BLAB: BRUTALLY LONG AUDIO BENCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Developing large audio language models (LMs) capable of understanding diverse spoken interactions is essential for accommodating the multimodal nature of human communication and can increase the accessibility of language technologies across different user populations. Recent work on audio LMs has primarily evaluated their performance on short audio segments, typically under 30 seconds, with limited exploration of long-form conversational speech segments that more closely reflect natural user interactions with these models. To address this gap, we introduce Brutally Long Audio Bench (BLAB), a challenging long-form audio reasoning benchmark that evaluates audio LMs on localization, duration estimation, emotion and counting tasks using audio segments averaging 51 minutes in length. BLAB consists of 833+ hours of diverse, full-length audio clips, each paired with human-annotated, text-based natural language questions and answers. Our audio data were collected from permissively licensed sources and underwent a human-assisted filtering process to ensure task compliance. We evaluate six open-source and proprietary audio LMs on BLAB, and find that all of them, including advanced models such as Gemini 2.0 Pro and GPT-4o, struggle with the tasks in BLAB. Our comprehensive analysis reveals key insights into the trade-offs between task difficulty and audio duration. In general, we find that audio LMs struggle with long-form speech, with performance declining as duration increases. They perform poorly on localization, temporal reasoning, speaker counting, and struggle to understand non-phonemic information, relying more on prompts than audio content. BLAB serves as a challenging evaluation framework to develop audio LMs with robust long-form audio understanding and reasoning capabilities.

1 INTRODUCTION

Comprehensive audio perception and reasoning are central to building intelligent agents capable of real-world interaction. Text and visual inputs alone cannot fully capture the richness of human communication, which is multimodal and requires reasoning over a wide range of auditory cues, including tone, pitch, and rhythm. Recent audio language models (LMs; Chu et al., 2024; Gemini Team et al., 2024; OpenAI et al., 2024; Ghosh et al., 2025; Microsoft et al., 2025) have demonstrated strong audio comprehension skills on short audio clips up to 30 seconds (Huang et al., 2024; Sakshi et al., 2024) or at most 5 minutes (Ghosh et al., 2025). This leaves a critical gap: real-world audio, such as meetings, music, podcasts, and other media content, routinely spans tens of minutes to hours. Segmenting long recordings into short chunks may appear practical, but it limits the ability to benchmark reasoning over broader contexts, produces fragmented or inconsistent predictions, and introduces substantial computational and post-processing overhead. While long-context modeling has been extensively studied for text (Dong et al., 2024) and visual inputs (Chen et al., 2025), in the audio domain, long-form analyses are limited to conversational speech recognition (Cornell et al., 2025b), underscoring the need for systematic evaluation of long-form audio reasoning.

To address this gap, we introduce Brutally Long Audio Bench (BLAB), the first benchmark dedicated to reasoning over long-form audio (15 minutes–2 hours). BLAB contains over 833 hours of conversational speech across eight tasks and evaluates four fundamental reasoning skills: temporal localization, speaker counting, emotion interpretation, and duration estimation (see Figure 1 for an overview). Our task selection was guided by practical real world scenarios in which audio LMs could be expected to perform end-to-end reasoning over long-form content. We focused on tasks that require not just recognition of words or sounds, but understanding temporal structure, contextual relationships, and

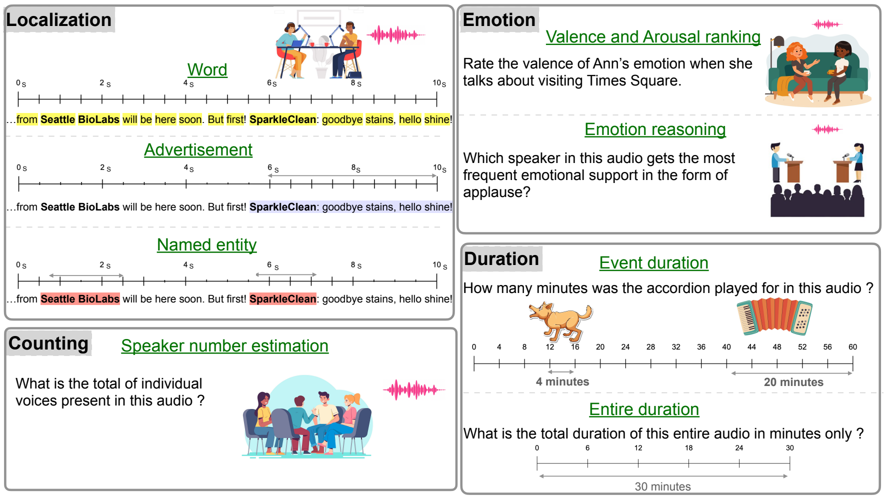


Figure 1: Overview of BLAB, designed to test true long-context multimodal understanding abilities of audio LMs. It contains eight distinct audio tasks across four categories, namely **localization**, **counting**, **emotion**, and **duration estimation**. †

acoustic cues across extended audio spans. These tasks have high impact applications. For instance, media platforms like YouTube create timestamped chapters and highlights to help users find relevant content quickly, and robust temporal localization abilities could make this process even more effective for long-form audio. Compared to existing benchmarks, BLAB is substantially more challenging due to the extensive length of the audio samples, which contain richer contextual information. Our data is entirely sourced from Creative Commons-licensed videos on YouTube, using a rigorous human-assisted filtering procedure to ensure diverse and high-quality content (more details in §2).

Using BLAB, we conduct a comprehensive evaluation (§4) and in-depth analysis (§5) of several frontier audio LMs. Our analysis reveals that even proprietary models achieve an average F_1 score up to **3.02** on localization tasks (Gemini 2.0 Flash) and average exact match accuracy up to **22.25** on the remaining tasks (Gemini 2.0 Pro), underscoring the complexity of our benchmark and the limitations of current modeling approaches. We thoroughly analyze model responses across all tasks and document key patterns, common errors, and areas where models struggle the most. We find that audio duration plays a large role in model performance in BLAB, as well as task complexity. Even though all models struggle to perform tasks in BLAB, we still observe considerable performance gaps between open-sourced and proprietary models, especially Gemini. These findings motivate new research on long-form audio. However, limited transparency in most models' training data and checkpoints makes it hard to probe their results. This underscores the need for open-source long-context multimodal LMs with fully documented data, checkpoints, and training methods.

2 BLAB: Brutally Long Audio Bench

The primary focus of BLAB is to evaluate perception and reasoning abilities of audio LMs on long-form audio derived from various real-world sources. Audio-grounded reasoning requires a model to identify and execute skills relevant for solving an audio task implicitly or explicitly. This motivates us to design eight tasks that demand extended reasoning over long-form audio. One may ask why these tasks cannot be addressed with short audio clips; however, long audio contain contextual dependencies and relationships that are difficult to capture through simple chunking or mere speech recognition. Existing long-form datasets such as CHiME (Barker et al., 2018; Watanabe et al., 2020), AMI (Carletta et al., 2005), and TED (Hernandez et al., 2018) focus on speech recognition. In contrast, BLAB shifts the focus to audio-grounded reasoning. To capture challenging scenarios that require reasoning, we collect audio with diverse properties, including multiple speakers with distinct profiles spread across the recording, overlapping sound events, occasional background music, etc.

Each task includes 200 Creative Commons-licensed audio files sourced from YouTube, alongside corresponding human-annotated questions and answers. Each audio was carefully and thoughtfully selected, prioritizing complexity, quality, diversity, and task relevance. The question-answer pairs were either manually generated by the authors of this paper or model-generated and verified by the authors to ensure quality. We provide a detailed description of each category and its tasks below.

2.1 Localization Tasks

The localization tasks require identifying the start and end timestamps of specific events within audio samples. They enable real-world applications such as audio indexing, retrieval, targeted advertising, and creating timestamped highlights for videos. allowing users to quickly locate moments in lectures or podcasts without relying on transcripts. In this category, we create three tasks, namely **word localization**, **advertisement localization**, and **named entity (NE) localization**. These tasks have received limited attention in long-form audio and are often restricted to very short audio clips (e.g., 30 seconds) (Huang et al., 2024; Fiscus et al., 1970). We describe our annotation procedure below.

Word Localization We used 200 audio files obtained from YouTube (total duration of 191 hours, samples are 57 minutes on average) and applied existing forced alignment (FA) tools to obtain word- and sentence-level alignments between audio samples and their corresponding transcripts. For word localization, we used WhisperX to generate word-level timestamps for each word spoken in each audio sample (Bain et al., 2023). Next, an annotator (one of the authors) manually reviewed a subset of our entire dataset to ensure that the forced alignments were error-free (only $\sim$1% of timestamps needed to be corrected). Each audio sample contains 10,500 word-timestamp pairs on average.

Named Entity Localization We defined nine entity categories to be localized: Event, Location, NORP (nationalities or religious or political groups), Organization, Person, Product, TV shows, Temporal, and Work of Art, as well as "All entities", which includes all of the above. For each category, we also define fine-grained subcategories, allowing us to evaluate whether a model understands the nuances of entity types within the same category. For instance, the example in Table 3 focuses on movie entities, which are a subcategory of Work of Art. Next, we crawled 200 audio files from YouTube and transcribed using WhisperX. Each transcript was fed into a text-only LM to extract plausible NE spans for all entity categories. We tested this part of the annotation process with GPT-4 (OpenAI et al., 2024) and Claude-3.5 Sonnet (Cla) and obtained a higher recall with Claude so we settled on Claude. After extracting the NE spans from text, we mapped these spans back to their timestamps (their location in the audio), also derived from WhisperX. We paired 49 audio files from our pool to the "All entities" category. For the single entity categories, we excluded audio files with fewer than 15 predicted entity spans or a duration of less than 20 minutes, leaving 69 audio files from the remaining 151 in our pool. These 69 audios were then paired to multiple "single entity" categories, resulting in 151 extra items for the NE localization task. This yields audio data with a total duration of 110 hours, each sample being 56 minutes on average. Compared to existing work by Huang et al. (2024), which reports an average of 2 entities per sample, our dataset contains an average of 46 entities per sample. The number of entities to be localized and their temporal position in the audio adds to the complexity of this task, as models often struggle to capture long-term dependencies.

Advertisement Localization We used 200 podcasts from YouTube. Similar to the NE localization annotation procedure, we transcribed each podcast using WhisperX. We fed the transcripts into GPT-4 to extract plausible advertisement segments. Since we obtained very long transcripts, we fed them to the model in chunks of 20 sentences at a time, instructing the model to extract segments that contained an advertisement of a product or promotion from the podcast host. We observed that GPT-4 sometimes incorrectly identified segments of the transcript as advertisements, while in other cases it missed spans that should have been detected. To remedy this, an annotator (one of the authors) reviewed all predicted advertisement segments for every audio sample, removed false positives and added segments that were missing. The percentage of advertisement segments that were kept was 84%. Subsequently, we aligned human-verified advertisement segments with their temporal location in the audio file using the sentence-level timestamps from WhisperX. Our final set of audio data consists of 232 hours of audio across 200 podcasts, with an average of three advertisement segments per podcast and 180 words per advertisement segment.

2.2 Counting Task

Speaker Number Estimation The task of this category is to count the total number of distinct voices in an audio sample. Existing work cover clips with fewer than 10 speakers (Huang et al., 2024; Sakshi et al., 2024; Yang et al., 2024; Cornell et al., 2025a), but BLAB includes samples with 4–80 speakers, averaging 53 minutes in length. Long audio introduce challenges such as overlapping speech and new speakers appearing later, testing a model's ability to track multiple speakers over extended periods. This task is crucial for applications like speaker diarization, meeting summarization. We included 200 YouTube audio samples (total duration of 177 hours with an average audio sample length of 53 minutes) in this task. Two annotators (also authors) listened to the entire audio samples and counted the number of distinct speakers. Due to the task complexity, some samples received different counts from the annotators. For these cases, a third annotator reviewed both annotations and selected a final count. In majority cases, annotators argued that multiple counts could be valid, so we retained a range of count labels and adjusted our evaluation metric (see Section B.2) to consider any number within the range as correct. Overall, audio samples with a range of values as their ground truth count labels account for 60% of our dataset, and the range does not exceed two speakers.

2.3 Emotion Tasks

The emotion tasks involve ranking emotions expressed in speech and non-verbal sounds, and reasoning over emotional expressions. They evaluate a model's ability to integrate semantic content with acoustic features that are strong indicators of emotions, and absent in text alone. They are critical for real-world applications such as analyzing customer service calls, monitoring sentiment in meetings, enhancing voice assistants' empathetic responses, and detecting emotional cues in podcasts or media.

Valence & Arousal Ranking Existing benchmarks typically structure emotion tasks around discrete emotion classification, using labels such as happy, sad, or angry (Huang et al., 2024; Sakshi et al., 2024; Yang et al., 2024). However, this design does not account for variations in emotional intensity, which is particularly relevant in speech-based emotional expressions (Martínez et al., 2014; Sethu et al., 2019). This motivates us to evaluate audio LMs' ability to rank ordinal emotional expressions in audio segments rather than to classify them. Yannakakis et al. (2017; 2021) presented strong evidence supporting the ordinal nature of emotions, demonstrating that ordinal labels enhance the validity, reliability, and overall performance of emotion recognition models in affective computing. Ordinal emotion labels from classical emotion theory (Russell, 1980; Lotfian & Busso, 2019) are based on ranking emotions by intensity, and are often analyzed along three dimensions: valence (the degree of negativity or positivity in the emotion), arousal (level of activation or intensity), and dominance (control or power over the emotion) with respect to another speaker in a conversation.

To curate our dataset, we used 28 audio samples obtained from YouTube (total duration is 18 hours, and samples are 39 minutes on average) and performed annotation through human-AI collaboration. To increase the complexity of the task, we focused on ranking speech segments in an audio file that are difficult to distinguish in terms of emotional content based off as transcript. This challenges cascaded systems and encourages end-to-end audio LMs to jointly consider the prosodic and semantic content in the audio. Each audio sample in our dataset is diarized into speaker-specific segments using Pyannote (Plaquet & Bredin, 2023; Bredin, 2023), obtaining 600 segments on average per audio sample. To create instances for which labels cannot be inferred from text alone, we input each audio segment into an emotion recognition model trained to predict valence and arousal scores (Wagner et al., 2023). Next, we feed corresponding transcripts to GPT-4 to obtain text-based valence and arousal scores. We filter segments where difference between the audio and text-based valence and arousal scores for the same segment was greater than a threshold (0.3 in our experiments). Two annotators reviewed each segment, verifying its alignment with the expressed emotion. For each sample, we randomly sample up to four segments for evaluating both valence and arousal. We crafted questions for each segment, prompting the model to rank the expressed emotion. This resulted in 156 high-quality segments with verified valence and arousal scores across 18 audio samples.

Emotional Reasoning Our motivation for this task is to evaluate an audio LM's ability to understand emotions beyond surface-level sentiments in long audio, focusing on deeper emotional states and pragmatic meanings of speech over extended periods. We manually identified 22 audio samples on YouTube that were suitable candidates for this task, such as those obtained from controversial debate podcasts, comedy shows, movie review podcasts, and emotionally charged interviews. One annotator listened to each audio sample and identified emotional patterns and shifts, such as transitions from calmness to agitation, or from happiness to frustration, and crafted reasoning questions based on verbal and non-phonemic cues, like changes in speech tempo, pauses, or arousal. We design this task as a multiple-choice question answering problem, where confounders are generated by extracting plausible, contextually relevant answers from the audio, increasing the task's difficulty. Overall we have 44 questions paired with 22 audio samples and plan to scale further in future work.

2.4 Duration Tasks

Event Duration This task involves predicting the duration of specific acoustic events (e.g., laughter in a comedy special, question-and-answer segments in a panel session, or a particular speaker's total speaking time in a meeting) within an audio sample, or the total duration of the sample itself in seconds. This task evaluates basic temporal reasoning abilities of audio LMs, including their capacity to identify, localize, and track both verbal and non-verbal acoustic events and activities. We used 87 YouTube audio samples (total duration of 77 hours with an average duration of 53 minutes). An annotator (one of the authors) reviewed each audio sample, detecting and labeling acoustic events and activities. The diverse range of genres in YouTube allowed us to identify multiple events within a single audio sample. After selecting these events, the annotator formulated questions and answers and added their specific durations, pairing 200 questions to all 87 audio files. Each audio sample was paired with at least one question and up to a maximum of four questions.

Entire Duration This task asks an audio LM to predict the total duration of an audio file in seconds. The dataset includes 200 YouTube samples ranging from 8 seconds to 92 minutes (92 hours total, 27 minutes on average), making it the only task with short audio. The total duration of audio samples is 92 hours, and samples are 27 minutes on average. While estimating total duration may seem like a stress test, it is key to temporal reasoning. Accurate event duration estimation relies on understanding the overall temporal structure to correctly associate acoustic events with their time frames.

3 Experimental Setup

3.1 Models

Almost all open-weight audio LMs (Chu et al., 2024; Ghosh et al., 2024b) can only process audio samples with a duration of 30 seconds or less. This limitation comes from their underlying training setup (Radford et al., 2022), which truncates audio samples longer than 30 seconds. Audio Flamingo (Ghosh et al., 2025) is an exception, as it supports audio samples with a duration of up to 5 minutes. In contrast, Gemini models (Gemini Team et al., 2024) support up to 9.5 hours of audio, while GPT-4o (OpenAI et al., 2024) handles up to 8 minutes of audio. Based on these model characteristics and the design of BLAB, our evaluations focus on four Gemini models: Gemini 2.0 Flash, Gemini 2.0 Pro. To enable broader comparisons and analyze the impact of audio duration on model performance, we also evaluate other models (Chu et al., 2024; OpenAI et al., 2024; Ghosh et al., 2025; Microsoft et al., 2025) on a curated short audio ($\leq$ 30 seconds) subset of our benchmark named BLAB-mini.

3.2 Evaluation Strategy

Setup Localization, counting, and entire duration tasks include 200 audio samples each, paired with 20 handcrafted, paraphrased questions. Each question is randomly paired with 10 audio samples to ensure diversity. For event duration and emotion tasks, each question is unique to an audio sample as they contain the event information present in the audio, resulting in 200 unique questions.

In our experiments, the audio LMs take a text prompt (instruction) and an audio file as input and produce text as output. In order to ensure that models truly understand the audio samples and avoid biases by providing options, we restrict our benchmark to free-form generation, except for emotion tasks. Also, prior work suggested that multiple-choice question answering is not always reliable, as distractor options are often either too plausible or models exploit shortcuts to arrive at the correct answer (Balepur et al., 2025). For emotion, confounders are generated by extracting plausible, contextually relevant answers from the audio.

Task	Metric (↑)	Gemini 2.0 Flash	Gemini 2.0 Pro
Word Localization	word F_1	1.12	0.19
Advertisement localization	Frame-level F_1	4.93	0.15
NE Localization	Frame-level F_1	2.97	2.14
Speaker Number Estimation	EMA	8.00	8.50
Valence and Arousal Ranking	EMA	26.28	32.00
Emotion Reasoning	EMA	54.54	64.29
Entire Duration	EMA (without / with ± 2 seconds offset)	0.50/3.50	0.00/2.50
Event Duration	EMA (with / without ± 2 seconds offset)	1.49/4.95	1.49/3.96

Table 1: Performance comparison of Gemini audio LMs across all BLAB long audio tasks. Both models exhibit similar performance, generally achieving low performance across tasks.

Prompt Formatting To ensure consistent outputs across different inputs and models, we append task-specific suffixes to the original questions. For localization tasks, we instruct the model to return JSON-formatted strings with start and end timestamps. For duration and counting tasks, models are instructed to output a number only, without additional explanation. In emotion tasks, which follow a multiple-choice format, the model is prompted to select the most appropriate option from the provided choices. We provide more details about our prompt formatting in Table 5 in the Appendix.

Metrics We use task-specific metrics to evaluate model performance on BLAB. Model outputs are post-processed to match the expected ground truth format. For localization tasks, we compute Frame-level F_1 (*Named Entity & Advertisement*), and Word-level (*Word*) F_1 scores. Counting, duration and emotion tasks are evaluated using exact match accuracy (EMA). For duration tasks, we also report EMA scores with an offset of ± 2 seconds to account for minor timing discrepancies. More details about our evaluation metrics are given in Appendix B.2.

4 Results and Discussion

In Table 1, we present the performance of Gemini 2.0 Flash and Gemini 2.0 Pro on BLAB.

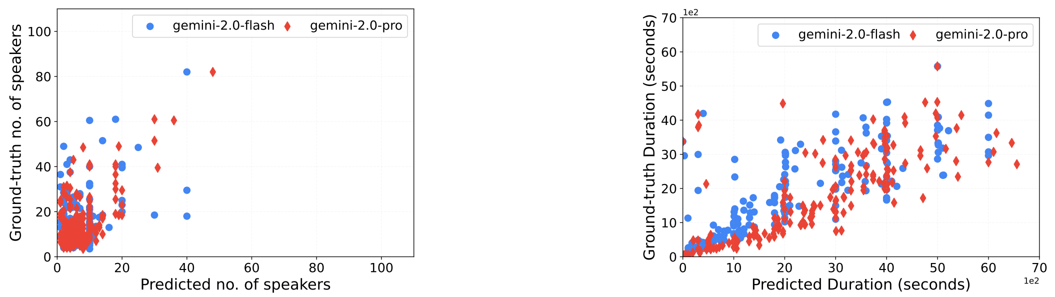
4.1 Localization Tasks

Word localization appears the most challenging task in BLAB with both models performing extremely poorly. Both Gemini models achieved F_1 scores below 2%. These scores are particularly noteworthy, as state-of-the-art word timing models typically achieve scores close to 99% on these last two metrics, as noted by Sainath et al. (2020). We note that each audio sample for this task contains an average of $\sim$10200 words. Gemini, due to its limited output context length of 8096 tokens, is able to generate only $\sim$261 word timestamps per sample, accounting for only about 2% of the ground truth. Gemini 2.0 flash achieves a precision of 24.37%, indicating that the model predictions are correct approximately 24.58% of the time. However, precision for Gemini pro is very low at 3.42%.

Models also perform poorly on NE and advertisement localization, with frame-F_1 scores below 5%. For NE localization, they detect 27% of ground truth entities, but fail to correctly locate them accurately. For advertisement localization, performance is better when ads are at the beginning of the audio files, and the Gemini models are more accurate at predicting start times than end times. This leads us to hypothesize that the models estimate rather than detect segments with advertisements.

4.2 Counting Task

The EMA on speaker number estimation for both models is below 9%. They typically underestimate the number of unique speakers (see Figure 2a) and struggle with overlapping voice. In some cases, we observe overestimation, likely due to the models considering the same speaker at different positions in the audio as distinct. These errors suggest that audio LMs lack the ability to track speakers



(a) Gemini often underestimates the number of speakers on speaker number estimation

(b) Gemini frequently overestimates the duration of audio samples on entire duration.

Figure 2: Predicted versus groundtruth for speaker number estimation and entire duration

consistently across turns, in conversations with overlapping speech, audio with music, commentary, other forms of extraneous content, or audio with varying prosodic features generally.

4.3 Duration Tasks

Entire Duration We find that Gemini struggles to predict the entire duration on full audio samples (EMA up to 3.50%). Compared to our observations in the speaker number estimation task, our analysis indicates that the models often overestimate duration, as shown in Figure 2b. However, in most cases where the predictions are correct, the actual duration is less than 60 seconds.

Event Duration The performance scores are low as well for this task (EMA up to 4.95%). From our observations, there are no clear trends regarding which acoustic events are predicted more accurately than others. Performance generally varies across different event types, and we observe that the model tends to underestimate event durations more frequently than it overestimates them.

4.4 Emotion Tasks

The highest scores in BLAB are on emotion, with Gemini 2 Pro outperforming Gemini 2 Flash. It achieves 32.00% EMA on emotion ranking and 63.63% on emotion reasoning. For emotion ranking, we find that even Gemini 2 Pro struggles to correctly rank scenarios with extremely calm emotions, often misclassifying them as neutral or highly aroused. However, it accurately predicts higher arousal in 80% of cases. Valence rankings show no clear trends, with accuracy near chance.

5 Analysis

What role does the duration of audio play? In this section, we analyze the impact of audio duration on task performance by conducting experiments on shorter audio. This lets us evaluate a broader set of LMs, in particular open-weights models that process audio inputs with a maximum of 30 seconds (Qwen 2, Chu et al., 2024, and Phi-4-Multimodal Instruct, Microsoft et al., 2025) and 5 minutes (Audio Flamingo, Ghosh et al., 2025, and GPT-4o, OpenAI et al., 2024). We conduct analyses on word and entity localization, speaker number estimation and duration tasks. We derive the data from samples in BLAB by extracting audio segments up to 30 seconds. We reuse pre-existing annotations for the localization task and re-annotate the segments for speaker number estimation and duration tasks using the same procedure described in §2. We refer to this as BLAB-mini. It contains 813 questions and 346 minutes of audio in total. More details are provided in Appendix Table 6.

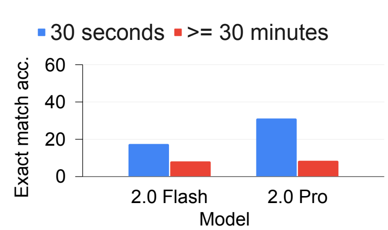


Figure 3: Comparison of long audio and short audio results on Gemini.

Task	Metric ↑	G2 Flash	G2 Pro	Q2	AF2	Phi-4	GPT-4o
Word Localization	Word F_1	**30.22**	8.61	2.43	–	2.73	–
NE localization	Frame-level F_1	45.49	**49.58**	12.07	–	7.63	–
Speaker Number Estimation	EMA	17.50	**31.00**	7.0	6.00	15.50	14.50
Entire Duration	EMA	5.00/31.00	3.50/**34.50**	6.5/27.5	2.5/20.50	3.50/22.00	**7.00**/27.00
Event Duration	EMA	9.45/**36.22**	4.72/29.13	3.15/18.90	1.57/16.54	3.9/24.21	1.57/18.11

Table 2: Performance comparison of audio LMs on BLAB-MINI audio tasks ($\leq$ 30 seconds). **G2** = Gemini 2.0, **Q2** = Qwen 2.0, **AF2** = Audio Flamingo 2. Gemini outperforms all others. Audio Flamingo 2 and GPT-4o refuse to perform any localization task, so we leave them blank. For event duration, we report scores without and with ± 2 seconds offset.

The results are summarized in Table 2 in the main paper and Figure 6 in the appendix. We display comparisons for speaker number estimation in (Figure 3). Gemini models consistently improve across all tasks as the duration is limited to 30 seconds, with the most visible gains observed in word and NE localization. Meanwhile, Table 2 presents a comparison of model performance on BLAB-MINI across multiple models, demonstrating that Gemini outperforms all others on every task.

Are audio LMs long-form zero-shot reasoners? So far, we query the model to directly generate the answer. Inspired by test-time compute research (Kojima et al., 2022; Wei et al., 2022), we explore zero-shot chain-of-thought approaches. We append an auxiliary *reasoning prompt* to the original prompt to guide the model in generating reasoning chains that could lead to better predictions. We test the following prompts that have been effective in text-only LMs: *"Let's think step by step"* and *"Explain your reasoning before making a prediction"*. We conduct this analysis on speaker number estimation and event duration with Gemini 2.0 Flash and use all long audio examples in BLAB.

For event duration, we observed accuracies of 6.93 and 6.44 with the reasoning prompts *Let's think step by step"* and *Explain your reasoning before making a prediction"*, respectively—an average improvement of 3% over the original accuracy of 3.96% without any reasoning prompts (cf. Table 1). For speaker number estimation, the accuracies with both reasoning prompts are 9.55% and 11% respectively, a slight performance boost compared to the original 8% accuracy.

We analyzed individual reasoning steps generated for both prompts, and observed that the model excels at planning but struggles with correctly executing each sub-task in its plan due to limited auditory understanding. For instance, in Table 4 in the Appendix, we can see that both reasoning prompts contain task decomposition, but sub-task results are often incorrect due to weak audio perception. While reasoning traces are coherent, limited auditory grounding leads to errors in the final output. Audio LMs are therefore not truly zero-shot reasoners, and more advanced test-time strategies may be needed to better align reasoning with accurate auditory perception.

Robustness of audio LMs on long-form audio Recent work has suggested that certain audio LMs rely only on text prompts, rather than jointly attending to text and audio input. Sakshi et al. (2024) report that GAMA (Ghosh et al., 2024a), Qwen2-Instruct (Chu et al., 2024) and Gemini Pro (Gemini Team et al., 2024) are more robust to noisy audio and are usually more attentive to audio content compared to other models like SALMONN (Tang et al., 2024) . However, these experiments were done on short audio samples (up to 30 seconds).

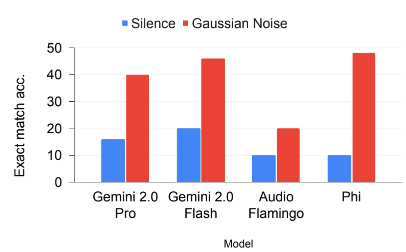
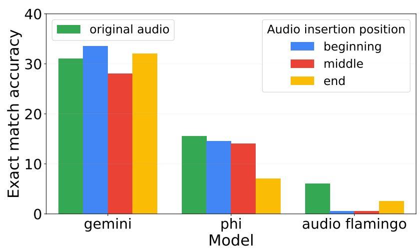


Figure 4: Performance comparison when the original audio input is replaced with silence or Gaussian noise. As the entire input is noisy, the ground truth label is 0

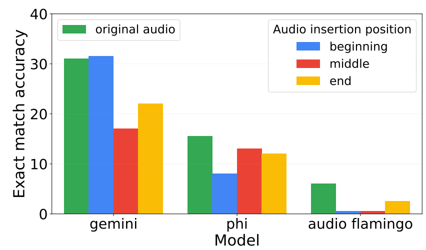
We follow Sakshi et al. (2024) and start by comparing the original model's prediction for speaker number estimation with its prediction when the audio input is replaced with random Gaussian noise. In a different setting, we also replaced the original audio input with silence. All experiments are performed on Gemini, Phi-4-Mini and Audio

Flamingo, since they support longer durations than other models. We generate 5 minute noisy audio samples for Audio Flamingo and Phi-4-Mini, while we generate one hour long noisy audios samples for Gemini. In contrast to previous work on short audio (Sakshi et al., 2024), which finds that audio (LMs) are robust to noisy short audio samples, our findings reveal a different trend for long noisy inputs. Specifically, our analysis shows in Figure 4 that the models are not robust to noisy inputs, and they are particularly less robust to silence than to Gaussian noise.

Next, we investigated how the positioning of noise affects the robustness of an audio LM. Unlike our previous analysis, where the entire input was replaced with Gaussian noise or silence, we now introduce 30-second audio clips from BLAB-MINI into 60-minute noisy recordings for Gemini and 5-minute noisy recordings for Phi and Audio Flamingo, placing the clips at various positions. Our goal is to measure the model's ability to disregard background noise and focus on meaningful content. We conducted this experiment for speaker number estimation, varying the placement of the clean audio clip based on the model's maximum input duration.



(a) Silence



(b) Gaussian

Figure 5: Placing a 30-second clean audio clip at different points within a long, noisy audio input impacts speaker number estimation performance. Proprietary models like Gemini perform better when the clean clip is positioned at the beginning or end of the noisy audio.

In Figure 5, we compare the performance of Gemini, Phi, and Audio Flamingo. Gemini's performance degrades when the 30-second audio clip is placed in the middle, suggesting that the model struggles to effectively use middle information in long input contexts. This is consistent with previous findings on text LMs, where performance peaks if relevant information is at the beginning or end and significantly drops when it is in the middle, even for models designed to handle long contexts (Liu et al., 2024). Degradation is worse with Gaussian noise than silence. For 5-minute inputs, Phi performs better when the 30-second clip is placed in the middle of noise, while Audio Flamingo shows consistent degradation, likely due to difficulty distinguishing noise from actual signal as seen in Figure 4.

6 Conclusion

In this paper, we introduce Brutally Long Audio Bench (BLAB), a challenging benchmark for evaluating long-form audio understanding and reasoning in audio language models across localization, duration estimation, emotion, and counting tasks. BLAB is the first benchmark to assess audio LMs on long-form audio, with durations ranging from over 15 minutes to up to 2 hours, with tasks that are both practical and readily applicable to real-world use cases. Our evaluation of six open-weight and proprietary audio LMs reveals that these models struggle substantially with long-form speech, with performance deteriorating as audio duration increases. Additionally, we find that audio LMs perform poorly on both temporal and counting tasks and struggle to process non-phonemic information in audio. Contrary to existing findings on short-form audio, our analysis suggests that audio LMs are not particularly robust when handling long-form speech. These models tend to rely more on prompts than on actual audio content, making them susceptible to distractions from noisy audio, such as Gaussian noise and silence. We provide a detailed report of our data curation and evaluation framework. Overall our findings underscore the need for more approaches to developing long-context multimodal language models with strong long-form audio understanding capabilities.

REFERENCES

Claude 3.5 sonnet model card addendum. URL https://api.semanticscholar.org/CorpusID:270667923.

Yushi Bai, Xin Lv, Jiajie Zhang, Hong Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. *Annual Meeting of the Association for Computational Linguistics*, 2023.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above, 2025. URL https://arxiv.org/abs/2502.14127.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In *Proc. Interspeech 2018*, pp. 1561–1565, 2018. doi: 10.21437/Interspeech.2018-1768.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, 2005. URL https://api.semanticscholar.org/CorpusID:6118869.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, ICLR '25, 2025.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin shan Lee. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. In *Interspeech 2020*, pp. 4168–4172, 2020. doi: 10.21437/Interspeech.2020-1570.

Samuele Cornell, Christoph Boeddeker, Taejin Park, He Huang, Desh Raj, Matthew Wiesner, Yoshiki Masuyama, Xuankai Chang, Zhong-Qiu Wang, Stefano Squartini, Paola Garcia, and Shinji Watanabe. Recent trends in distant conversational speech recognition: A review of chime-7 and 8 dasr challenges, 2025a. URL https://arxiv.org/abs/2507.18161.

Samuele Cornell, Christoph Boeddeker, Taejin Park, He Huang, Desh Raj, Matthew Wiesner, Yoshiki Masuyama, Xuankai Chang, Zhong-Qiu Wang, Stefano Squartini, et al. Recent trends in distant conversational speech recognition: A review of chime-7 and 8 dasr challenges. *arXiv preprint arXiv:2507.18161*, 2025b.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2086–2099, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.188/.

K. Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

Jonathan Fiscus, Jerome Ajot, John Garofolo, and George Doddington. Results of the 2006 spoken term detection evaluation. ACM SIGIR Conference, Amsterdam, NL, 1970.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika

Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank

V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem

Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, S Ramaneswaran, S. Sakshi, Oriol Nieto, R. Duraiswami, and Dinesh Manocha. Compa: Addressing the gap in compositional reasoning in audio-language models. *International Conference on Learning Representations*, 2023.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.361. URL https://aclanthology.org/2024.emnlp-main.361/.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities, 2024b. URL https://arxiv.org/abs/2406.11768.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities, 2025. URL https://arxiv.org/abs/2503.03983.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, pp. 198–208. Springer International Publishing, 2018.

Chien-yu Huang, Ke-Han Lu, Shi Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2023.

Chien-yu Huang, Wei-Chih Chen, Shu wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, Fabian Ritter-Gutierrez, Ming To Chuang, Kuan-Po Huang, Siddhant Arora, You-Kuan Lin, Eunjung Yeo, Kalvin Chang, Chung-Ming Chien, Kwanghee Choi, Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I-Hsiang Chiu, Heitor R. Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Chang, Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, Kuan-Yu Fang Chiang, Chi An Fu, Hsien-Fu Hsiao, Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian-Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Jui-Chiang Wei, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huck Yang, Chieh-Chi Yang, Jia Qi Yip, Shao-Xiang Yuan, Vahid Noroozi, Zhehuai Chen, Haibin Wu, Karen Livescu, David Harwath, Shinji Watanabe, and Hung yi Lee. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, 2024. URL https://arxiv.org/abs/2411.05361.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. *North American Chapter of the Association for Computational Linguistics*, 2019.

A. S. Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE transactions on multimedia*, 2021.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

S. R. Livingstone and F. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 2018.

R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October-December 2019. doi: 10.1109/TAFFC.2017.2736999.

Héctor P. Martínez, Georgios N. Yannakakis, and John Hallam. Don't classify ratings of affect; rank them! *IEEE Transactions on Affective Computing*, 5(3):314–326, 2014. doi: 10.1109/TAFFC.2014.2352268.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Elliot Nelson, Georgios Kollias, Payel Das, Subhajit Chaudhury, and Soham Dan. Needle in the haystack for memory based large language models. *arXiv preprint arXiv:2407.01437*, 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980. doi: 10.1037/h0077714.

Tara N. Sainath, Ruoming Pang, David Rybach, Basi García, and Trevor Strohman. Emitting word timings with end-to-end models. In *Interspeech*, 2020. URL https://api.semanticscholar.org/CorpusID:226200377.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth S. Narayanan. The ambiguous world of emotion representation. *ArXiv*, abs/1909.00360, 2019. URL https://api.semanticscholar.org/CorpusID:202540864.

Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung yi Lee, and Shinji Watanabe. ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets. In *Interspeech 2024*, pp. 1230–1234, 2024. doi: 10.21437/Interspeech.2024-2248.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8906–8937, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.496. URL https://aclanthology.org/2023.acl-long.496/.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

Joseph P. Turian, Jordie Shier, H. Khan, B. Raj, Björn Schuller, C. Steinmetz, C. Malloy, G. Tzanetakis, Gissel Velarde, K. McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, J. Salamon, P. Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. Hear: Holistic evaluation of audio representations. *Neural Information Processing Systems*, 2022.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-bench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu hsien Huang, W. Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdel rahman Mohamed, and Hung yi Lee. Superb: Speech processing universal performance benchmark. *Interspeech*, 2021.

G.N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, pp. 248–255, San Antonio, TX, USA, October 2017. doi: 10.1109/ACII.2017.8273608.

G.N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, 12(1):16–35, January-March 2021. doi: 10.1109/TAFFC.2018.2879512.

A Appendix

Limitations

Audio data for BLAB is sourced from various real-world recordings, including interviews, podcasts, and political speeches. Our annotation framework leverages human-AI collaboration. However, we observe that overlapping speech is common, which can impact the accuracy of automatic annotation tools like WhisperX and Pyannote (Bain et al., 2023; Bredin, 2023), as they may not be robust in handling such cases. To address this, our framework includes human verification steps to improve annotation quality.

Ethics Statement

BLAB is entirely sourced from Creative Commons-licensed videos on YouTube, using a rigorous human-assisted filtering process to ensure diverse and high-quality content. Human speech is a particularly sensitive modality, as it is deeply personal and can convey not only language but also emotions and mental states. Each audio file in BLAB is carefully selected, with deeply sensitive material excluded to protect privacy. We have also ensured that no child sexual abuse material is present in our dataset.[1] We recognize that speech recordings can be used to track or identify individuals without their knowledge or consent. To address this, we have built our dataset using publicly available YouTube data that complies with ethical guidelines regarding privacy and data usage. However, we also acknowledge the potential risks of data misuse, such as the unintended identification of individuals or the reinforcement of biases in downstream audio language models due to potential contamination. Therefore, we encourage responsible use of our dataset and highlight the importance of considering privacy and ethical concerns when applying it to evaluate real-world applications. To promote transparency and reproducibility, we will make our benchmark publicly available, along with associated evaluation metrics and data curation framework, allowing the research community to contribute and build upon our work.

B Related work

Audio Benchmarking Audio benchmarks can be broadly classified into two main categories based on their scope and purpose: *Task-Specific Benchmarks* focus on evaluating models' performance on particular audio tasks. Examples include text-audio retrieval (Koepke et al., 2021), compositional audio reasoning (Ghosh et al., 2023), automatic speech recognition (Panayotov et al., 2015; Shi et al., 2024), audio captioning (Drossos et al., 2019; Kim et al., 2019), and emotion recognition (Livingstone & Russo, 2018). Several benchmarks combine such tasks into a collection such as SUPERB (Yang et al., 2021), HEAR (Turian et al., 2022), among others. Our work falls under the umbrella of *instruction following benchmarks* which assess model capabilities to understand audio signals and follow instructions in a conversation format. Dynamic-SUPERB was one of the first benchmarks of this kind (Huang et al., 2023), followed by AIR-Bench (Yang et al., 2024) and MMAU (Sakshi et al., 2024). However, almost all of these benchmarks contain samples with a maximum duration of 30 seconds. Most closely related to our work is Ghosh et al. (2025), who release LongAudioBench containing tasks with samples with a maximum duration of 5 minutes. There are existing long-form speech benchmarks such as CHiME Barker et al. (2018); Watanabe et al. (2020), AMI Carletta et al. (2005), and TEDHernandez et al. (2018), which often span more than one hour and primarily focus on speech recognition. Our work focuses on much longer long audio samples which can be up to 2 hours long with a large focus on audio reasoning.

Long Context Benchmarks As the context length for text-based LMs has increased, new benchmarks have emerged to test how well the models handle long-form text. These include tasks like information retrieval (Needle-in-a-Haystack; Nelson et al., 2024) and synthetic long-range reasoning (Long Range Arena; Tay et al., 2020). Newer benchmarks such as LongBench (Bai et al., 2023) evaluate models on tasks over inputs with up to 128K tokens. While these efforts have advanced

[1] https://www.missingkids.org/cybertiplinedata

long-context evaluation in text, similar benchmarks are largely missing for audio. Our work fills this gap by introducing a benchmark for evaluating models on extremely long-form audio understanding.

Task	TD (hrs.)	AD (mins.)	Question Example
Word Localization	191	57	Align the provided audio file with each word in its transcript. For each word spoken, predict the start and end timestamps in seconds and milliseconds.
NE Localization	110	56	Your task is to analyze an audio file and detect all **Movie** named entities present within it. **Movie** entity refers to the name of **feature films or animated movies**. For each detected **Movie** entity, provide the **start** and **end** timestamps (in seconds and milliseconds) that indicate the location and duration of the entity in the audio.
Advertisement Localization	232	70	Given the audio file, your task is to detect the presence of promotions and advertisements within the audio. You are required to predict the start and end timestamps (in seconds and milliseconds) for the entire span for all promotion and advertisement segments you have identified.
Speaker Number Estimation	176	53	What's the number of unique voices detected in this sound file?
Valence & Arousal Ranking	18	39	How would you assess the valence of the speaker in the fourth apology clip within this audio file?
Emotion Reasoning	16	44	Throughout the course of this debate, which speaker demonstrates the most consistent emotional intensity, particularly in their tone and delivery?
Entire Duration	92	27	What is the total duration of this audio file in seconds only?
Event Duration	174	51	A harmonica is being played fifteen seconds towards the end of this audio. How long in seconds was this instrument played for ?

Table 3: Statistics and examples of questions for each task in BLAB. TD and AD represent total and average duration, respectively. Each task consists of 200 question-audio-answers.

Question	Ground Truth	Let's think step by step	Explain your reasoning
What is the total time in seconds that the first panelist spoke for throughout the entire podcast?	1160	`Here's how we can calculate the total speaking time for the first panelist:1. **Identify the first panelist:** The first panelist is Dan Mogulof.2. **Locate the start and end times of Dan's first speaking segment:** Dan starts speaking at [00:02:00] and ends at [00:14:50].3. **Calculate the duration of Dan's first speaking segment:** 14 minutes and 50 seconds - 2 minutes and 0 seconds = 12 minutes and 50 seconds.4. **Locate the start and end times of Dan's second speaking segment:** Dan starts speaking at [00:48:30] and ends at [00:56:30].5. **Calculate the duration of Dan's second speaking segment:** 56 minutes and 30 seconds - 48 minutes and 30 seconds = 8 minutes and 0 seconds.6. **Convert the durations to seconds:** * 12 minutes and 50 seconds = (12 * 60) + 50 = 720 + 50 = 770 seconds * 8 minutes and 0 seconds = (8 * 60) + 0 = 480 seconds7. **Add the durations together:** 770 seconds + 480 seconds = 1250 seconds.Therefore, the first panelist, Dan Mogulof, spoke for a total of 1250 seconds throughout the entire podcast.`	`Here's how we can estimate the speaking time of the first panelist:1. **Identify the Speaker:** The first panelist is Dan Mogulof.2. **Locate the Start Time:** The first panelist begins speaking at 00:01:55.3. **Locate the End Time:** The first panelist stops speaking at 00:18:00.4. **Calculate the Duration:** 00:18:00 - 00:01:55 = 00:16:05.5. **Convert to Seconds:** (16 minutes * 60 seconds/minute) + 5 seconds = 965 seconds.**Prediction:** The first panelist spoke for approximately 965 seconds.`

Table 4: Comparison of reasoning prompts and model responses for estimating the duration of a person speaking in a panel discussion. While the model effectively decomposes the task and generates relevant reasoning, it often lacks the auditory understanding needed to execute sub-tasks accurately.

B.1 Prompt formatting

Task	Prompt Format
Word Localization	Align the provided audio file with each word in its transcript. For each word spoken, predict the start and end timestamps in seconds and milliseconds. Format the response as: {{’word’: ’<word>’, ’start’: <start timestamp of the word>, ’end’: <end timestamp of the word>}}}
NE Localization	Your task is to analyze an audio file and detect all **movie** named entities present within it. **Movie** entity refers to the name of **feature films or animated movies.** For each detected **movie** entity, provide the **start** and **end** timestamps (in seconds and milliseconds) that indicate the location and duration of the entity in the audio. ∗∗Expected Output Format∗∗ Your response should be formatted as a list of dictionaries: {{ [’entity’: ’<movie name>’, ’category’: ’movie’, ’start’: <start timestamp>, ’end’: <end timestamp>, ’entity’: ’<movie name>’, ’category’: ’movie’, ’start’: <start timestamp>, ’end’: <end timestamp>] }}}
Advertisement Localization	Given the audio file, your task is to detect the presence of promotions and advertisements within the audio. You are required to predict the start and end timestamps (in seconds and milliseconds) for the entire span for all promotion and advertisement segments you have identified. Format the response as: {{ "advertisement": <advertisement text>, "start": <start timestamp of the advertisement>, "end": <end timestamp of the advertisement>}}}
Speaker Number Estimation	What’s the number of unique voices detected in this sound file? Provide only the numeric value without any explanation.

Continued on next page

Table 5 – continued from previous page

Task	Prompt Format

Valence & Arousal Ranking	`How would you assess the valence of the speaker in the fourth apology clip within this audio file?`
	
	`Listen to the audio and select one option from the provided choices that best matches the answer. Return only that option.`
	`Options:`
	`(A) Very Pleasant`
	`(B) Pleasant`
	`(C) Neutral`
	`(D) Very Unpleasant`
	`(D) Unpleasant`

Emotion Reasoning	`Throughout the course of this debate, which speaker demonstrates the most consistent emotional intensity, particularly in their tone and delivery?`
	
	`Listen to the audio and select one option from the provided choices that best matches the answer. Return only that option. Options:`
	`(A) The Tory Party leader`
	`(B) The Labour Party leader`
	`(C) Both speakers exhibit similar levels`
	`(D) It is difficult to determine`

Entire Duration	`What is the total duration of this audio file in seconds only?`
	
	`Provide only the numeric value without any explanation.`

Event Duration	`A harmonica is being played fifteen seconds towards the end of this audio. How long in seconds was this instrument played for ? Provide only the numeric value as an integer without any explanation. Do not use the MM:SS format.`

Table 5: Exact prompt formats used for evaluating each task in BLAB.

B.2 METRICS

Given a model output, we post-process it to match the expected ground-truth format. For localization tasks, we use the `json_repair` library.[2] For the remaining tasks with numeric outputs, we use regular expressions to extract relevant numerical values. The regular expression are designed to identify and format various numeric formats, including integers and numbers expressed with units (e.g., ”35 seconds” or 00:23:00).

Localization Tasks For these tasks, the models are expected to generate JSON outputs. Word localization is typically evaluated using metrics that compare start and end timestamp differences for matching words in the audio transcript, detecting delays in word onset or offset Sainath et al. (2020). For ease of evaluation, we report F_1 scores on the number of correctly aligned words.

For NE and advertisement localization, which are span-localization tasks, we use Frame-level F_1 as mentioned in Shon et al. (2023). This metric is derived from question-answering evaluation frameworks (Chuang et al., 2020), and measures the overlap between the predicted and the ground truth answer spans.

Counting and Duration and Emotion tasks These tasks require numeric answers and are evaluated using exact match accuracy (EMA).

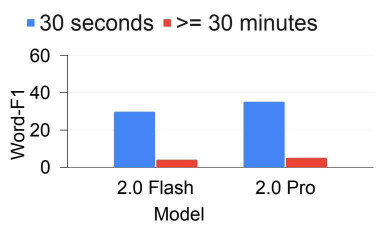
B.3 DATASET STATISTICS OF BLAB-MINI

BLAB-MINI is a subset of BLABthat contains audio samples less than or equal to 30 seconds of audio on average.

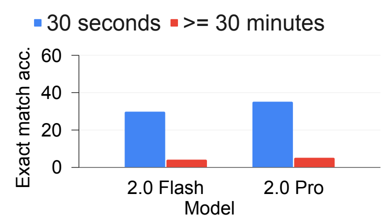
Task	TD (mins.)	AD (seconds.)	Number of questions
Word Localization	89.00	30.00	178
NE localization	36.00	30.00	107
Speaker Number Estimation	99.17	30.00	200
Entire Duration	58.78	17.64	200
Event Duration	63.55	29.79	128
Total	**346.5**		**813**

Table 6: Statistics and examples of questions for each task in BLAB-MINI. TD and AD represent total and average duration, respectively.

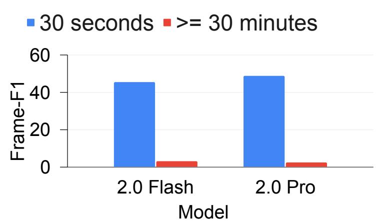
[2]https://pypi.org/project/json-repair/



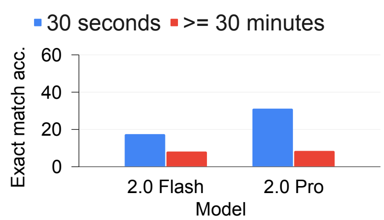
(a) Word Localization



(b) Event Duration



(c) Entity Localization



(d) Speaker Number Estimation

Figure 6: Comparison of long audio and short audio results across Gemini Models