2 for metric abbreviations and detailed explanations.



(a) Samples Processed per Second (↑)

(b) Real-time Factor (↓)

Figure 3: **Efficiency comparison across evaluation frameworks and runtime scenarios.** (a) Samples Processed per Second (↑ better) and (b) Real-time Factor (↓ better) measured across three datasets (MELD-Emotion, LibriSpeech-test-clean, ClothoAQA) and three runtime conditions: Individual (dataset-specific), Sequential (worst-case serialized execution), Parallel (optimal concurrent execution) and Total (complete execution). AU-Harness consistently outperforms existing toolkits across all scenarios, with most significant gains in parallel and total execution, demonstrating effective utilization of concurrent processing capabilities.

Table 4, the naive vLLM design improves throughput relative over HuggingFace-based counterparts, achieving the 0.98 gains of average SPS. However, it remain less efficient than our proposed AU-Harnesswith a throughput deficit of 0.41 of average SPS. Under the realistic settings of concurrent task processing (*Reality*), we observe more significant gap when compared to LMMS-Eval (1.41 points in RTF reduction and 0.56 points of improvement in average SPS).

EvalKit	RTF (↓)	Samples Processed per Second (↑)
LMMS-Eval-HF	8.6	1.54
LMMS-Eval-vLLM (Optimal)	3.58	2.52
LMMS-Eval-vLLM (Reality)	5.51	1.64
AU-Harness (Optimal)	3.07	2.93
AU-Harness (Reality)	4.10	2.20

Table 4: **Parallel runtime efficiency comparison between AU-Harness and LMMS-Eval.** We conduct controlled experiment following the previously presented setups in Section 5.1. As both LMMS-Eval and AU-Harness support multi-task parallel evaluation setups, besides the *Optimal* parallel runtime, we report *Reality* parallel runtime variant where additional realistic overheads are taken into account during evaluation.

5.2 Instruction Modality Gap

When text-based benchmarks are converted to the audio-based counterparts, the impact of instruction modality is often overlooked. However, this distinction can have a significant impact on the downstream task evaluation performance, especially for more complex instruction-following tasks. As observed in Table 5, leveraging audio instruction modality instead text can have a major impact on the performance evaluation. For instance, on

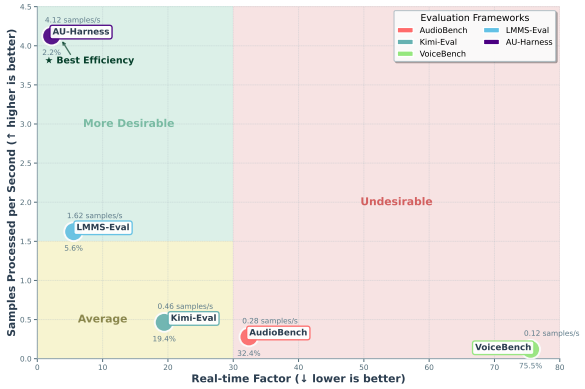


Figure 4: **Total runtime efficiency analysis across evaluation frameworks.** Scatter plot comparing frameworks under optimal parallel execution conditions, plotting Real-time Factor (x-axis, ↓ better) against Samples Processed per Second (y-axis, ↑ better). Our framework (top-left-most cluster) achieves superior performance in both dimensions, demonstrating the effectiveness of token-based request scheduling, dataset sharding, and vLLM integration for large-scale LALM evaluation.

Instruction Modality	IFEval (↑)	MTBench (↑)
Text	87.56	81.06
Audio	80.39	76.19

Table 5: **Empirical evaluations to assess the impact of different instruction modalities on complex instruction following tasks with proprietary Qwen3-Omni-30B-A3B-Thinking** reveals the significant performance gap between Audio and Text instructions, highlighting the need for a more thorough investigation when instruction-following benchmarks are converted from text to audio.

challenging task of Instruction Following (i.e. IFEval), we observe a performance degradation of up to 7.2 points. This observation highlights a potential core limitation of the contemporary LALMs in following audio instructions. Therefore, a careful and thorough reassessment of different instruction modality is needed to accurately measure a model's true reasoning capabilities in a multimodal context.

6 Conclusion

We introduced a modular and extensible evaluation framework for large audio-language models that emphasizes broad task coverage, ease of use, and adaptability. Its modular design enables researchers and practitioners to extend the codebase, customize benchmarks, and integrate new models or tasks without major restructuring. The efficiency gains of our AU-Harness are realized through the aggregation of dataset sharding and effective token request orchestration. More importantly, the broader value of our framework lies in enabling flexible, large-scale evaluations that were previously difficult to conduct in a reproducible and accessible manner. By lowering the barrier to benchmarking and fostering customization, we aim to support both systematic research and practical deployment, contributing a more standardized and transparent evaluation ecosystem for LALMs.

Limitations

Backend dependency and reproducibility. Our efficiency gains are evaluated with vLLM integration; hence, models without mature backends might revert to conventional execution with reduced throughput. Support for closed-source endpoints depends on chat-completions APIs, limiting batching control and introducing provider rate limits. Even with deterministic configs, runs may vary due to endpoint queuing and transient failures, requiring documentation of capacity and request budgets for cross-institutional comparability.

Standardization vs. task fidelity. Standardized prompting improves reproducibility but cannot eliminate prompt sensitivity. For open-ended tasks, canonical prompts may bias results toward specific behaviors. The community needs multiple documented prompt families and complementary temporal measures to triangulate performance fairly.

Evaluation Framework vs Benchmark: Our work presents a unified and efficient evaluation framework targeting specifically for LALM evaluation. Despite the wide coverage presented in Table 3, our ultimate objective is to create an extensible framework where new benchmarks, tasks and metrics can be seamlessly integrated.

Coverage and generalization gaps. Our coverage remains skewed toward English and common domains. Environmental audio, music understanding, and low-resource languages are underrepresented. Moreover, the relationship between standardized benchmark performance and real-world audio-language capabilities where contexts are noisier, more diverse, and less structured requires further empirical validation.

These limitations highlight challenges in audio-language evaluation. Achieving reproducible, comprehensive, and valid assessment requires community coordination around prompting standards, temporal diagnostics, and multilingual breadth. Our framework is designed to enable practical, systematic progress in these areas across the broader ecosystem.

Ethics Statement

Our work focuses on responsible development of audio language model evaluation infrastructure. We have taken care to ensure that all audio datasets used in our benchmarks respect copyright and privacy guidelines, with particular attention to speaker consent in diarization tasks. While our framework enables large-scale evaluation of LALMs, we cannot guarantee that models evaluated through AU-Harness will not generate harmful or biased audio-related outputs. Researchers and practitioners are strongly encouraged to implement appropriate content filtering and bias detection when deploying LALMs in production environments. Our speech synthesis components for creating reasoning benchmarks use only publicly available, ethically sourced voice models. Additionally, we acknowledge that our current task coverage is skewed toward English and common domains, which may inadvertently reinforce existing representational biases in audio AI systems. We encourage the community to extend our framework to include more diverse languages and cultural contexts.

Regarding language model usage in manuscript preparation, we utilize them solely to refine the language used in paper to improve clarity and correctness, without generating any substantial content or claims.

Reproducibility Statement

We are committed to full reproducibility of our evaluation framework and experimental results. All AU-Harness code, configuration files, evaluation scripts, and documentation will be publicly released under an open-source license upon acceptance. We provide comprehensive implementation details including all hyperparameters, model endpoints, dataset preprocessing steps, and evaluation metrics in our appendices. For efficiency comparisons, we document exact hardware specifications, vLLM versions, concurrent request limits, and retry policies used across all experiments. Our newly introduced reasoning benchmarks include complete details on text-to-speech synthesis parameters and prompt templates. To ensure consistent reproduction, we provide Docker containers with fixed dependency versions and detailed setup instructions for multi-node evaluation. All random seeds, sampling parameters, and LLM-as-judge configurations are specified to enable identical result replication across different research groups.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Carlos Arriaga, Alejandro Pozo, Javier Conde, and Alvaro Alonso. 2024. Evaluation of real-time transcriptions using end-to-end asr models. *arXiv preprint arXiv:2409.05674*.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, and 1 others. 2024a. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024b. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, and 1 others. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, and 1 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, Junbo Zhang, and Jian Luan. 2025. Mecat: A multi-experts constructed benchmark for fine-grained audio understanding tasks. *Preprint*, arXiv:2507.23511.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Sebastián Ramirez. 2018. Fastapi. `https://fastapi.tiangolo.com/`.

Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, and 1 others. 2025. Versa: A versatile evaluation toolkit for speech, audio, and music. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 191–209.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025b. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Lu Wang, Hao Chen, Siyu Wu, Zhiyue Wu, Hao Zhou, Chengfeng Zhang, Ting Wang, and Haodi Zhang. 2025c. Codecbench: A comprehensive benchmark for acoustic and semantic evaluation. *Preprint*, arXiv:2508.20660.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. Superb: Speech processing universal performance benchmark. *Interspeech 2021*.

Junbo Zhang, Heinrich Dinkel, Yadong Niu, Chenyu Liu, Si Cheng, Anbei Zhao, and Jian Luan. 2025a. X-ares: A comprehensive framework for assessing audio encoder performance. *Preprint*, arXiv:2505.16369.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2025b. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Appendix

A.1 Comprehensive Audio Evaluation

A.1.1 Benchmark Details

We present a comprehensive benchmark suite comprising 56 diverse datasets spanning six funda-

mental task categories in audio and speech understanding. Our benchmark encompasses *Audio Understanding* (6 datasets), evaluating models' capabilities in audio scene analysis and music comprehension; *Paralinguistics* (12 datasets), assessing speech characteristics including emotion, gender, accent recognition, and speaker-related tasks; *Safety and Security* (2 datasets), examining robustness against adversarial inputs and spoofing; *Spoken Language Reasoning* (5 datasets), testing complex reasoning abilities from mathematical problem-solving to code generation from speech; *Spoken Language Understanding* (21 datasets), the largest category covering speech question-answering, intent classification, and translation tasks; and *Speech Recognition* (15 datasets), establishing baselines for automatic speech recognition across multiple languages and acoustic conditions.

A.1.2 Metric Details

- **Word Error Rate (WER)** – Measures automatic speech recognition (ASR) errors via insertions and deletions in transcribed text. Lower is better.
- **LLM-Judge (MJ)** – LLM-based evaluation of response quality. Higher is better. Reported metrics:
  - **Binary (LB)** – Binary LLM-based pass/fail correctness judgment.
  - **Detailed (LD)** – Detailed multi-level llm judgement across multiple dimensions.
  - **BigBench Audio (LBBA)** – LLM-based evaluations for BigBench-like audio tasks.
  - **RedTeaming (SafetyJudge)** – LLM-based evaluations for red-teaming and safety.
  - **MT-Bench (MTJudge)** – LLM-based evaluation for multi-turn systems.
- **BLEU** – N-gram overlap score for comparing generated and reference text. Higher is better.
- **Instruction Following Score (IFScore)** (Zhou et al., 2023) – Measuring instruction following capability in natural language tasks via averaging accuracy across (1) strict-prompt, (2) strict-instruction, (3)loose-prompt and (4) loose-instruction scenarios.
- **Joint Goal Accuracy (JGA)** - A strict holistic measure for dialogue state tracking systems requiring a perfect alignment between predicted states and reference states for every slot-value pair at every single turn in multi-turn conversations.

A.2 Inference Efficiency Evaluation Settings

To provide a comprehensive and fair comparison with other evaluation kits, regardless of their underlying implementation, we introduce two additional runtime scenarios beyond individual dataset runtimes, namely *Sequential* and *Parallel*. First, *Sequential* runtime represents the most inefficient runtime by assuming each benchmark is executed in a sequential manner, where no data or model parallelization algorithms are introduced. On the other hand, *Parallel* presents the theoretical upper-bound for optimal runtime. The final runtime is calculated by taking the longest runtime among all evaluated datasets. This scenario presumes an ideal, zero-overhead parallelization environment where communication protocols among parallel processes and other overheads do not impact the runtime. This is considered a best-case runtime for our framework and existing evaluation kits across all presented datasets and models.

In our experimental settings, for fair comparison , we allocate 3xH100 GPUs to all of the evaluation frameworks and maximize the throughput designed by the frameworks either through multi-processing or supported concurrent parallel multi-task evaluations.

A.3 Contemporary Evaluation Kits

There are a few evaluation kits that we have built upon and been inspired by, both in evaluation framework design and task coverage.

- **AudioBench** (Wang et al., 2025a): A comprehensive open-source audio evaluation framework encompassing eight core tasks and more than twenty-six curated datasets, with coverage continuing to expand. AudioBench supports both open and closed-source models and provides standardized evaluation pipelines using conventional metrics such as Word Error Rate (WER) and METEOR, alongside LLM-as-a-judge scoring for instruction-following and reasoning tasks.
- **Kimi-Eval** (Ding et al., 2025): A multilingual and multi-model evaluation suite designed to assess leading Chinese and English large language models, including the

Table 6: **Comprehensive Audio and Speech Datasets Overview.** Listing of 56 datasets across 6 task categories: Speech Recognition, Paralinguistics, Audio Understanding, Spoken Language Understanding, Spoken Language Reasoning, and Safety & Security.

Task Category	Task Type	Dataset Name	Task	Description	License
Speech Recognition	ASR	AISHELL-1	1	High-quality Mandarin speech recognition dataset	Apache 2.0
	ASR	AMI Meeting Corpus	2	Multispeaker meeting recordings for ASR and diarization	CC BY 4.0
	ASR	CallHome	5	Conversational speech corpus across multiple languages	LDC User Agreement for Non-Members
	ASR	Common Voice	100	Crowdsourced multilingual speech dataset from Mozilla	CC0 1.0 Universal
	ASR	FLEURS EN-US	102	Multilingual speech dataset for ASR and translation	CC BY 4.0
	ASR	GigaSpeech	1	Large-scale audio and transcription corpus for end-to-end ASR	Apache 2.0
	ASR	GigaSpeech2	2	Large-scale audio and transcription corpus for end-to-end ASR (v2)	Apache 2.0
	ASR	LibriSpeech	2	Audiobook-derived speech corpus with clean and noisy subsets	CC BY 4.0
	ASR	Multilingual LibriSpeech (MLS)	7	Extension of LibriSpeech with multiple European languages	CC BY 4.0
	ASR	MNSC	6	Large-scale multitask speech corpus	MNSC: Publicly released
	ASR	People's Speech	1	Large-scale open-source English speech recognition dataset	CC-BY-SA
	ASR	SPGISpeech	1	Transcriptions of financial meeting recordings	Kensho User Agreement
	ASR	TEDLIUM3	1	Transcribed TED talks for ASR and speaker adaptation	CC BY-NC-ND 3.0
	ASR	VoxPopuli	17	Multilingual speech corpus from European Parliament recordings	CC0
	Code-switching ASR	SEAME	2	Mandarin-English code-switching speech dataset	LDC2015S04
	Long-form ASR	TEDLIUM3	1	Transcribed TED talks for ASR and speaker adaptation (long-form version)	CC BY-NC-ND 3.0
	Long-form ASR	Earnings21	1	Long-form earnings call dataset for speech recognition	CC-BY-SA-4.0
	Long-form ASR	Earnings22	1	Long-form earnings call dataset for speech recognition	CC-BY-SA-4.0
Paralinguistics	Accent Recognition	MNSC AR Dialogue	1	Dataset for accent recognition in dialogue speech	MNSC: Publicly released
	Accent Recognition	MNSC AR Sentence	1	Dataset for accent recognition in sentence-level speech	MNSC: Publicly released
	Accent Recognition	VoxCeleb Accent	1	Speech dataset with diverse speakers for accent recognition	CC BY 4.0
	Emotion Recognition	IEMOCAP Emotion	1	Multimodal dataset for emotion recognition in speech	GPL-3.0
	Emotion Recognition	MELD Emotion	1	Multi-party conversation dataset for emotion recognition	GPL-3.0
	Emotion Recognition	MELD Sentiment	1	Multi-party conversation dataset for sentiment analysis	GPL-3.0
	Gender Recognition	IEMOCAP Gender	1	Multimodal dataset for gender recognition in speech	GPL-3.0
	Gender Recognition	MNSC GR Dialogue	1	Dataset for gender recognition in dialogue speech	MNSC: Publicly released
	Gender Recognition	MNSC GR Sentence	1	Dataset for gender recognition in sentence-level speech	MNSC: Publicly released
	Gender Recognition	VoxCeleb Gender	1	Speech dataset with diverse speakers for gender recognition	CC BY 4.0
	Speaker Recognition	MMAU-mini	1	Multi-modal audio dataset for speaker recognition	Apache 2.0
Audio Understanding	Music Understanding	MuChoMusic	1	Benchmark for music understanding for LALMs	CC-BY-SA-4.0
	Scene Understanding	AudioCaps	1	Large-scale dataset for open-domain audio captioning	MIT
	Scene Understanding	AudioCaps QA	1	Dataset for question answering over natural audio scenes	MIT
	Scene Understanding	Clotho AQA	1	Dataset for answering natural-language questions about audio signals	MIT
	Scene Understanding	WavCaps	1	Large-scale weakly labeled dataset for audio captioning	CC-BY-NC 4.0
	Scene Understanding	WavCaps QA	1	Large-scale dataset for audio question answering	CC-BY-NC 4.0
Spoken Language Understanding	Intent Classification	SLURP	1	Multi-domain spoken dialogue understanding benchmark	CC BY-NC 4.0
	Speech QA	Alpaca Audio	1	Speech dataset for question answering with audio instructions	Apache-2.0
	Speech QA	CN College Listen MCQ	1	Multispeaker dataset for listening-based multiple-choice questions	MERaLiON Public License
	Speech QA	Dream TTS MCQ	1	Dialogue-based multiple-choice comprehension dataset with audio	MIT
	Speech QA	MNSC SQA	4	Benchmark for reasoning and understanding in spoken language	NSC License
	Speech QA	OpenHermes	1	Speech dataset for question answering with audio instructions	CC-BY-NC
	Speech QA	Public-SG	1	Speech question answering benchmark	NSC License
	Speech QA	SLUE SQA	1	Spoken Language Understanding Evaluation benchmark	CC-BY-4.0
	Speech QA	Spoken Squad	1	Speech dataset for extraction-based question answering	CC-BY-SA-4.0
	SQQA	Big Bench Audio	2	Benchmark for reasoning with audio and text input	MIT
	SQQA	MMSU	12	Multi-choice question answering dataset	Apache-2.0
	SQQA	OpenBookQA	1	Multi-choice question answering dataset	Apache-2.0
	SQQA	SD-QA	22	Multi-choice question answering dataset	Apache-2.0
	Translation	CoVoST2 (zh→en)	36	Large-scale multilingual dataset for speech translation	CC-BY-NC-4.0
Spoken Language Reasoning	Speech Instruction Following	IFEVAL	2	Speech dataset for complex instruction following	Apache-2.0
	Speech Instruction Following	MTBench	2	Speech dataset for multi-turn instruction following	Apache-2.0
Safety & Security	Safety	Advbench	1	Speech dataset for testing resistance to adversarial or harmful prompts	Apache 2.0
	Spoofing	ASVpoof2017	1	Speech dataset for spoofing attack detection in real-world conditions	CC BY-NC 4.0
		Total Tasks	**363**		

Table 7: **Experimental setup for efficiency comparison across evaluation frameworks.** We conduct controlled experiments using 500 samples from three diverse datasets: MELD-Emotion (short emotional speech), LibriSpeech-clean (medium-length read speech), and ClothoAQA (long-form descriptive audio). Total audio duration varies from 1,476 to 11,376 seconds, enabling assessment across different audio characteristics and evaluation modalities (LLM-judge vs. traditional metrics).

	MELD-Emotion	Librispeech-clean	ClothoAQA
# Samples	500	500	500
Audio Duration (seconds)	1,476	3,780	11,376
Evaluation Metric	LLM-Judge	WER	LLM-Judge

Baichuan series, Qwen, GLM, and Kimi itself. The benchmark spans automatic speech recognition (ASR), multiple choice question answering (MQA), open question answering (OpenQA), and reference-based question answering (RefQA), enabling a broad assessment of both comprehension and generative audio capabilities.

- **VoiceBench** (Chen et al., 2024): A focused benchmark evaluating thirty-five-plus state-of-the-art speech models across seven carefully selected datasets. While the total number of datasets is smaller than in AudioBench, the high task complexity and distinctive challenge of each dataset provide a useful test suite.

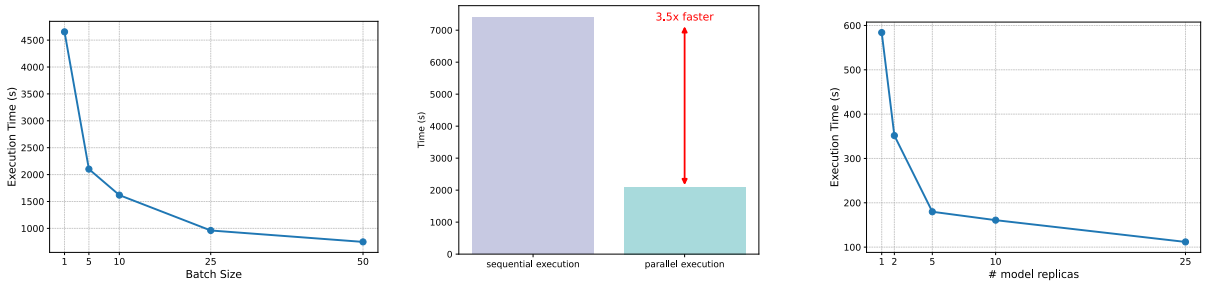
- **LMMS-Eval** (Zhang et al., 2025b): A comprehensive evaluation kit designed to assess multimodal frontier models across vision, audio and video modalities. Despite its broad coverage, audio-centric evaluation is comparatively limited as compared to other modalities.

A.4 Inference Efficiency Ablations

To assess the scalability and efficiency of AU-Harness, we conduct three controlled ablations: (a) varying batch size, (b) throughput gains from parallel execution, and (c) latency trade-offs with replica scaling. The experimental setup follows Table 7, except for (c), where we use the full LibriSpeech-clean dataset to ensure sufficient workload for scalability analysis.

Figure 5 presents the results. Increasing batch size reduces execution time substantially, though benefits taper off at higher scales. Parallel execution yields up to a 3.5× improvement in throughput over sequential execution, confirming the efficiency of concurrent scheduling. Replica scaling further lowers latency, with near-linear improvements observed up to 25 replicas.

Overall, these ablations highlight that AU-Harness is both scalable and adaptable. By leveraging batching, parallelism, and replica scaling, it can be tuned for diverse deployment scenarios ranging from high-throughput evaluation to low-latency inference.



(a) Batch size effect (b) Throughput scaling (c) Latency trade-offs

Figure 5: **Inference efficiency ablations in AU-Harness.** We examine three factors: (a) impact of batch size on execution time, (b) throughput gains from parallel execution, and (c) latency reduction through replica scaling.