Massive Sound Embedding Benchmark (MSEB)

Georg Heigold
Google Research
Munich, Germany
heigold@google.com

Ehsan Variani
Google Research
California, USA
variani@google.com

Tom Bagby
Google Research
California, USA
tombagby@google.com

Cyril Allauzen
Google Research
New York, USA
allauzen@google.com

Ji Ma
Google Research
London, UK
maji@google.com

Shankar Kumar
Google Research
New York, USA
shankarkumar@google.com

Michael Riley
Google Research
New York, USA
riley@google.com

Abstract

Audio is a critical component of multimodal perception, and any truly intelligent system must demonstrate a wide range of auditory capabilities. These capabilities include transcription, classification, retrieval, reasoning, segmentation, clustering, reranking, and reconstruction. Fundamentally, each task involves transforming a raw audio signal into a meaningful 'embedding'—be it a single vector, a sequence of continuous or discrete representations, or another structured form—which then serves as the basis for generating the task's final response. To accelerate progress towards robust machine auditory intelligence, we present the Massive Sound Embedding Benchmark (MSEB): an extensible framework designed to evaluate the auditory components of any multimodal system. In its first release, MSEB offers a comprehensive suite of eight core tasks, with more planned for the future, supported by diverse datasets, including the new, large-scale Simple Voice Questions (SVQ) dataset. Our initial experiments establish clear performance headrooms, highlighting the significant opportunity to improve real-world multimodal experiences where audio is a core signal. We encourage the research community to use MSEB to assess their algorithms and contribute to its growth. The library is publicly hosted at `https://github.com/google-research/mseb`.

1 Introduction

Audio is a critical component of multimodal perception, and any truly intelligent system must demonstrate a wide range of auditory capabilities. These capabilities are extensive, including transcription of speech, classification of sounds, retrieval of information from spoken queries, reasoning over provided contexts, segmentation of key moments, clustering of similar acoustic objects, reranking of hypotheses, and even generative tasks like audio reconstruction. While these capabilities appear distinct, they can be unified by a fundamental process: transforming a raw audio signal into a meaningful 'embedding'—be it a continuous vector, a sequence of discrete tokens, or another structured representation—which then serves as the basis for generating the task's final response. This conceptual framework allows for a systematic evaluation of various modeling paradigms, from traditional cascaded systems to end-to-end multimodal architectures.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Historically, different applications have evolved specialized forms of embeddings. Technologies involving human speech, like Voice Search [1, 2] or Voice Assistants [3], have focused on creating discrete, textual embeddings through automatic speech recognition (ASR). In contrast, monitoring technologies such as Speaker Verification [4, 5] or general sound classification [6] have concentrated on learning dense, continuous vector representations. While effective for their specific purposes, this fragmentation has made it difficult to assess progress toward a more generalized form of machine auditory intelligence.

Thorough research in this direction requires more than just new models; it demands a standardized way to measure their holistic impact. Specifically, we need to determine: 1) the current state of end-to-end performance at specific compression rates and computational complexities; 2) the maximum performance headroom available; and 3) the existence of general-purpose audio embedding techniques capable of capturing rich acoustic information relevant to multiple tasks. To enable this research, we need a modern benchmark with diverse task coverage, mirroring the role of the massive text embedding benchmark (MTEB) [7] for text and similar benchmarks for images [8, 9]. We introduce the Massive Sound Embedding Benchmark (MSEB) to fill this critical gap, providing a comprehensive forum for the community to compare methodologies and push the boundaries of sound understanding in meaningful, real-world contexts.

Our work makes several key contributions. First, MSEB introduces a suite of eight crucial tasks that cover a wide spectrum of sound-centric technologies, including retrieval, reasoning, and reranking, which have been underrepresented in prior benchmarks. Second, we provide an extensible, open-source library designed to support diverse modeling paradigms relevant to multimodal systems, allowing researchers to seamlessly evaluate their own algorithms. Third, we introduce the Simple Voice Questions (SVQ) dataset, a novel, large-scale resource featuring spoken queries across 26 locales and 17 languages, uniquely designed to support evaluation across multiple tasks from a single source. Finally, our initial experiments establish clear performance headrooms, highlighting the significant opportunity to improve real-world multimodal experiences where audio is a core signal.

MSEB is distinct from prior benchmarks like HARES [10], NOSS [11], and SUPERB [12] in four key ways: 1) it prioritizes tasks with direct real-world technological applicability; 2) it emphasizes measuring embedding efficiency via compression ratio and computational complexity; 3) it includes critical, user-centric tasks such as retrieval and reasoning; and 4) its evaluation methodology does not rely on task-specific fine-tuning, offering a true test of an embedding's generalizable capabilities. We envision MSEB as a dynamic platform and encourage the community to contribute data and tasks to facilitate collaborative advancement in sound technology.

2 Massive Sound Embedding Benchmark (MSEB)

The Massive Sound Embedding Benchmark (MSEB) is designed to be the core platform for evaluating the auditory capabilities of any intelligent system, regardless of whether it processes sound as a single modality or as part of a broader multimodal context. MSEB operationalizes these capabilities through defined tasks, carefully selected as verifiable proxies for core challenges in widespread applications. For each task, the model being evaluated must provide a prediction for a given input object—which includes the sound signal and potentially other modalities—drawn from our curated datasets. A fundamental principle of MSEB is verifiability: every task is grounded in established ground truth and a standardized evaluation methodology. To facilitate widespread adoption, the benchmark is supported by a flexible library designed to streamline the evaluation of diverse modeling approaches, from conventional uni-modal models and cascade systems that convert between modalities, to native multimodal models such as Large Language Models (LLMs). While the long-term vision of MSEB encompasses the full spectrum of auditory intelligence, the initial release prioritizes eight tasks with demonstrated real-world utility in rapidly advancing technologies such as voice search, intelligent assistants, and bioacoustic monitoring.

2.1 Datasets

A core tenet of MSEB is that robust evaluation requires high-quality, accessible data. We curate our datasets based on five rigorous criteria: (a) public availability with easy-to-use licensing for research; (b) data complexity, ensuring derived tasks are non-trivial and not easily solved by simple baselines; (c) large scale, sufficient to report performance metrics with high statistical confidence;

(d) high-quality labeling, providing verifiable ground truth; and (e) diverse coverage, spanning a reasonable portion of the vast spectrum of sound.

For its initial release, MSEB includes four datasets meeting these criteria, all currently featuring sound and text modalities. We are actively working to expand this coverage to include other modalities, such as images.

Simple Voice Questions (SVQ): We introduce SVQ as a novel, large-scale dataset specifically collected for MSEB. It comprises over 177k short spoken queries across 26 locales and 17 languages, recorded in four distinct environments (clean, background speech, traffic, and media noise). Each query is paired with rich textual metadata derived from the XTREME-UP benchmark [13] (including aligned Wikipedia pages and passages), as well as comprehensive speaker information and fine-grained temporal annotations. Uniquely, SVQ is released as a single, undivided collection. Extensive details on data collection, specific split definitions, and statistical distributions are provided in Appendix A.

Speech-MASSIVE [14]: This multilingual Spoken Language Understanding (SLU) dataset extends the text-based MASSIVE corpus [15] with spoken utterances across 12 languages. It features highly granular annotations, covering 18 domains, 60 intents, and 55 slots. We utilize its established test split for standard benchmarking.

FSD50K [16]: To cover general sound events, we include FSD50K, containing over 51k clips totaling over 100 hours of audio. It features established development and evaluation splits meticulously designed to prevent data leakage, and uses 200 sound classes drawn directly from the AudioSet ontology [6]. We utilize the standard evaluation split for all MSEB tasks.

BirdSet [17]: For bioacoustics, BirdSet provides a massive-scale collection with over 6,800 hours of training data covering nearly 10,000 bird species. It offers seven distinct test sets featuring over 400 hours of soundscape recordings derived from Passive Acoustic Monitoring (PAM) devices. These recordings present realistic challenges like high background noise and overlapping calls. Crucially, they come with strong, temporally precise labels that include rich metadata such as bird type, call type, gender, and geospatial coordinates. MSEB utilizes all seven of these test sets.

2.2 Tasks

The initial release of MSEB introduces eight foundational super tasks Figure 1, selected to provide comprehensive coverage of the auditory capabilities expected of a modern intelligent system. To ensure granular and robust assessment, each super task is composed of numerous specific sub-variations, which we refer to simply as tasks. Structurally, every task defines a strict input—comprising the primary audio signal and potentially side information from other modalities—and a required output format. These tasks are selected for their direct alignment with realistic, high-utility applications deployed at scale.

The super tasks are presented below, ordered by a logical progression from user-centric Information Access (Retrieval, Reranking, Reasoning), to fundamental Core Perception (Classification, Transcription, Segmentation), and finally to higher-level Organization Generation (Clustering, Reconstruction).

Retrieval (Finding information): This super task mimics the human process of answering questions by consulting a vast body of knowledge. It serves as a verifiable proxy for Voice Search, where a model is given a spoken query and has access to a "knowledge index"—a bank of information in the form of documents. The task is to find the most relevant information for the query. The target information may be present in the index, or the model may need to correctly determine that no answer is available. We define six sub-tasks based on the granularity of the search and the language match between the query and the index. When searching for a whole document (like a web page) in the same language as the query, the task is DocumentInLang. This is mirrored by PassageInLang (searching for a passage) and SpanInLang (searching for a span of words). Alternatively, a model might hear the question in one language but search across a corpus in another, which defines the DocumentCrossLang, PassageCrossLang, and SpanCrossLang tasks. The SVQ dataset enables evaluation of this super task across all six sub-tasks, 26 language locales, and four different recording environments. It provides a knowledge index (a subset of Wikipedia) and, for each audio query, the corresponding ground-truth pages, passages, and spans.

Reranking (Refining information): Intelligent systems, much like humans, often encounter ambiguity where initial perception yields multiple plausible interpretations—for example, when speech is obscured by noise or contains phonetically confused terms. The Reranking super task serves as a proxy for resolving this ambiguity, requiring models to reorder a provided list of candidate hypotheses based on their actual relevance to the original audio signal. This capability is a core component of many practical systems, used to refine candidate documents in search engines or to select the best transcription from an N-best list in automatic speech recognition. In the current MSEB release, we focus on Acoustic Hypothesis Reranking. Leveraging the SVQ dataset, each audio query is paired with a set of text candidates that sound similar to the true utterance. The task is to utilize the audio signal to sort these hypotheses by their true relevance to what was actually spoken.

Reasoning (Extracting precise answers): Profound understanding requires more than just locating relevant documents; it demands synthesizing information to derive precise answers. This super task serves as a proxy for next-generation Intelligent Assistants. Assuming relevant knowledge has already been retrieved (e.g., a passage or document), the Reasoning task challenges models to analyze this provided multimodal context in response to a spoken query and pinpoint the exact text span that satisfies the user's information need. Crucially, this requires the robust capability to correctly determine when the provided context contains no answer. In the current MSEB release, leveraging the SVQ dataset, we define two sub-tasks based on language alignment: SpanInLang, where the spoken query and text context share a language, and SpanCrossLang, where they differ.

Classification (Identifying sounds): Upon hearing a sound, an intelligent system might need to immediately identify its source, the characteristics of its speaker, or even the intent behind it. This super task evaluates a model's ability to categorize an audio signal into predefined classes based on varied acoustic features. While the potential scope is vast, the initial release of MSEB focuses on four key areas: speaker properties, user intent, environmental sounds, and bioacoustics. For speaker classification (identity, gender, age) and recording environment classification, we utilize the SVQ dataset, which provides rich labels across 26 language locales. Intent classification is evaluated using the Speech-MASSIVE dataset, offering a standard benchmark across 12 languages. For general environmental sound classification, we employ FSD50K with its 200 AudioSet-derived classes. Finally, bioacoustic classification is assessed using the massive-scale BirdSet benchmark.

Transcription (Converting sound to text): Upon hearing a sound, a natural response is to transcribe it into a textual representation. While this is most commonly associated with Automatic Speech Recognition (ASR) for spoken language, the concept extends to describing any auditory event in text. This super task evaluates how effectively an intelligent system can translate audio input into a verbatim or descriptive textual output. In the current MSEB release, we focus on ASR, utilizing the SVQ dataset to assess transcription quality across 26 language locales. Furthermore, SVQ's four distinct recording environments allow for a precise comparison of a model's robustness to varied acoustic conditions, such as background noise or media.

Segmentation (Locating key information in time): Often, we don't need to hear an entire audio recording; we only need the key information. This super task mimics that need by evaluating a model's ability to identify the most salient moments within an audio signal and pinpoint their precise timestamps. In the current MSEB release, this task is defined using the SVQ dataset across all its 26 language locales. For each audio query, we have identified the top-three most salient terms and their corresponding start and end times. The goal for a model is to correctly predict both these key terms and their exact temporal locations.

Clustering (Organizing unknown sound): Beyond just identifying known categories, a truly intelligent system must be able to organize unknown, unstructured sound data into a coherent structure. This super task evaluates a model's ability to discover latent structure by grouping audio segments that share similar properties into homogeneous clusters, without relying on predefined labels. In the initial MSEB release, we define sub-tasks across different domains: speaker, gender, and age clustering using the SVQ dataset; environmental sound clustering using FSD50K; and bioacoustic clustering using BirdSet.

Reconstruction (Generating sounds): True mastery of a modality implies the ability not just to understand it, but to generate it. This super task evaluates a system's generative capabilities by requiring it to reconstruct the original audio signal solely from its internal embedding representation. High-fidelity reconstruction serves as a rigorous test that an embedding retains essential acoustic details—beyond just high-level semantics—vital for applications like neural audio compression

Figure 1: The eight MSEB super tasks, spanning Information Access (Retrieval, Reasoning, Reranking), Core Perception (Transcription, Classification, Segmentation), and Organization Generation (Clustering, Reconstruction).

and speech synthesis. In the initial release, we focus on reconstructing speech in diverse noise environments using the SVQ dataset, alongside general environmental sound reconstruction using FSD50K dataset.

2.3 Evaluators

The MSEB evaluator is designed to be strictly model-agnostic. We assume only that a model generates some form of representation for each input example. The evaluator's primary role is to standardise assessment by mapping these representations—optionally, if they are not already in the task's target space—to that space to calculate metrics. While multiple metrics are computed for granular analysis, one primary metric is designated for each task for leaderboard reporting. Beyond task-specific performance, MSEB universally measures efficiency for all entries using two key metrics: Compression Ratio (CR), quantifying the memory footprint of the representation relative to the original audio signal, and Computational Complexity, measured in floating-point operations per second (FLOPS). The following sections detail the specific evaluation protocols for each super task.

Retrieval, the evaluator ranks all candidates in the knowledge index based on their proximity to the query in the representation space. The primary metric is Mean Reciprocal Rank (MRR), which rewards models that place the correct target higher in the ranked list. We also report Exact Match (EM), measuring the frequency with which the correct target is the absolute top-ranked candidate.

Reranking, the evaluator requires the model to assign a relevance score to each item in a provided list of candidates, which are then sorted by these scores. The primary metric is Mean Average Precision (mAP), assessing the overall quality of the ranked list. We also report Mean Reciprocal Rank (MRR) to measure the rank of the most relevant candidate. Furthermore, to evaluate the utility of the reranking for downstream tasks (e.g., selecting the best ASR transcript), we measure the content quality of the top-ranked candidate using Word Error Rate (WER) and Candidate Error Rate (CER).

Reasoning, the model must identify a precise answer span within a provided context, or correctly indicate if no answer exists. The primary metric is F1 score, which treats the predicted and ground truth spans as bags of tokens to measure their overlap. This metric provides a balanced assessment by rewarding both exact matches and partial overlaps, while also effectively evaluating the model's ability to correctly identify unanswerable queries.

Classification, the evaluation strategy adapts to the nature of the task's label space. For standard multi-class problems (single label per example), the primary metric is Accuracy. For multi-label problems, where an example may have multiple simultaneous ground truth labels, the primary metric is Mean Average Precision (mAP) (macro-averaged). To ensure a comprehensive assessment, we also report a diverse set of secondary metrics, including top-k accuracy, balanced accuracy, and weighted F1-scores for multi-class tasks, alongside micro/macro F1-scores, Hamming loss, and subset accuracy for multi-label scenarios.

Transcription, the evaluator measures the divergence between the model's generated text and the verbatim ground truth. The primary metric is Word Error Rate (WER), a standard measure assessing the proportion of word-level errors (substitutions, deletions, insertions) relative to the reference. To provide finer-grained insights, particularly for languages with complex morphology or to assess spelling accuracy, we also report Character Error Rate (CER).

Segmentation, the evaluator assesses the model's ability to identify and localize key events in time. The primary metric is Normalized Discounted Cumulative Gain (NDCG), which evaluates the quality of the predicted sequence by rewarding correct terms found in the correct temporal order. To provide a detailed breakdown of performance, we report three specific accuracy metrics: Content Accuracy (matching the label irrespective of time), Temporal Precision (matching start/end timestamps within a specified tolerance), and the strictest Overall Accuracy (requiring both content and temporal alignment). Additionally, we report Mean Average Precision (mAP) to evaluate detection performance based on confidence scores, and Word Error Rate (WER) to measure overall sequence divergence.

Clustering, the evaluator assesses the inherent structure of the embedding space. Given the ground truth number of clusters, we employ a standard MiniBatch K-Means algorithm to group the provided representations. The primary metric is V-measure [18], a score derived from the harmonic mean of homogeneity and completeness, which evaluates how successfully the embeddings separate distinct ground truth categories (e.g., speakers, acoustic environments) into pure and complete clusters.

Reconstruction, the evaluator assesses the fidelity of the generated audio by comparing its spectrogram to that of the original reference signal. The primary metric is Fréchet Audio Distance (FAD) [19], which measures the distance between the distributions of the original and reconstructed signal embeddings. To provide a comprehensive view of generation quality, we also report Kernel Audio Distance (KAD) [20], an alternative distribution-based metric using Maximum Mean Discrepancy (MMD), and Embedding MSE, which measures the average per-example mean squared error between the embedding frames.

2.4 Library design

MSEB is designed as a flexible library capable of running a diverse set of benchmarks across many types of encoders. The core design principles focus on:

- **Modular Tasks and Encoders**: The library supports modular tasks and encoders that are easy to add and configure.
- **Bulk Encoder Inference**: To accommodate computationally expensive models and large datasets, the library supports bulk inference. Sound datasets often involve significant data sizes and expensive preprocessing (e.g., audio resampling), making efficient bulk processing a critical feature.
- **Diverse Task Types**: MSEB supports a mix of task complexities, ranging from those with expensive setup stages (e.g., retrieval corpus generation) to complex inference steps (e.g., clustering).

The main components and execution flow of the library are illustrated in Figure 2. Task represent instances of a particular benchmark Dataset paired with an Evaluator to score the outputs of an encoder model. The task maps the Dataset (e.g., the Simple Voice Questions dataset) to the multimodal encoder inputs expected by the system (e.g., audio, text, or label pairs).

For bulk inference, the task explicitly exposes the set of inputs that need to be encoded. A `Runner` interface exists to execute the encoder over this set in various ways, such as through a beam pipeline. The output of the runner is a key-value store (the Embedding Cache) mapping unique identifiers used

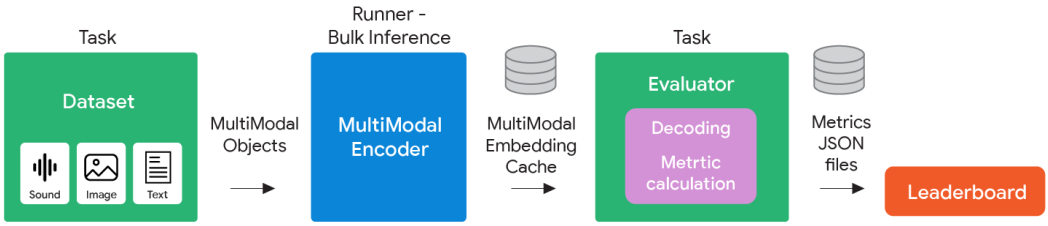

Figure 2: The MSEB library architecture, illustrating the flow from Task Dataset to Leaderboard via bulk inference and evaluation.

by the task to the encoder outputs. This design allows separating bulk encoding steps into multiple stages with different resource requirements.

Encoder models implement a `MultiModalEncoder` interface, which supports an extensible set of input and output types with runtime checks. While currently oriented around Sound and Text modalities, it is designed to be extensible to others, such as images, in future iterations. The definition of an encoder output encompasses many styles of "embedding": fixed-size vectors, sequences of embeddings, or discrete tokens.

Encoders can also be combined; we currently support three main types:

1. **End-to-End Encoder**: Takes multimodal input and directly produces an embedding.
2. **Cascade**: A sequence of encoders chained together (e.g., ASR $\rightarrow$ text embedding).
3. **Collections**: Multi-tower encoders or combinations of encoders for different modalities.

The Evaluator handles the correct usage of these diverse encoder outputs to compute metrics. For instance, in classification, an end-to-end encoder might output a class label text token directly, while a fixed-size vector encoder might need to be paired with label vectors and evaluated using cosine distance.

The core function of the library is to run a benchmark instance of a `(task, encoder)` pair. Registries are provided for listing and selecting these instances. Each benchmark run generates a JSON file containing complete metadata and metrics. Leaderboard submissions are facilitated by submitting pull requests to add these output JSON metric files, which are then collated into the final display.

3 Experiments

We utilized the MSEB framework to establish initial standard baselines and quantify the performance "headroom" across all eight super tasks. These widespread experiments provide a unified view of the current state of machine auditory capabilities, highlighting both existing strengths and significant opportunities for improvement.

Figure 1 presents a comprehensive overview of these results. In this figure, each bar represents the primary metric for a specific task, identified in brackets next to the task name. The bars range from 0 to the metric's maximum value. For bounded metrics (e.g., Accuracy, F1), the bar ends at 1.0. For unbounded metrics (e.g., WER, FAD), a dashed line at the end indicates their open-ended nature, with each internal bin representing one-tenth of the displayed range (the scale for these is noted next to the bar). Furthermore, each bar visualizes performance variability: the labels indicate the specific sound types (e.g., language locales or domains) that achieved the minimum and maximum scores, while the gradient color between them represents the distribution of all other sound types relevant to that task.

To quantify this performance gap, we adopted a unified comparison where applicable. Green bars show performance when sound was used as the primary input to the task, representing typical current systems. Blue bars show performance when the corresponding ground-truth text transcript was used instead, representing an "oracle" scenario with perfect perception. The gap between these bars

quantifies the immediate headroom available for improving auditory intelligence, particularly for tasks that mainly rely on the semantic content of the sound rather than its other acoustic characteristics.

Two main observations can be drawn from this overview: a) significant opportunity for improvement exists across all tasks to reach the maximum possible score, and b) current sound-based approaches significantly lag behind those using the text modality, suggesting that more research is needed to design better sound representations for robust auditory understanding.

Next, we provide further details on the specific baselines evaluated for each super task. For comprehensive comparisons across all sub-tasks, including secondary metrics and the latest submissions, we refer readers to the live MSEB leaderboard and the detailed appendices.

Retrieval: For this super task, we adopted a cascade encoding approach, which remains the state-of-the-art for current industrial voice search systems [21]. Our standard baseline (Sound Input) utilized Whisper Large v3 [22] for automatic speech recognition (ASR), followed by a text embedding model to encode the resulting transcription. We evaluated two distinct text encoders: GeminiEmbedding [23], chosen for its state-of-the-art performance on the MTEB leaderboard at the time of our experiments, and Gecko [24], selected to provide a comparative baseline using a smaller, less powerful text model.

Evaluation was conducted across all SVQ locales. For each example, the spoken query was transcribed and then embedded. Simultaneously, the document index was explicitly encoded using the same text embedding model. Retrieval was performed via dot product between the query embedding and document embeddings, with quality measured by Mean Reciprocal Rank (MRR).

To establish the performance headroom (Text Input in Figure 1), we conducted an "oracle" experiment where ground-truth human transcriptions were fed directly to the text encoders, bypassing the ASR step.

Our primary observations are threefold. First, as shown in Figure 1, sound-based models consistently lag behind their text-only counterparts across all languages, establishing a clear gap for future end-to-end audio models to close. Second, ASR quality (WER) does not perfectly correlate with retrieval performance (MRR) across different languages (detailed in Appendix B). This likely arises because standard ASR objectives treat all words equally, whereas effective retrieval depends disproportionately on semantically salient terms. Third, comparing in-language and cross-language tasks reveals a clear performance gap for both sound and text modalities. This suggests that current embedding vectors have not yet fully achieved universal semantic representation and remain bound by language. Interestingly, this gap is relatively smaller for passage retrieval compared to full-page retrieval, implying that shorter document contexts may facilitate better cross-lingual alignment.

Reranking: For this task, we focused on Acoustic Hypothesis Reranking using the SVQ dataset. Our baseline (Sound Input) employed the same cascade approach as used in Retrieval: audio queries were transcribed by Whisper Large v3, and both these transcripts and the provided candidate hypotheses were embedded using GeminiEmbedding. Candidates were then re-ordered based on their cosine similarity to the query's embedding. The Text Input headroom was established by using the ground-truth transcript as the query for reranking.

Results show extreme variability across locales, heavily correlated with the underlying ASR quality. While English locales achieve high performance (e.g., en-AU at 0.86 mAP), challenging languages like Malayalam (ml-IN) see severe degradation (down to 0.11 mAP). This confirms that current text-based reranking is heavily bottlenecked by the initial 1-best ASR output. A direct sound-based reranker that leverages acoustic nuances lost in transcription could potentially bypass this limitation.

Reasoning: We employed the Gemma 3 foundational model [25] for this task, prompting it to extract precise answer spans from the provided text context. We compared a Sound Input baseline (feeding Whisper ASR transcripts to Gemma) against a Text Input oracle (feeding ground-truth transcripts).

Results indicate significant headroom, with the performance gap heavily dependent on the language and task complexity. While some languages like Arabic show resilience, others suffer dramatic performance losses due to ASR errors. For instance, in Cross-Language reasoning for Malayalam (ml-IN), F1 scores plummet from 0.54 (Text) to just 0.24 (Sound). Even in simpler In-Language scenarios, clear gaps persist (e.g., Bengali drops from 0.66 to 0.59 F1), demonstrating that current ASR systems often fail to preserve the precise semantic details necessary for robust downstream reasoning.

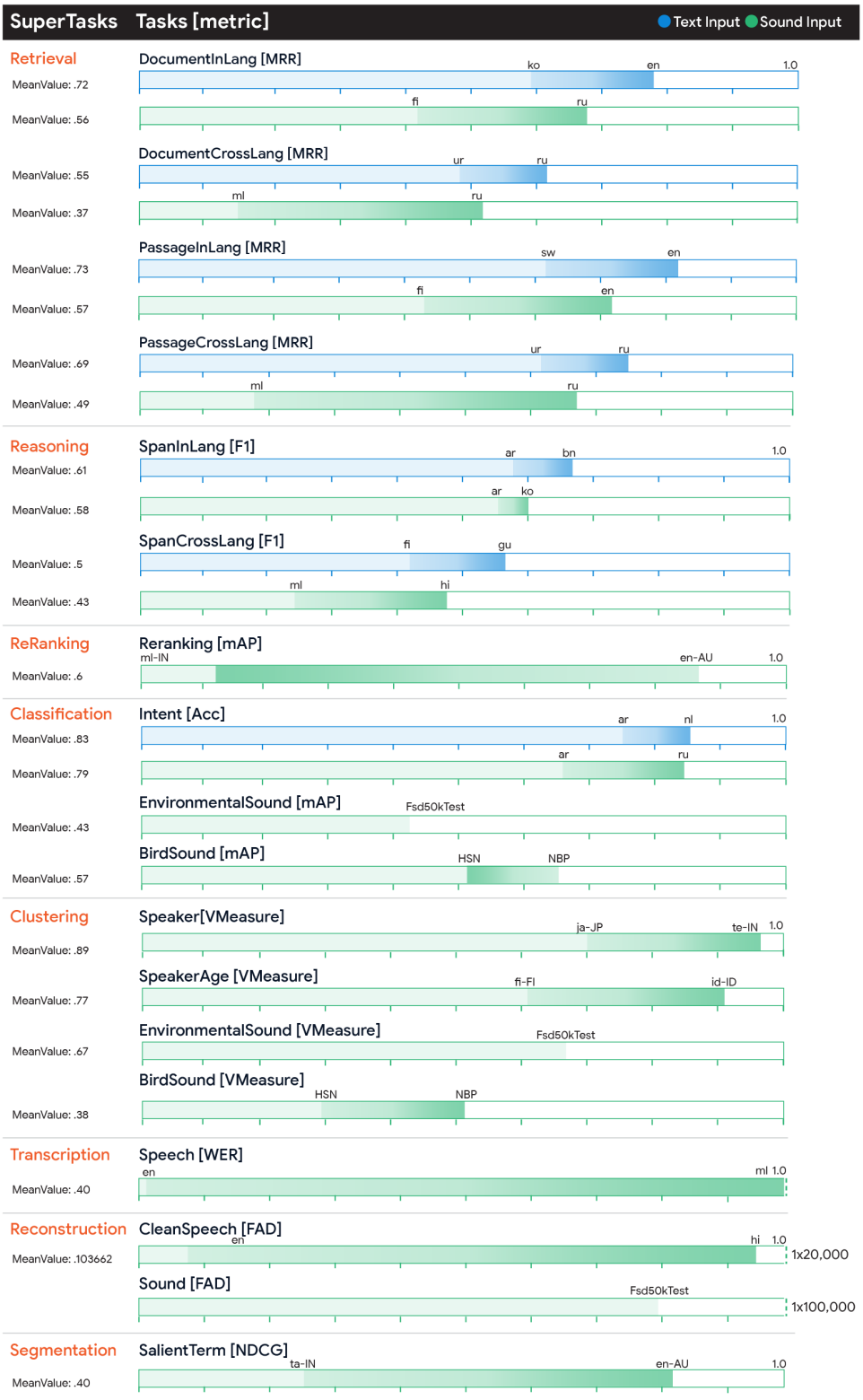

Figure 3: Performance overview of the eight MSEB super tasks. Green bars show performance when sound was used as input to the task, while blue bars show performance using the corresponding ground-truth text transcript. Range markers indicate variability across locales or domains.

Classification: We evaluated diverse modeling strategies across different domains for this task. For Intent Classification on Speech-MASSIVE, we employed the same cascade baseline as in Retrieval (Whisper ASR $\rightarrow$ GeminiEmbedding), where predictions were made by finding the class label embedding with the smallest cosine distance to the input embedding. Comparing Sound Input (cascade) against Text Input (oracle transcriptions) reveals varying degrees of headroom; while some languages like German show minimal loss (0.84 to 0.83 accuracy), others like Arabic suffer more significant degradation (0.77 to 0.67), highlighting ASR's varying impact on downstream understanding.Conversely, for non-speech tasks, we utilized direct audio encoders to demonstrate MSEB's breadth. For Environmental Sound Classification on FSD50K, we evaluated the CLAP encoder [26], leveraging its dual-encoder architecture to perform zero-shot classification by matching audio embeddings directly to class label text embeddings (achieving 0.43 mAP). For Bioacoustics on BirdSet, we evaluated Perch [27], a specialized wildlife audio model, using its direct logit outputs (achieving, for example, 0.66 mAP on the NBP test set).

Transcription: We evaluated the robustness of state-of-the-art ASR by transcribing all 26 SVQ locales using Whisper Large v3. Results show significant variability across languages, with Word Error Rates (WER) ranging from as low as 7.7% for Arabic in clean conditions to over 100% for Malayalam, indicating severe challenges for some long-tail languages even with powerful models. Furthermore, environmental noise has a measurable but varied impact; for example, English WER degrades from 1.6% in clean conditions to 3.8% in background speech, while other languages show different sensitivities to specific noise types.

Segmentation: We established a baseline for this task using a cascade approach on the SVQ dataset. Audio was transcribed by Whisper Large v3, and the top-three salient terms were identified using a predefined IDF table. We then utilized Whisper's inherent word-level timestamps to localize these terms. Performance, measured by NDCG, is highly variable and inextricably linked to ASR quality. While English locales achieve high precision (e.g., en-AU at 0.83), performance collapses for challenging languages like Bengali (0.05) and Malayalam (0.002), illustrating the brittleness of relying entirely on explicit text transcription for localizing semantic content.

Clustering: We evaluated discovering latent structure across diverse domains using domain-appropriate encoders. For Bioacoustics (BirdSet), Perch embeddings [27] proved highly effective, achieving V-measures up to 0.53 on complex soundscapes (NBP), significantly outperforming general audio baselines. For Environmental Sounds (FSD50K), the CLAP encoder [26] demonstrated strong performance (V-measure 0.68), successfully organizing diverse audio events without explicit labels. In contrast, for Speaker Clustering on SVQ, simple raw spectrogram features surprisingly matched or even outperformed sophisticated pre-trained models like HuBERT [28] and Wav2Vec2 [29] in many locales (e.g., achieving 0.97 V-measure for speaker ID in te-IN), suggesting that fundamental acoustic properties are often sufficient for basic speaker discrimination in clean conditions.

Reconstruction: To establish a baseline for this task, we evaluated EnCodec [30], a widely adopted neural audio codec. We measured the fidelity of the reconstructed audio against the original signal utilizing Fréchet Audio Distance (FAD) on both SVQ and FSD50K.Results indicate that current models are highly optimized for specific languages and conditions but lack universal robustness. On SVQ, even for clean speech, performance varies drastically by locale, ranging from $\sim$15k FAD for English to over 190k for Hindi. Quality degrades further in noisy environments (e.g., English in traffic noise jumps to $\sim$490k FAD). General environmental sound reconstruction on FSD50K proves even more challenging, yielding the highest FAD score ($\sim$778k) and highlighting the substantial need for improved universal audio generative models.

4 Conclusion

he Massive Sound Embedding Benchmark (MSEB) aims to be the definitive platform for evaluating the sound capabilities of intelligent systems. Its initial release, featuring eight diverse super tasks and four large-scale datasets, covers a significant portion of the sound spectrum and essential real-world functionalities. MSEB's flexible, model-agnostic library enables seamless evaluation of any architecture, from conventional models to LLMs. Our initial experiments reveal substantial performance headrooms between current sound-based methods and text-based oracles, alongside significant variability across different sound types, languages, and noise conditions. This highlights the urgent need for robust, universal sound representations. We invite the research community to contribute to MSEB, fostering collaborative progress toward truly general machine sound intelligence.

References

[1] Charles T Hemphill and Philip R Thrift. Surfing the web by voice. In *Proceedings of the third ACM international conference on Multimedia*, pages 215–222, 1995.

[2] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. "your word is my command": Google search by voice: A case study. In *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, pages 61–90. Springer, 2010.

[3] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[4] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4052–4056. IEEE, 2014.

[5] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5115–5119. IEEE, 2016.

[6] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous. Unsupervised learning of semantic audio representations. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 126–130. IEEE, 2018.

[7] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[9] Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. *arXiv preprint arXiv:2504.10471*, 2025.

[10] Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, et al. Towards learning universal audio representations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4593–4597. IEEE, 2022.

[11] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix De Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*, 2020.

[12] 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, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[13] Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, et al. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint arXiv:2305.11938*, 2023.

[14] Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. Speech-massive: A multilingual speech dataset for slu and beyond. *arXiv preprint arXiv:2408.03900*, 2024.

[15] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*, 2022.

[16] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

[17] Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, et al. Birdset: A large-scale dataset for audio classification in avian bioacoustics. *arXiv preprint arXiv:2403.10380*, 2024.

[18] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

[19] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Proc. INTERSPEECH 2019*, pages 2350–2354, 2019.

[20] Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. Kad: No more fad! an effective and efficient evaluation metric for audio generation. *arXiv preprint arXiv:2502.15602*, 2025.

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[23] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini embedding: Generalizable embeddings from gemini, 2025.

[24] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large language models, 2024.

[25] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[26] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[27] Bhuvanesh Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel-rahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[29] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[30] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The MSEB benchmark is introduced in abstract and introduction and further described in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: This is a benchmark paper, its main goal is providing a framework to progress in speech embedding. We do not see any clear limitation with the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical result is presented.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the code, are open-sourced and experiments conducted by the paper are reproducible

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: SVQ dataset and benchmakr evaluation code are all open sourced/

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Reported numbers are all inference time performance of state-of-the-art embedding models. All details are presented either in the paper, appendices or the codebase

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Due to space limitation, the error bars are mostly presented in the appendices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Available in open source library and also described in appendices

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The positive impact of researching in sound embedding and understanding are discussed. We do not see any negative impacts

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All the data collected are anonymously collected, speaker information is replaced by unique code per speaker

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: SVQ dataset is introduced and documented in the paper, appendices and open source library

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not see any risk

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: core method development in this research does not involve LLMs

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A SVQ Dataset

The SVQ dataset is an open-source collection, accessible at `https://huggingface.co/datasets/google/svq`, which was specifically curated to support a variety of tasks within the MSEB benchmark. This section provides a high-level statistical overview and descriptive details of this dataset.

Splits: The audio data in this release is presented as a single, comprehensive collection, rather than being pre-divided into training, validation, or testing subsets. This decision stems directly from the design of the data acquisition process. Specifically, text prompts and recording environments were randomly allocated across the speaker cohort. While this approach promotes a rich variety of conditions, it introduces a complexity for traditional data splitting: creating partitions that ensure no overlap of speakers and no overlap of text material between splits (a common best practice) would lead to a substantial data reduction, estimated at around 40% of the total recordings.

The primary goal guiding this release strategy is to maximize the utility and volume of the data available to users. Therefore, to avoid this significant data loss and provide the fullest possible dataset, the data is released in its entirety as an undivided evaluation set. Users intending to train models with this data will need to devise and implement their own splitting strategies, keeping in mind the inherent trade-offs between data volume and strict speaker/text disjointness if they attempt to replicate such conditions.

Linguistic Composition: The dataset includes 26 language locales, representing 17 different languages. Languages, sourced in correspondence with the XTREME-UP benchmark [13], were recorded across 1 to 5 locales each to ensure representation of various regional dialects and accents. Recordings from all identified locales are incorporated into the final dataset. Languages with multiple locales include:

- English (en): en-AU, en-GB, en-IN, en-PH, en-US
- Arabic (ar): ar-EG, ar-x-gulf, ar-x-levant, ar-x-maghrebi
- Bengali (bn): bn-BD, bn-IN
- Urdu (ur): ur-IN, ur-PK

Note: uk-PK in the original text is presumed to be ur-PK for consistency with the language code.

Dataset Scale and Audio Characteristics: Mean recording duration: 5.1 seconds
Minimum recording duration: 1 second
Maximum recording duration: 62.4 seconds
Total individual recordings: 171,434
Distinct text transcripts (prompts): 25,549
Total unique speakers: 700

An overview of the SVQ dataset's high-level statistics is presented in Table 1.

Speaker information: The SVQ dataset offers valuable speaker-specific metadata, including anonymous identifiers, self-reported gender, and age, for its 700 unique participants. An examination of the gender distribution reveals a diverse composition: the predominant group consists of 376 speakers identifying as female (approximately 53.7%), followed by 308 speakers identifying as male (approximately 44.0%). Additionally, the dataset includes a smaller segment of 2 individuals identifying as non-binary (around 0.3%) and 14 speakers (2.0%) who opted not to provide an answer regarding their gender. Collectively, these figures depict a generally balanced representation between the two largest gender categories, albeit with a slightly greater prevalence of female participants. For a more detailed, locale-specific breakdown of this gender distribution, refer to Table 2.

The overall age profile of the SVQ dataset, as detailed in Table 3, indicates a participant pool primarily composed of adults. Across all locales, the median age of speakers is 29 years. The full age spectrum represented in the dataset is quite broad, with the youngest participant being 18 years old and the oldest being 71 years old. This range suggests that while the central tendency leans towards young to

Table 1: SVQ dataset: general statistics.

Language	Recordings	Speakers	Locale	Recordings	Speakers
ar	52453	229	ar-EG	13811	59
			ar-x-gulf	13452	58
			ar-x-levant	12600	55
			ar-x-maghrebi	12590	57
bn	8845	42	bn-BD	4119	22
			bn-IN	4726	20
en	28004	121	en-AU	5506	25
			en-GB	5586	25
			en-IN	5632	23
			en-PH	5647	24
			en-US	5633	24
fi	10365	51	fi-FI	10365	51
gu	3689	16	gu-IN	3689	16
hi	3805	18	hi-IN	3805	18
id	5726	28	id-ID	5726	28
ja	2833	13	ja-JP	2833	13
kn	3453	15	kn-IN	3453	15
ko	6520	28	ko-KR	6520	28
ml	3353	16	ml-IN	3353	16
mr	3699	16	mr-IN	3699	16
ru	11912	48	ru-RU	11912	48
sw	5878	25	sw	5878	25
ta	3308	16	ta-IN	3308	16
te	10410	45	te-IN	10410	45
ur	7181	32	ur-IN	3688	16
			ur-PK	3493	16

middle-aged adults, there is also inclusion of both younger adults at the cusp of adulthood and more senior individuals.

Examining the age distributions within specific language locales reveals considerable variability, highlighting the diverse demographic makeup of the dataset. Median ages at the locale level fluctuate, ranging from a younger median of 24 years for speakers in the fi-FI (Finnish in Finland) locale, to notably older medians such as 39 years for en-AU (English in Australia), and 37 years for both ru-RU (Russian in Russia) and ja-JP (Japanese in Japan). This indicates that the typical age of participants can differ substantially from one region or language group to another.

Furthermore, the age spans (from minimum to maximum age) within locales also show diversity. For instance, the ru-RU locale includes speakers up to 71 years old, mirroring the overall maximum age in the dataset, and fi-FI includes speakers up to 67 years. In contrast, some locales exhibit a more constrained age range, such as gu-IN (Gujarati in India), where the maximum participant age is 34, and bn-BD (Bengali in Bangladesh) with a maximum age of 36. This variation underscores that while some locales capture a very wide age demographic, others focus on a more specific, often younger, adult age group.

Table 2: SVQ dataset: speaker gender statistics.

Language	Locale	Female	Male	No Answer	Non-Binary
ar	ar-EG	28	31		
	ar-x-gulf	38	16	4	
	ar-x-levant	36	16	3	
	ar-x-maghrebi	28	29		
en	en-AU	16	9		
	en-GB	15	9		1
	en-IN	12	8	3	
	en-PH	20	4		
	en-US	10	12	1	1
bn	bn-BD	4	18		
	bn-IN	4	16		
fi	fi-FI	12	38	1	
gu	gu-IN	8	8		
hi	hi-IN	9	8	1	
id	id-ID	23	5		
ja	ja-JP	8	4	1	
kn	kn-IN	11	3	1	
ko	ko-KR	19	9		
ml	ml-IN	9	7		
mr	mr-IN	8	7	1	
ru	ru-RU	31	16		1
sw	sw	12	13		
ta	ta-IN	10	6		
te	te-IN	18	27		
ur	ur-IN	6	10		
	ur-PK	9	7		
total		376	308	14	2

Table 3: SVQ dataset: speaker age statistics.

Language	Locale	Median Age	Min Age	Max Age
ar	ar-EG	32	20	56
	ar-x-gulf	30	19	48
	ar-x-levant	30	20	52
	ar-x-maghrebi	27	18	46
en	en-AU	39	22	65
	en-GB	27	20	49
	en-IN	27	21	46
	en-PH	31	20	39
	en-US	28	20	62
bn	bn-BD	28	21	36
	bn-IN	30	24	51
fi	fi-FI	24	20	67
gu	gu-IN	28	23	34
hi	hi-IN	31	19	47
id	id-ID	28	22	46
ja	ja-JP	37	28	52
kn	kn-IN	28	22	49
ko	ko-KR	35	20	55
ml	ml-IN	28	22	41
mr	mr-IN	28	22	45
ru	ru-RU	37	21	71
sw	sw	27	22	40
ta	ta-IN	32	22	48
te	te-IN	26	19	46
ur	ur-IN	31	21	4
	ur-PK	32	22	43
total		29	18	71

B Retrieval

We define retrieval tasks at three distinct levels of granularity: page retrieval, passage retrieval, and span retrieval. Each of these tasks is addressed in two variants: in-language and cross-language retrieval.

For passage retrieval, we utilize the indexes provided by XTREME-UP. A single index containing 271,711 passages, including those from 9 languages and additional hard negatives, supports all in-language retrieval tasks. For cross-language retrieval, we use a separate index of 112,426 passages from the English Wikipedia.

For page retrieval, we constructed indexes in two sizes: 'small' and 'full'. The 'small' in-language and cross-language page indexes are created by mapping passages from the previously described XTREME-UP passage indexes to their corresponding page titles, followed by deduplication. The pages referenced by these titles are drawn from the wikipedia/20190301.[ar|bn|en|fi|id|ko|ru|sw|te] TensorFlow Datasets (TFDS) snapshot[1]. This process yields 8,568 distinct pages for the 'small' in-language index (aggregating these nine languages) and 4,559 for the 'small' cross-language index (derived from English page titles).

In contrast, the 'full' indexes utilize all pages from each available language edition within the wikipedia/20190301.<lang> TFDS snapshot. Unlike the multilingual 'small' in-language page index, the 'full' in-language page indexes are language-specific. Furthermore, the 'full' English (en) index serves as the 'full' cross-language retrieval index. The sizes of these 'full' indexes vary significantly, ranging from several thousand to several million pages per language, thereby spanning three orders of magnitude, see Table 4.

Table 4: MSEB page index sizes (full).

ar	bn	en	fi	id	ko	ru	sw	te
5,824,596	1,272,226	87,566	619,207	947,627	980,493	2,449,364	48,434	91,857

The textual content of pages is substantially greater than that of passages and can thus exceed the maximum context length of the employed embedding models. For instance, the Gecko and Gemini models impose an 8,000-token limit. Fig. 4 illustrates the page text length distribution for a selection of languages. As depicted, while most pages (say, 99%) are shorter than this limit, a small portion is considerably longer. Where page length surpasses this capacity, we segment the text into chunks and utilize the mean of their respective embeddings as the page's final representation.

For span retrieval, indexes are constructed from the corresponding passage-level indexes, using a methodology similar to that for the 'small' page retrieval indexes detailed previously. We recommend limiting spans to a maximum of ten tokens.

Comprehensive per-language results for cascade models, including headroom analysis using ground truth transcripts, are presented in Fig. 6 (Gecko) and Fig. 7 (Gemini Embedding). We observe that the headroom varies across retrieval tasks based on factors such as language, in- vs. cross-lingual conditions, and index size.

We highlight the following key observations:

- Whisper demonstrates significant performance differences across languages (see Fig. 5), with word error rates (WERs) ranging from 10% (en) to 100% (ml).
- Both Gecko and Gemini Embedding exhibit consistent performance across (most) languages.
- Gemini Embedding consistently outperforms Gecko.
- Notably, headroom persists even with nearly 'perfect' text embedders because answers cannot always be reliably inferred from noisy text, as exemplified by page retrieval tasks using the 'small' index.
- Conversely, headroom tends to be limited when text embedder performance is low, which is evident in cross-language page retrieval tasks with the 'full' index.

[1] `https://www.tensorflow.org/datasets/catalog/wikipedia`

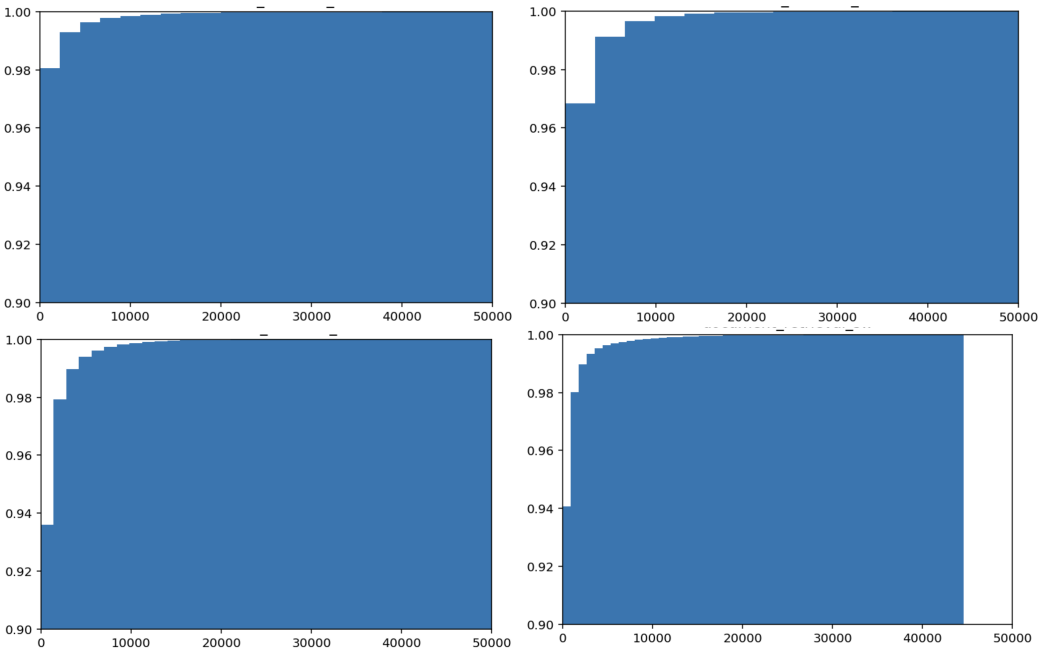

Figure 4: Typical page length distributions (in Gemini Embedding tokens).

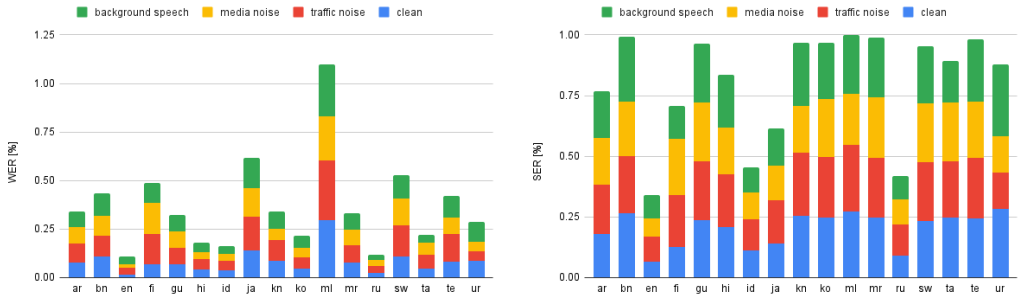

Figure 5: Whisper word error rates (left) and sentence/query error rates (right) across different languages and environments.

Overall, a substantial headroom exists between ASR-based models and ground truth, indicating considerable potential for improved modeling.

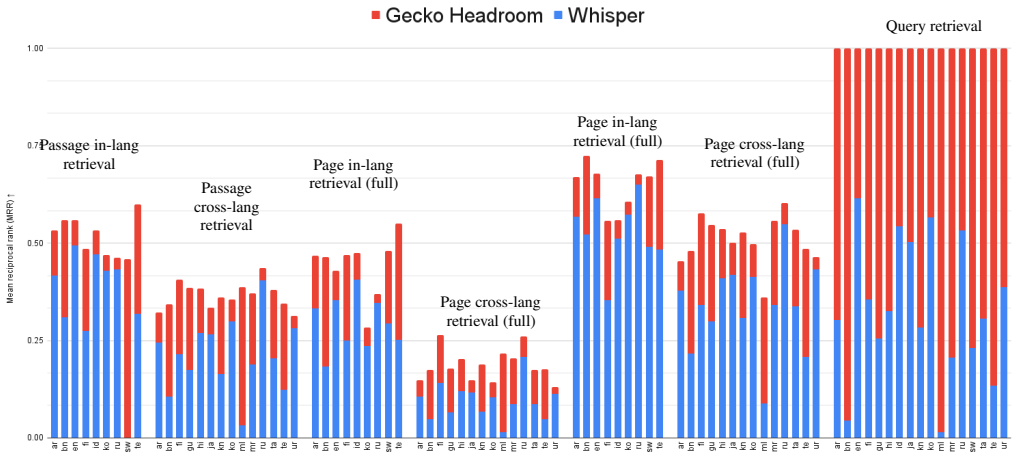

Figure 6: MSEB retrieval tasks: head room analysis for Gecko.

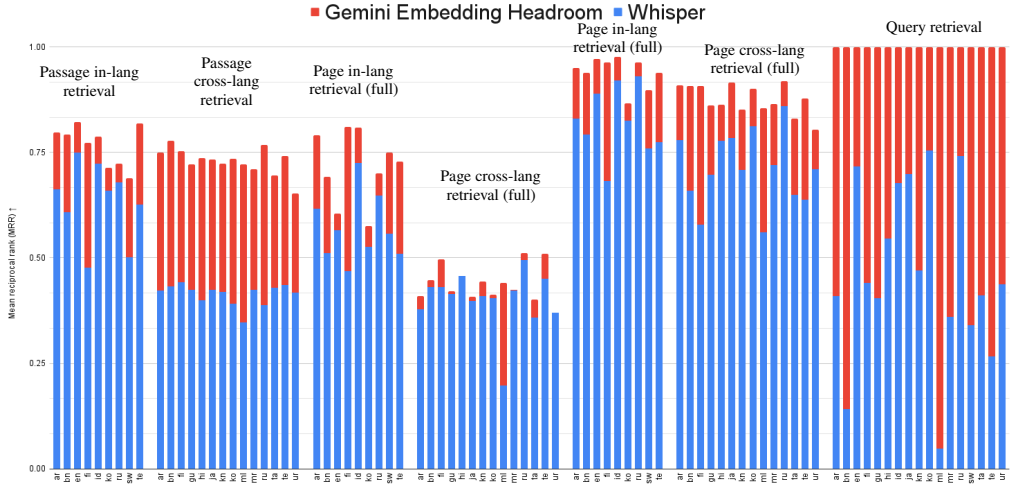

Figure 7: MSEB retrieval tasks: head room analysis for Gemini Embedding.

C Reranking

For the query reranking task, our objective is to generate a challenging set of candidate queries. This set is designed to include queries phonetically similar to the ground-truth query, thereby testing acoustic discrimination, and is augmented with semantically similar candidates to serve as plausible distractors to make guessing the answer more difficult. These candidates are initially generated using a large language model (LLM) guided by a prompt that incorporates these phonetic and semantic constraints, see Table 5.

Subsequently, the generated list undergoes a few post-processing steps: addition of the ground-truth query, duplicate removal, sorting by increasing Word Error Rate (WER) relative to the ground-truth query, and truncation to retain up to the top 250 candidates. Given that a single ground-truth query may correspond to multiple recorded utterances (differing in noise conditions or locales), all such utterances are assigned this consistent set of candidate queries. An illustrative example of these generated candidates is presented in Table 6.

Table 5: System instruction templates for generation of phonetically and semantically similar candidates in query reranking task.

Given the sentence '{text}', generate {num_candidates} sentences in {language} that sound phonetically similar to it. Focus on similar sounding words and rhythm. Do not consider the semantic meaning of the sentences, only the sound. Output the result as a list, with one sentence per line, without reasoning and comments.
Given the sentence '{text}', generate {num_candidates} sentences in {language} that are semantically similar to it. Focus on synonyms and similar semantic meaning. Do not consider the sound of the sentences, only the semantic meaning. Output the result as a list, with one sentence per line, without reasoning and comments.

Table 6: MSEB query reranking candidates for the query "What artists are part of Clamp?"

1. "What artists are part of clamb?"	9. "What artists are affiliated with Clamp?"
2. "What artists are members of Clamp?"	10. "What arses are tart of slump?"
3. "What start are part of arm?"	11. "What actors are start of trump?"
4. "What artists are cut of plane?"	12. "What arced ships part of sum?"
5. "Which designers are part of Clamp?"	13. "What artists impart to clamp?"
6. "What designers are a part of Clamp?"	14. "What farts are part, plum?"
7. "Which animators are part of Clamp?"	15. "What artists part card swarm?"
8. "What individuals are members of Clamp?"	...

Per-language results for our cascade model using Gemini Embedding are presented in Figure 8. The text-based baseline inherently achieves optimal performance (e.g., MRR=1) in this evaluation because the ground-truth query will always yield the highest similarity score when compared against itself.

We observe that headroom varies across languages. Key observations include:

- Whisper demonstrates significant performance differences across languages (see Fig. 5), with word error rates (WERs) ranging from 10% (en) to 100% (ml).
- Gemini Embedding consistently outperforms Gecko.

Overall, a substantial headroom exists between ASR-based models and ground truth, indicating considerable potential for improved modeling.

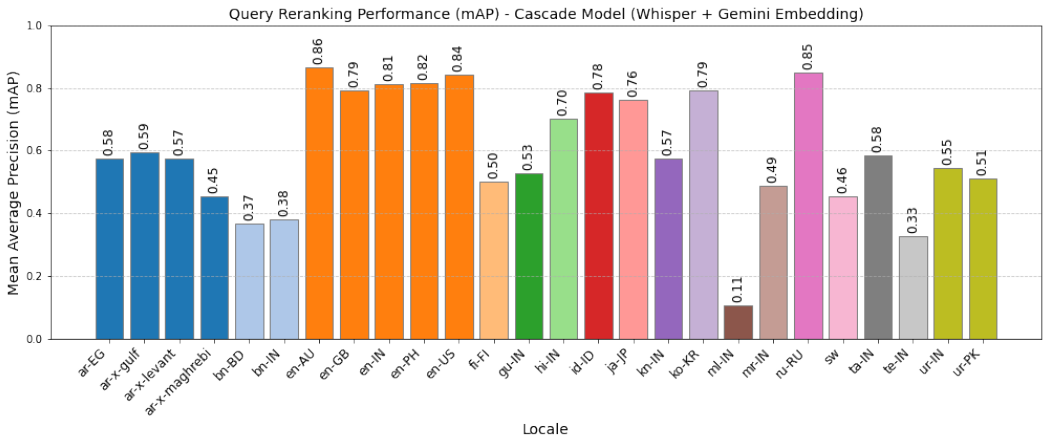

Figure 8: Query Reranking performance (mAP) across SVQ locales using the cascade baseline (Whisper ASR + Gemini Embedding).

D Reasoning

This paper frames reasoning tasks as follows: given a query and a context, the goal is to predict the corresponding answer span from that context or, if no answer is found, to indicate this. We further distinguish between two types of reasoning tasks based on the context provided: passage-level and page-level. In passage-level reasoning, the context consists of a title and a specific Wikipedia passage, for which we reuse the passages from the original XTREME-UP dataset. For page-level reasoning, the context includes a title and the full Wikipedia page referenced by that title. This wiki-page information is sourced from the wikipedia/20190301 collection.

We provide baseline results for a generative (Gemma 3) and a retrieval (Gemini embedding) approach. In the retrieval approach we use a threshold of 0.8 to discriminate between a real answer and 'No Answer'. The per-language results for these two reasoning methods are presented in Fig. 9. The retrieval approach does not work well, supposedly because both Gemini embedding and Gecko have not been fine-tuned for this type of task.

Table 7: System instruction for reasoning tasks.

Task: Answerability Determination and Exact Answer Extraction

Goal: Determine if a question can be answered using the provided context and title, and if so, extract the **exact** answer verbatim from the text.

Input: You will receive a question, title and context each in a new line.

* "question": The question being asked (string).
* "title": The title of the document the context is from (string).
* "context": A text passage which can either be a wiki page or a wiki paragraph that may or may not contain the answer.

Output: You will produce a single JSON object as a plain text string (no markup). The structure depends on answerability:

If the question IS answerable:

* "rationale": (string) A concise explanation of why the provided answer is correct. Be specific, referencing sentences or phrases.
* "answer": (string) The answer to the question, copied **exactly** from the title or context. Do not paraphrase or summarize. Prefer concise answers (shortest possible while complete).

If the question IS NOT answerable:

* "No Answer": (string) A clear and concise explanation of why the question cannot be answered. Specify what information is missing.

Important Considerations:

* **Code change**: The question may be in a language differ from context and title. In that case, answer the question with the same language as context and title.
* **Exact Matches**: Prioritize using the exact words within the provided text. Do not rephrase or summarize. Do not translate to English.
* **Specificity**: Be as specific as possible in your rationale and no_answer explanations.
* **Title and Context**: Consider both.
* **Direct Answers**: Only use the text from title or context. Do not infer or conclude.
* **Ambiguity**: If a question could have multiple different answers based on the context, the answer should be considered "No Answer" as the context is not specific enough. If multiple answers are equally supported and equally correct, select 'No Answer'.
* **Plain Text JSON Output**: The output must be a valid JSON string, but it must be a plain text string – no markup of any kind.

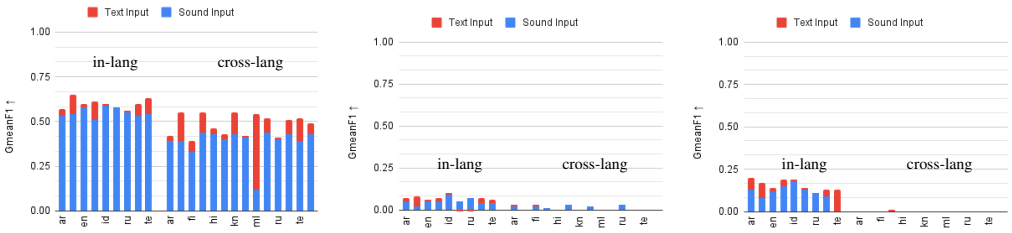

Figure 9: MSEB span reasoning tasks: Generative (Gemma) vs retrieval (Gemini embedding and Gecko) approach.

E Classification

Table 8 outlines state-of-the-art classification performance benchmarks on the Speech-MASSIVE and BirdSet datasets. For intent classification on Speech-MASSIVE, utilizing text-based embeddings from a Whisper model, accuracy varies significantly across the twelve listed language locales. Performance ranges from 61.22% for Arabic (ar) and 63.93% for Hungarian (hu) at the lower end, to a high of 85.93% for French (fr), with other languages such as Spanish (es) and Portuguese (pt) achieving accuracies around 80%. In the domain of bird-song classification on the BirdSet dataset, a wav2vec model with vector embeddings is employed. The class-mean Average Precision (cmAP) metric shows considerable differences across seven distinct recording subsets, with scores ranging from 0.18 for the PER subset to a peak of 0.63 for the NBP subset, and notable scores like 0.45 for HSN. These results provide a detailed look at current capabilities in these respective classification tasks.

Table 8: MSEB classification tasks

name			embedding		metric		
type	dataset	split	model	type	type	value	reference
intent	Speech-MASSIVE	ar de es fr hu ko nl pl pt ru tr vi	Whisper	text	acc [%]	61.22 78.64 80.59 85.93 63.93 74.09 77.49 76.88 80.02 79.51 71.14 68.71	[14]
bird-song	BirdSet	PER NES UHH HSN NBP SSW SNE	wav2vec	vector	cmAP	0.18 0.39 0.27 0.45 0.63 0.28 0.29	[16]

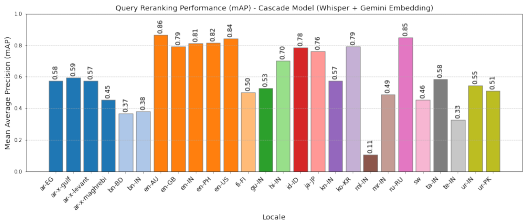

Figure 10: Query Reranking performance (mAP) across SVQ locales using the cascade baseline (Whisper ASR + Gemini Embedding).

F Transcription

For the transcription task, we evaluated the Whisper Large v3 model [22] across all 26 language locales available in the SVQ dataset. The evaluation encompassed all four acoustic environments—clean, background speech, traffic noise, and media noise—to assess the model's robustness. Performance was quantified using two standard metrics: Word Error Rate (WER) and Sentence Error Rate (SER).

Figure 11 illustrates the overall WER and SER for each language, averaged across all environments. The results highlight severe performance disparities between languages. While high-resource languages like English (`en`) and Russian (`ru`) achieve excellent performance with low WERs, other languages such as Malayalam (`ml`) and Bengali (`bn`) exhibit extremely high error rates. This underscores the significant headroom remaining for achieving truly universal ASR capabilities.

To further analyze the impact of acoustic conditions, Figures 12 and 13 present the cumulative error rates broken down by environment. As expected, the `clean` condition consistently yields the lowest errors. However, susceptibility to specific noise types varies; for example, `traffic_noise` causes notable degradation in many locales, while others are more heavily impacted by `background_speech`.

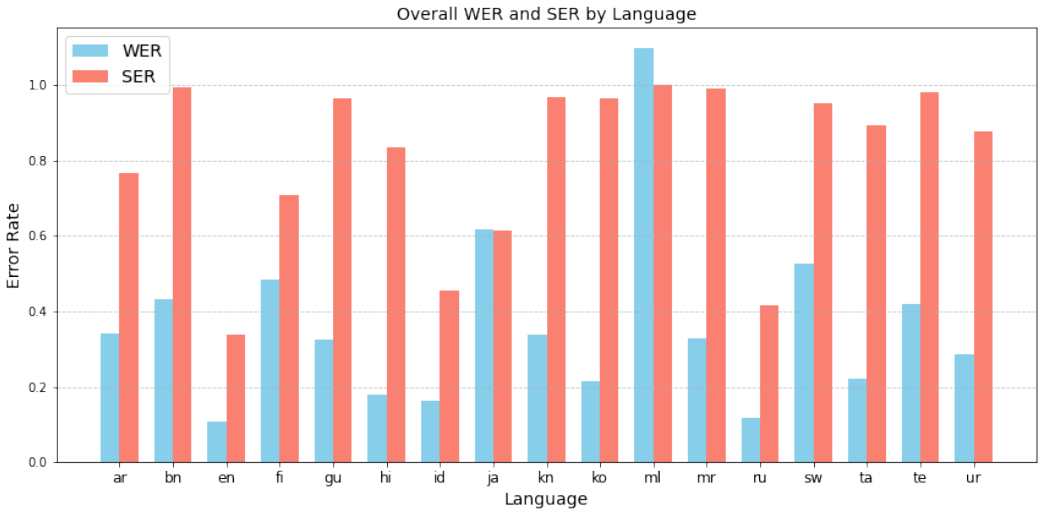

Figure 11: Overall Word Error Rate (WER) and Sentence Error Rate (SER) across all 26 SVQ language locales, averaged over all acoustic environments.

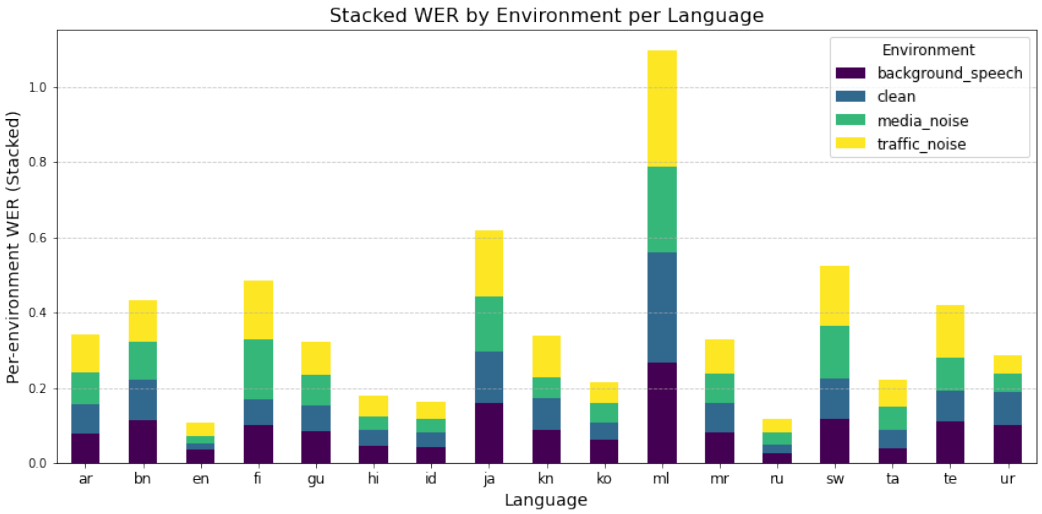

Figure 12: Cumulative Word Error Rate (WER) by environmental condition for each language.

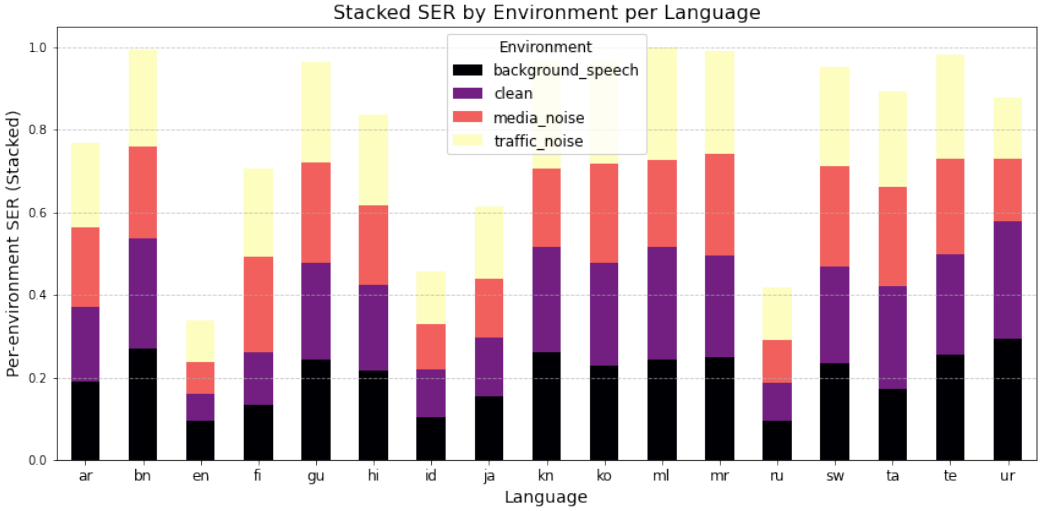

Figure 13: Cumulative Sentence Error Rate (SER) by environmental condition for each language.

G Segmentation

For the segmentation task, we employed a cascade baseline designed to locate key information in time. First, the audio query was transcribed using **Whisper Large v3**. Next, we identified the top-three most salient terms in the resulting transcript using a predefined Inverse Document Frequency (IDF) table computed from Wikipedia. Finally, we utilized the word-level timestamps provided by the Whisper model to determine the start and end times for these salient terms.

The primary metric for this task is **Normalized Discounted Cumulative Gain (NDCG)**, which evaluates the quality of the predicted sequence of salient terms and their temporal order. As shown in Figure 14, performance is highly variable across locales. High-resource languages with strong ASR models, such as English (`en-AU`, `en-US`), achieve high NDCG scores ($\sim$0.80). Conversely, languages with poor ASR performance, such as Malayalam (`ml-IN`) and Bengali (`bn-BD`), show near-zero scores.

To understand the sources of error, we broke down performance into three accuracy metrics (Figure 15):

- **Content Accuracy**: Identifying the correct salient term, regardless of timing.
- **Temporal Accuracy**: Identifying the correct time interval, regardless of the term's content.
- **Overall Accuracy (Strict)**: Correctly identifying BOTH the term and its exact time interval.

Figure 15 reveals that the primary bottleneck is often **Content Accuracy**. If the ASR fails to transcribe the salient term correctly (common in difficult languages), strict accuracy naturally falls to zero. Interestingly, for some languages like Korean (`ko-KR`), even when content is identified (high blue bar), temporal alignment remains challenging, leading to a notable drop in strict overall accuracy (green bar).

This strong dependency on ASR quality is confirmed in Figure 16, which shows a clear negative correlation between Word Error Rate (WER) and NDCG.

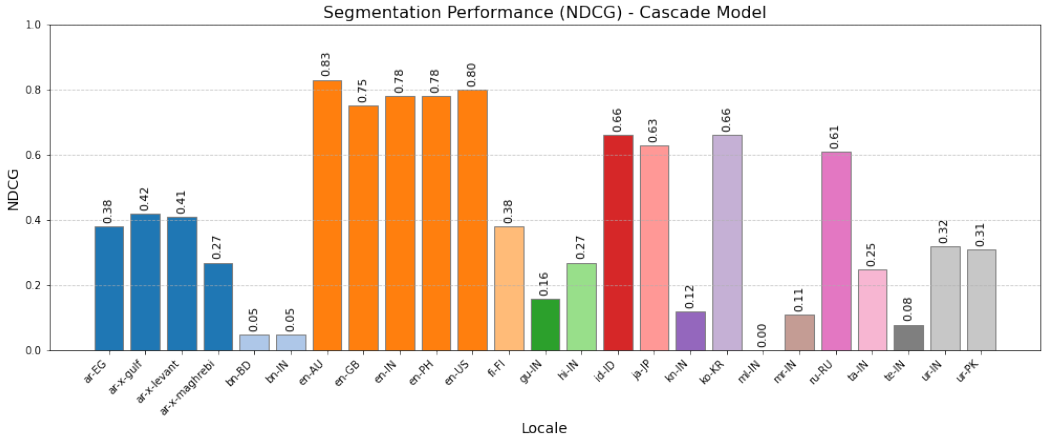

Figure 14: Segmentation performance (NDCG) across all SVQ locales using the cascade baseline.

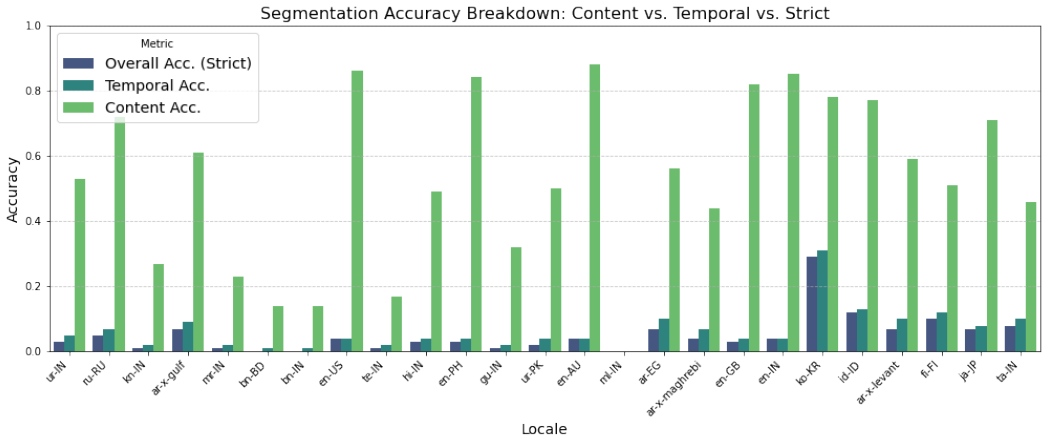

Figure 15: Breakdown of segmentation accuracy. 'Overall Acc. (Strict)' requires both correct content identification and precise temporal localization (within 0.1s tolerance).

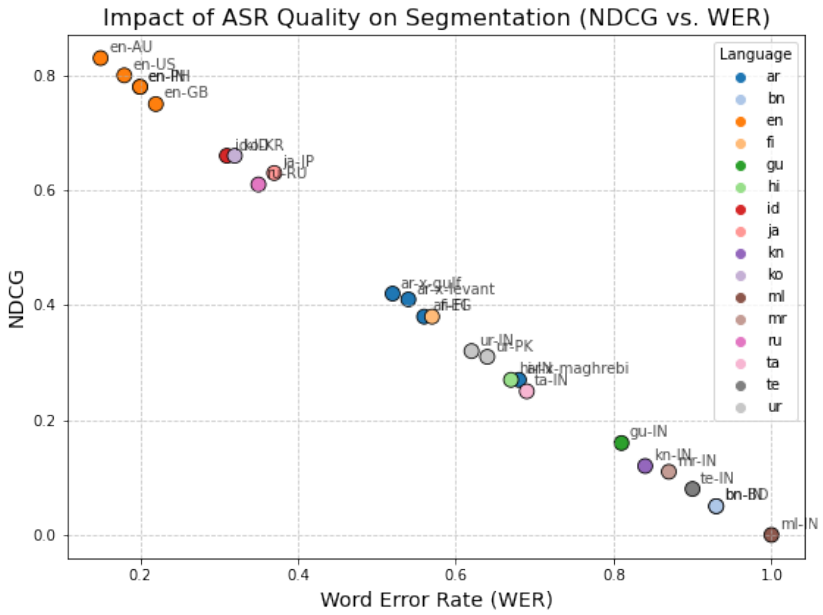

Figure 16: Correlation between ASR quality (WER) and downstream segmentation performance (NDCG), highlighting the cascade model's bottleneck.

H Clustering

For the clustering super task, we evaluated the capacity of various embeddings to discover latent acoustic structure without supervision. Performance was measured using **V-Measure**, a standard metric that balances homogeneity and completeness given the ground truth number of clusters.

We conducted experiments across three distinct domains, employing both general-purpose and domain-appropriate encoders:

- **Bioacoustics**: Evaluated on three BirdSet test splits (HSN, NBP, POW) using **Perch** (specialized for bioacoustics), **CLAP** (general audio), and a raw **Spectrogram** baseline.
- **Environmental Sounds**: Evaluated on FSD50K using **CLAP**.
- **Human Speech**: Evaluated on SVQ across 26 locales for three sub-tasks (Speaker ID, Gender, Age) using self-supervised speech models (**HuBERT Large**, **Wav2Vec2 Large**) and a raw **Spectrogram** baseline.

Figure 17 illustrates performance in non-speech domains. As expected, the specialized **Perch** model dominates in bioacoustics, achieving V-measures above 0.5 on complex soundscapes (NBP). However, **CLAP** demonstrates strong value as a generalist encoder, performing respectably on BirdSet while achieving a high V-measure (0.68) on FSD50K environmental sounds.

Figure 18 presents a notable finding in the speech domain. While large pre-trained SSL models (HuBERT, Wav2Vec2) generally perform well, the simple **Spectrogram** baseline proves highly competitive, particularly for **Speaker ID**. In many locales, it matches or even exceeds the performance of complex models. This suggests that for the clean, close-talking queries typical of Voice Search (as represented in SVQ), fundamental acoustic features are often sufficient for robust unsupervised speaker discrimination.

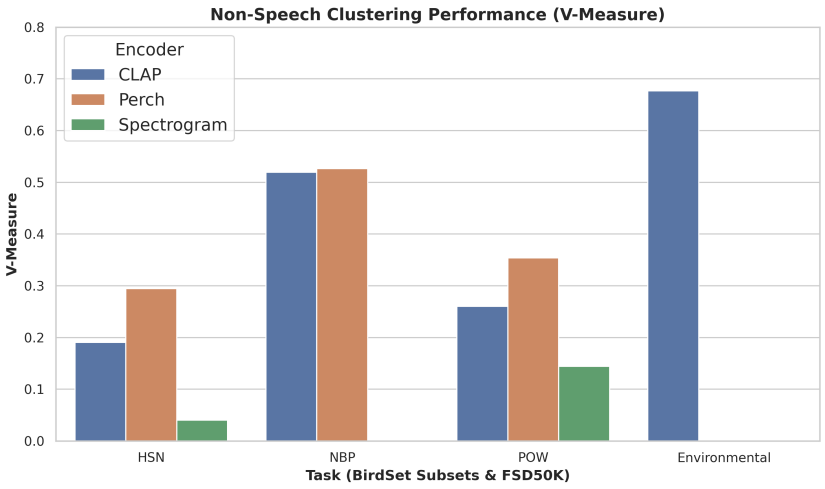

Figure 17: Clustering performance on non-speech domains. Perch excels in its specialized domain of bioacoustics (BirdSet), while CLAP demonstrates strong generalist capabilities across both environmental and bioacoustic tasks.

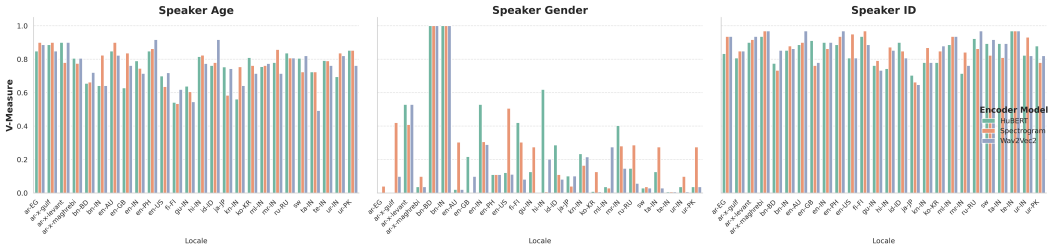

Figure 18: Speaker clustering performance across SVQ locales for Speaker Age, Gender, and ID. Surprisingly, the raw Spectrogram baseline (orange) remains highly competitive with large pre-trained models (HuBERT, Wav2Vec2), especially for Speaker ID.

I Reconstruction

For the reconstruction task, we established a baseline using the EnCodec model [30], a widely adopted neural audio codec. The goal was to assess the model's ability to faithfully reconstruct the input audio signal from its compressed representation. We evaluated reconstruction quality using the Fréchet Audio Distance (FAD) [19] metric, calculated by comparing the spectrograms of the original and reconstructed audio. Spectrograms were computed using a frame length of 1024 samples and a frame step of 320 samples.

Experiments were conducted on two distinct domains: speech under various conditions using the SVQ dataset, and general environmental sounds using the FSD50K dataset. The results, summarized in Table 9, reveal significant challenges for current generative audio models.

As shown in Table 9, EnCodec performs best on clean speech, although even here, substantial variation exists across languages, with Hindi (`hi`) showing significantly higher reconstruction error than English (`en`). Performance degrades considerably in the presence of background noise, with traffic noise proving the most detrimental condition. The reconstruction of general environmental sounds from FSD50K yielded the highest FAD score by a large margin, indicating that generating complex, non-speech acoustic scenes remains a particularly difficult task for current models. These results establish a clear baseline and highlight the need for generative models with improved robustness to noise and greater universality across diverse sound types.

Table 9: Reconstruction performance (FAD, lower is better) using EnCodec.

Dataset	**Condition**	**Mean FAD**	**Min FAD (Example)**	**Max FAD (Example)**
SVQ	Clean Speech	103,662	15,532 (en)	193,799 (hi)
SVQ	Background Speech	163,984	20,094 (en)	310,327 (hi)
SVQ	Media Noise	211,774	125,050 (ko)	346,577 (hi)
SVQ	Traffic Noise	348,885	133,411 (ko)	596,491 (hi)
FSD50K	Environmental Sound	778,232	N/A	