Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs

Microsoft

Abstract

We introduce **Phi-4-Mini** and **Phi-4-Multimodal**, compact yet highly capable language and multimodal models. **Phi-4-Mini** is a 3.8-billion-parameter language model trained on high-quality web and synthetic data, significantly outperforming recent open-source models of similar size and matching the performance of models twice its size on math and coding tasks requiring complex reasoning. This achievement is driven by a carefully curated synthetic data recipe emphasizing high-quality math and coding datasets. Compared to its predecessor, `Phi-3.5-Mini`, **Phi-4-Mini** features an expanded vocabulary size of 200K tokens to better support multilingual applications, as well as group query attention for more efficient long-sequence generation. **Phi-4-Multimodal** is a multimodal model that integrates text, vision, and speech/audio input modalities into a single model. Its novel modality extension approach leverages LoRA adapters and modality-specific routers to allow multiple inference modes combining various modalities without interference. For example, it now ranks first in the OpenASR leaderboard to date, although the LoRA component of the speech/audio modality has just 460 million parameters. **Phi-4-Multimodal** supports scenarios involving (vision + language), (vision + speech), and (speech/audio) inputs, outperforming larger vision-language and speech-language models on a wide range of tasks. Additionally, we experiment to further train `Phi-4-Mini` to enhance its reasoning capabilities.[1]. Despite its compact 3.8-billion-parameter size, this experimental version achieves reasoning performance on par with or surpassing significantly larger models, including `DeepSeek-R1-Distill-Qwen-7B` and `DeepSeek-R1-Distill-Llama-8B`.

1 Introduction

The Phi family of models [AJA+24, AAB+24] have shown that carefully curated and synthesized data enables Small Language Models (SLMs) to achieve highly competitive performance despite having a significantly smaller number of parameters. These models demonstrate comparable results to much larger models. Building on the success of the Phi family of language models, we extend their capabilities to handle additional modalities — such as vision and audio, achieving significant progress akin to private models like GPT [HLG+24], Claude [Ant24], and Gemini [TGL+24].

In this report, we introduce `Phi-4-Multimodal`, a unified multimodal SLM that supports multiple inference modes combining various modalities (e.g., text-only, text + image, speech/audio, speech + image) within a single model checkpoint. `Phi-4-Multimodal` employs a novel "mixture of LoRAs" technique, enabling multimodal capabilities by integrating modality-specific LoRAs while keeping the base language model entirely frozen. Our findings show this technique outperforms existing approaches (e.g., cross-attention designs [ADL+22, AI23]) and achieves comparable performance to fully fine-tuned models on multimodal benchmarks. Additionally, the design of `Phi-4-Multimodal` is highly extensible, allowing seamless integration of new LoRAs to support additional modalities without impacting existing ones.

[1] Please note that reasoning-enhanced `Phi-4-Mini` is a separate model and currently in a preview stage and will not be released concurrently with `Phi-4-Mini` and `Phi-4-Multimodal`.

Our training process comprises multiple stages, including language training (encompassing both pretraining and post-training) and then expansion of the language backbone to vision and speech/audio modalities. For the language model, we train `Phi-4-Mini` using high-quality, reasoning-rich text data. Notably, we include curated, high-quality code datasets to enhance performance on coding tasks. Once the language model training is complete, we freeze the language model and implement our "Mixture of LoRAs" technique to proceed with the multimodal training stage. Specifically, we train two additional LoRA modules alongside modality-specific encoders and projectors to enable vision-related tasks (e.g., vision-language and vision-speech) and speech/audio-related tasks (e.g., speech-language). Both of them contain pretraining and post-training stages for modality alignment and instruction finetuning, respectively.

We also explore the reasoning potential of `Phi-4-Mini` to create a compact yet powerful model that rivals substantially larger state-of-the-art reasoning systems, such as `DeepSeek-R1-Distill-Qwen-7B` and `DeepSeek-R1-Distill-Llama-8B` [GYZ+25].

The key contributions of this model are listed below.

1. **Unified Multi-Modality Support**: In contrast to existing methods [Tea25b, CWW+24] that employ separate models for different modalities, `Phi-4-Multimodal` is designed as a unified model capable of efficiently handling multiple modality scenarios. By leveraging the Mixture of LoRAs [HSW+22], `Phi-4-Multimodal` extends multimodal capabilities while minimizing interference between modalities. This approach enables seamless integration and ensures consistent performance across tasks involving text, images, and speech/audio.

2. **Remarkable Language Performance for the size**: The language models achieve state-of-the-art performance in natural language understanding and generation for its size category. It demonstrates exceptional reasoning and mathematical capabilities, making it well-suited for complex problem-solving and knowledge-based tasks.

3. **Outstanding Code Understanding and Generation for the size**: The language models achieve state-of-the-art performance on code-related tasks within its size category. The model excels at tasks such as code synthesis, debugging, and documentation generation, empowering developers and aiding in software engineering workflows.

4. **Superior Multi-Modal Capabilities for the size**: The model delivers state-of-the-art performance across multi-modal tasks for its size category, demonstrating robust integration of diverse data types. This includes tasks that involve combining images with text and speech modalities, enabling multi-modal reasoning.

5. **Exceptional Speech and Audio Performance**: The model achieves strong performance especially on multilingual speech recognition and translation tasks, and is the first open-sourced model with speech summarization capability.

6. **Enhanced Reasoning Capabilities**: The reasoning-optimized version of `Phi-4-Mini` demonstrates superior reasoning abilities for a model in its size category.

2 Model architecture

The `Phi-4-Mini` series comprises two state-of-the-art small models: a language model (`Phi-4-Mini`) and a multimodal model (`Phi-4-Multimodal`) that integrates language, vision, and speech/audio

modalities. All Phi-4-Mini models use the tokenizer `o200k_base` tiktoken [2] with a vocabulary size of 200,064 intended to support multilingual and multimodal input and output more efficiently. All models are based on decoder-only Transformer [VSP+17] and support 128K context length based on LongRoPE [DZZ+24a].

2.1 Language model architecture

`Phi-4-Mini` and `Phi-4-Multimodal` share the same language model backbone. `Phi-4-Mini` consist of 32 Transformer layers with hidden state size of 3,072 and tied input / output embedding which reduces the memory consumption significantly while providing much wider coverage of vocabularies compared **Phi-3.5**. Each Transformer block includes an attention mechanism based on Group Query Attention (GQA) [ALTdJ+23], which optimizes key and value memory (KV cache) usage for long-context generation. Specifically, the model employs 24 query heads and 8 key/value heads, reducing KV cache consumption to one-third of its standard size. Additionally, in the RoPE configuration [SAL+24], a fractional RoPE dimension is used, ensuring that 25% of the attention head dimension remains position-agnostic. This design supports smoother handling of longer contexts. To determine the peak learning rate, we follow [BBC+24] with $LR^*(D) = BD^{-0.32}$ where B is a constant we tune for this specific model and D is the total number of training tokens. We fit B by tuning across $D = 12.5B, 25B, 37.5B, 50B$.

2.2 Multimodal model architecture

To integrate vision as an input modality, numerous vision-language models have been developed, including the LLava series [LLWL24, LLL+24, LZG+24], QWenVL series [BBY+23, WBT+24], InternVL series [CWT+24, CWW+24, CWC+24], InternLM-XComposer series [ZDW+23, DZZ+24b], Molmo [DCL+24], and NVLM [DLW+24]. Similarly, for audio input, notable contributions include Qwen2-Audio [CXY+24], InternLM-XComposer2.5-Omnilive [ZDC+24], InternOmni, Mini-Omni [XW24], and GLM4-Voice [ZDL+24].

However, in order to enable modality-specific functionality, these multimodal models generally require fine-tuning the base language model, which often diminishes its original language capabilities. Consequently, supporting diverse input signals without compromising quality necessitates deploying multiple models—a particularly challenging limitation for resource-constrained devices. To address this, LLama-Vision [DJP+24] adopts a strategy inspired by Flamingo [ADL+22], adding extra cross-attention layers while preserving the core language model. However, this approach will result in reduced performance on vision-language benchmarks compared to fully fine-tuned models. To fill the performance gap, NVLM [DLW+24] further explores a hybrid framework, employing joint supervised fine-tuning with high-quality text SFT data. Yet, this approach only examines limited language benchmarks and does not address additional training stages often required after SFT.

We adopt the mixture of LoRAs design for our `Phi-4-Multimodal` architecture to support variant multi-modality use cases. Different LoRAs are trained to handle interactions between different modalities. Our `Phi-4-Multimodal` supports a vast range of tasks, including single/multiple images QA/summarization, video QA/summarization, vision-speech tasks, speech QA/summarization/translation/recognition, and audio understanding, while maintains the original language model performance.

2.2.1 Modality Details

Vision modality. The vision modality is implemented with an image encoder, a projector to align the vision and text embeddings and a LoRA adaptor. The vision encoder is based on SigLIP-400M that is finetuned with LLM2CLIP [HWY+24] on large scale image-text pairs with resolution 448×448 . The

[2] `https://github.com/openai/tiktoken`

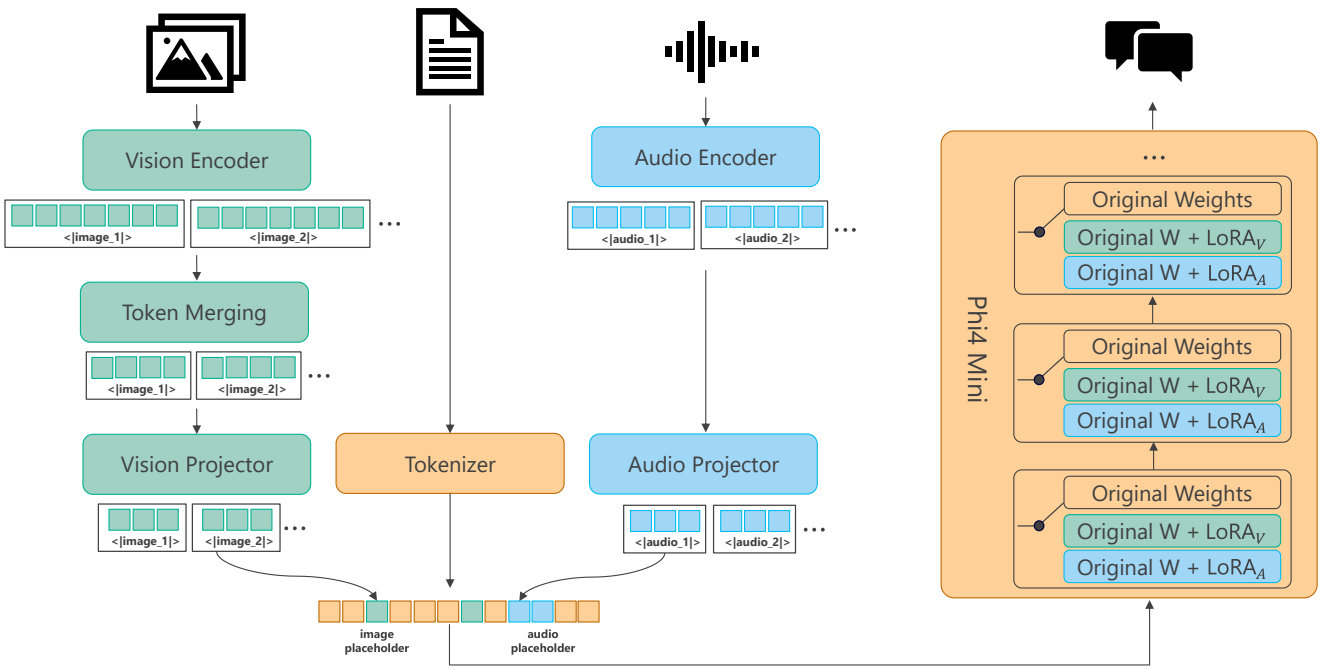

Figure 1: A overview of the Multimodal architecture for `Phi-4-Multimodal`

projector is a 2-layer MLP that maps the vision features dimension to the text embedding dimension. Extra LoRA is added on all the linear layers in the language decoder and only deployed in the supervised fine tuning (SFT) stage. The image encoder and projector introduce 440M model parameters while the vision adapter $LoRA_V$ consumes another 370M model parameters.

In order to enable the model to process images with diverse resolution effectively and efficiently, we proposed a new dynamic multi-crop strategy. Specifically, given a target image, we first compute the crop number for each side by dividing the original size by the crop-size, i.e. $\lceil \frac{H}{C} \rceil \times \lceil \frac{W}{C} \rceil$, where H, W, C are the image height, width and crop size respectively. If the total crop number is within the maximum number, i.e., 16 in the pretraining stage and 36 in SFT, we just slightly resize the image to let it fit the size given by the computed image crops. Otherwise, we will leverage the strategy proposed in InternVL2 [CWW+24] that find the crop number by matching the best aspect ratio. Compared to InternVL2, the key benefits of our strategy is to avoid resizing one small image (e.g., 28×448) to unreasonable large size when looking for the closest image aspect ratio.

Speech and Audio Modality: The speech/audio inputs we used are 80-dim log-Mel filter-bank features with the frame rate of 10ms. To enable `Phi-4-Multimodal` speech and audio functions, we connect a pre-trained audio encoder and `Phi-4-Mini` through an audio adapter. In addition, LoRA is applied on the language decoder to improve the performance of speech/audio benchmarks while preserving the text capability. The introduced modules for the speech/audio modality include:

- An audio encoder, which consists 3 convolutions layers and 24 conformer blocks [GQC+20] with 1024 attention dimensions, 1536 feed-forward dimensions, and 16 attention heads. The convolution layers contribute to a sub-sampling rate of 8, and thus 80ms token rate for the language decoder.
- An audio projector, which is a 2-layer MLP that maps the 1024-dim speech features to the text embedding space of 3072 dimensions, similar to the vision projector.

- $LoRA_A$ that has been applied to all attention and MLP layers in `Phi-4-Mini` with a rank of 320.

The audio encoder and projector introduce 460M parameters while $LoRA_A$ consumes another 460M parameters. Note that the speech token rate is 80ms, indicating 750 tokens for 1-minute audio.

2.2.2 Training Pipeline

The multimodal training stages include vision training, speech/audio training and vision-speech joint training.

Vision Training. The overall training pipeline for multimodal learning consists of vision training, speech and audio training, and joint vision-audio training. Vision training follows a four-stage process: 1) Projector Alignment stage: initially, only the projector is trained using caption data to align vision and text embeddings while preserving the pretrained representation of the vision encoder. 2) Joint Vision Training stage: Next, the projector and vision encoder are jointly trained on the full vision pretraining dataset to enhance key vision capabilities, such as OCR and dense understanding. 3) Generative Vision-Language Training stage: LoRA is then deployed on the language decoder and trained alongside the vision encoder and projector using curated single-frame SFT data, equipping the model with generative capabilities for vision-language inputs. 4) Multi-Frame Training stage: Finally, the model is trained on multi-frame SFT data with the vision encoder frozen, extending the context length coverage to 64k and enabling multi-image and temporal understanding.

Speech and Audio Training. With the `Phi-4-Mini` language model, we conduct a two-stage paradigm for speech and audio training, also known as speech/audio pre-training and post-training. In the pre-training stage, we use large-scale automatic speech recognition (ASR) data to align the audio encoder and `Phi-4-Mini` in the semantic space. In this stage, the encoder and projector is updated with a learning rate of 4e-5 for 50k steps while the language decoder is frozen. We initialize the audio encoder with a pre-trained encoder from the attention-based encoder decoder (AED) ASR model.

After the pre-training stage, the model can only perform the ASR task. To unlock the instruction following capability of `Phi-4-Multimodal` for variety of speech and audio tasks, we continue to train the model with about 100M curated speech and audio SFT samples (after weighted up) as the speech post-training stage. Please refer to Section 3.4.2 for data details. In speech/audio post-training, the audio encoder is frozen. We update the audio projector and $LoRA_A$ with a learning rate of 1e-4 for another 50k steps. We consider different maximum audio lengths for different tasks in post-training. For speech summarization task, we train up to 30-minute audio (22.5k tokens). For other tasks, the maximum audio exposed in training is 30s (375 tokens). If we consider the 128k context length for language decoder, theoretically `Phi-4-Multimodal` can support a maximum 2.8 hours of audio as out of the box inference. It is worth noting that we have not fine tuned the model on such long audio data and it may need further fine tuning to practically support such use cases.

Vision-speech Joint Training. The vision-speech joint training is conducted after vision post-training and speech post-training. We freeze the language base model, audio encoder, and audio projector, while finetuning the vision adapter $LoRA_V$, vision encoder, and the vision projector. In this stage, we train the model mainly on vision-speech SFT data but we also include a mixture of language and vision post-training data to maintain the corresponding performance.

Reasoning Training Recent studies have suggested that training a robust reasoning model only requires a small amount of high-quality data, such as LIMO [YHX+25] and S1K [MYS+25]. However, we propose a fundamentally different training paradigm for SLM: first, we need to conduct a pre-training phase on extensive reasoning data to capture general reasoning chains, and then perform careful fine-tuning on curated SFT or preference data. The continued training of `Phi-4-Mini` for reasoning proceeds in three distinct stages. First, building on `Phi-4-Mini`, the model is pre-trained on approximately 60 billion reasoning CoT tokens generated by frontier reasoning LLMs, after which rejection sampling is employed to filter out incorrect outputs. This allows the reasoning extension of `Phi-4-Mini` to learn the reasoning chains produced by these models. In the second stage, the model is fine-tuned on a smaller but carefully curated dataset of around 200K high-quality CoT samples, chosen to cover diverse domains and varying difficulty levels. Finally, in the third stage, we label filtered incorrect outputs as "dis-preferred" and their corrected counterparts as 'preferred', compiling a new dataset of 300K preference samples for DPO training.

3 Data and training details

3.1 Language training data

3.1.1 Pre-training data

Compared with Phi-3.5-Mini, we improved the quality of the pre-training data from several key aspects:

1. *Better data filtering*: By using an enhanced quality classifier, which is trained on a larger curated dataset consisting of cleaner positive and negative samples, we end up with better filtering quality across multiple languages with various aspects (e.g. toxic, obscure, scientific, etc.), leading to a more comprehensive and controllable filtering strategy overall.

2. *Better math and coding data*: For the math and coding data, we have augmented our original data with a specific instruction-based math and coding data set. This enhancement has resulted in effective results in math, coding and reasoning.

3. *Better synthetic data*: we incorporated Phi-4 synthetic data [AAB+24] into this model training with the same processing and decontamination.

4. *Better data mixture*: With the better classifiers, we re-tuned the data mixture with ablation experiments. Especially we increased the ratio for the reasoning data. That gives us a boost for the model quality.

With these techniques, we built the 5 trillion pre-training data corpus, which is larger and in higher quality compared to the **Phi-3.5-Mini**.

3.1.2 Post-training data

Compared to Phi-3.5-Mini, `Phi-4-Mini` includes a significantly larger and more diverse set of function calling and summarization data. Additionally, we synthesize a substantial amount of instruction-following data to enhance the model's instruction-following capabilities. For coding, we incorporate extensive code completion data, including tasks that require the model to generate missing code in the middle of an existing code snippet. This challenges the model to understand both the requirements and the existing context, leading to significant performance improvements.

3.1.3 Reasoning training data

We generate a large volume of synthetic chain-of-thought (CoT) data from larger reasoning models, covering diverse domains and difficulty levels. During sampling, we employ both rule-based and model-based rejection methods to discard incorrect generations and feed them back for resampling. Also, we label correct sampled answers as preferred generations and incorrect ones as dis-preferred, and create the DPO data. This data has been utilized exclusively for the experimental reasoning model and has not been applied to the officially released checkpoint `Phi-4-Mini`.

3.2 Vision-language training data

The `Phi-4-Multimodal` model's pre-training phase involves a rich and varied dataset, encompassing interleaved image-text documents, image-text pairs, image grounding data, synthetic datasets from OCR of PDFs and realistic images, and synthesized datasets for chart comprehension. During this phase, the model's primary focus is on predicting the next token, concentrating solely on text tokens and disregarding any loss associated with image tokens. The pre-training process involves a total of 0.5T tokens, combining both visual and textual elements. Additionally, the maximum image resolution is capped at 1344x1344, as most training images are smaller than this size. For supervised fine-tuning (SFT), we utilized a combination of a text SFT dataset, publicly available multimodal instruction tuning datasets, and large-scale in-house multimodal instruction tuning datasets that we developed. These datasets span diverse domains and tasks, including general natural image understanding, chart, table, and diagram comprehension and reasoning, PowerPoint analysis, OCR, multi-image comparison, video summarization, and model safety. Collectively, the multimodal SFT data comprises approximately 0.3T tokens.

3.3 Vision-speech training data

For vision-speech data, `Phi-4-Multimodal` model is trained on a diverse set of synthetic vision-speech data, covering single-frame and multi-frame scenarios. Specifically, we reuse a subset of vision-language SFT data and run in-house text-to-speech (TTS) engine to convert the user queries from texts to audios. This subset is carefully selected to avoid certain datasets where the queries are not suitable to read out in speech. We also measure the quality of the synthetic speech by transcribing the audio with in-house ASR model and calculating the word error rate (WER) between original text and transcription. Our final vision-speech data is generated with the WER-based filtering to ensure the quality.

3.4 Speech and Audio Training Data

The training data for speech/audio functions can be categorized into two types: 1) pre-training data with ASR transcriptions to provide a strong alignment between the speech and text modalities; 2) post-training data to unlock the instruction-following capability of `Phi-4-Multimodal` with the speech/audio modality involved. The post-training data covers a variety of tasks, including automatic speech recognition (ASR), automatic speech translation (AST), speech question answering (SQA), spoken query question answering (SQQA), speech summarization (SSUM), and audio understanding (AU).

3.4.1 Pre-training Data

Despite that the audio encoder is initialized from a well-trained ASR model as mentioned in Sec. 2.2, the speech and text latent spaces differ. To pre-train the adapter and reduce the modality gap between

the speech and text sequences, we curate a dataset of approximately 2M hours of anonymized in-house speech-text pairs with strong/weak ASR supervisions, covering the eight supported languages [3].

3.4.2 Post-training Data

Following language post-training paradigm, we curate SFT data for speech/audio post-training, aiming for unlocking the instruction-following capability with speech/audio as query or context. We use both the real and synthetic speech/audio data during speech post-training, covering the majority of speech and audio understanding tasks. All the SFT data are formatted as:

$<|user|><audio>$ {task prompt} $<|end|><|assistant|>$ {label} $<|end|>$
where task prompt is to describe each task in the natural language description and it is null for the SQQA task.

Speech Recognition Data. ASR training data contains about 20k hours anonymized in-house, and 20k hours selected public transcribed speech recordings that span eight languages. The weighted ASR training data contributes to 28M SFT examples.

Speech Translation Data. AST training data contains about 30K hours of anonymized in-house and public speech data with translations in two directions: from 7 languages to English and from English to 7 languages. This data contains both supervised and synthetic translation from a machine translation model. The AST data is created with two formats: direct ST and ASR + translation in a Chain-of-thoughts (CoT) manner, contributing to 28M weighted training examples in post-training.

Speech and Spoken Query Question Answering Data. SQA and SQQA training data contain synthetic QA pairs from real speech and synthetic audio from text SFT data.

- Synthetic QA pairs for SQA: To enable SQA capability, we reuse the speech-transcript pairs in the ASR training data and prompt the language model to generate multiple text QA pairs for each transcript. The low-quality QA pairs are filtered during training.
- Synthetic spoken query (audio) for SQQA: SQA is tasked to respond speech context plus text query. Responding to spoken query directly is also an important capability for `Phi-4-Multimodal`. Consequently, We sample from the language post-training data and convert the text query to audio query using our internal zero-shot TTS system.
- Synthetic LM response for SQQA: Similar to [FWL+24], we synthetically generate responses for speech prompts by prompting the language model with the ASR transcripts of those prompts. The LM response data can improve the SQQA robustness of `Phi-4-Multimodal` in real scenarios because of more diverse spoken queries sampled from the ASR training data.

The total SQA and SQAQA data contribute to 26M weighted SFT examples.

[3] The speech interface supports the following 8 languages: Chinese, English, French, German, Italian, Japanese, Portuguese, and Spanish.

Speech Summarization Data. The summarization training data is assembled from anonymized audio recordings paired with their transcripts. The audio consists of multi-speaker conversational speech that spans a range of topics. Rather than dividing the audio into shorter segments, we maintain its full length up to a maximum of 30 minutes. To construct query-summary pairs for each audio clip, we use GPT-4 to generate a variety of queries and their respective summaries based on the transcripts. For each audio clip, the summarization queries address specific or general aspects of the conversation and vary in format, including length (number of words or sentences) and structure (summaries formatted as bullet points, JSON, or email). The weighted dataset contributes to 1M SFT examples with English speech only.

Audio Understanding Data. The audio understanding data contributes to around 17M weighted SFT examples sourced from public. The dataset is created in the form of (audio, question, answer) tuples, where "audio" contains speech, audio, and music inputs. Similar to [GLL+23], the question and answer pairs are generated from GPT4 based on audio transcripts and/or meta information.

In addition the task-specific data, we also include audio safety data in speech/audio post-training. Please refer to Sec. 5.2 for the details of audio safety data. For all the public data, we utilize our Azure PII Detector[4] to identify and handle Personally Identifiable Information (PII). The training examples with PII detected are removed for privacy concerns.

4 Evaluation

4.1 Multimodal Benchmarks

4.1.1 Vision Benchmarks

We report in Table 1 the evaluation results of `Phi-4-Multimodal` on 13 open-source academic single-image vision-language benchmarks, 2 open-source multi-image/video vision-language benchmarks, and 4 vision-speech benchmarks. Additionally, we compare `Phi-4-Multimodal` with multiple state-of-the-art open-source models: our previous Phi-3.5-Vision [AJA+24], Qwen2.5-VL-3B & 7B [Tea25b], InternVL2.5-4B & 8B [CWC+24], and close-sourced multimodal models Gemini [TAB+23], Claude-3.5 [Ant24][5], and GPT-4o [HLG+24][6]. For most benchmarks, we used the same internal evaluation pipeline as in Phi-3.5-Vision [AJA+24] to ensure fair comparisons across all baseline methods. For benchmarks (e.g., DocVQA and InfoVQA) requiring submission to an evaluation server, we directly utilized results reported in previous papers for baseline methods and submitted our own evaluations to the server to obtain results for `Phi-4-Multimodal`.

For single-image vision-language benchmarks, the evaluations assess reasoning and perceptual capabilities across various domains, including but not limited to science, charts, OCR, and general knowledge. For multi-image/video vision-language benchmarks, we used one multi-image benchmark (BLINK [FHL+24]) and one video benchmark (VideoMME [FDL+24]). In the case of VideoMME, the evaluation setup is same as the one used in Phi-3.5-Vision [AJA+24], where 16 frames are extracted from each video by sampling frames at a rate ensuring uniform time coverage. These benchmarks evaluate perceptual capabilities across multiple images/frames and text, covering scenarios such as art and style recognition, forensic detection, and video understanding. For vision-speech benchmarks, we adopted four existing

[4] `https://learn.microsoft.com/en-us/azure/ai-services/language-service/personally-identifiable-information/overview`

[5] Claude-3.5-Sonnet-2024-10-22

[6] GPT-4o-2024-11-20 and GPT-4o-mini-2024-07-18

	Phi-4-Multimodal 5.6B	Phi-3.5-Vision 4.2B	Qwen2.5-VL-3B 3.8B	InternVL2.5-4B 3.7B	Qwen2.5-VL-7B 8.3B	InternVL2.5-8B 8.1B	Gemini-2.0-Flash-Lite-prv-02-05	Gemini-2.0-Flash	Claude-3.5-Sonnet	GPT-4o-mini	GPT-4o -
MMMU (val) [YNZ+23]	55.1	43.0	47.0	48.3	51.8	50.6	54.1	64.7	55.8	52.1	61.7
MMMUPro (standard/vision) [YZN+24]	38.5 (39.7/37.3)	21.8 (25.5/18.0)	29.9 (31.8/28.0)	32.4 (36.1/28.6)	38.7 (39.5/37.9)	34.4 (39.0/29.8)	45.1 (45.8/44.3)	54.4 (57.1/51.6)	54.3 (56.5/52.1)	40.8 (44.0/37.7)	53.0 (55.3/50.7)
ScienceQA (test) [LMX+22]	97.5	91.3	79.4	96.2	87.7	97.3	85.0	88.3	81.2	84.0	88.2
MathVista (testmini) [LBX+24]	62.4	43.9	60.8	51.2	67.8	56.7	57.6	47.2	56.9	38.8	56.1
Inter-GPS (test) [LGJ+21]	48.6	36.3	48.3	53.7	52.7	54.1	57.9	65.4	47.1	39.9	49.1
MMBench (dev-en) [LDZ+24]	86.7	81.9	84.3	86.8	87.8	88.2	85.0	90.0	86.7	83.8	89.0
POPE (test) [LDZ+23]	85.6	86.1	87.9	89.4	87.5	89.1	87.5	88.0	82.6	83.6	86.5
AI2D (test) [KSK+16]	82.3	78.1	78.4	80.0	82.6	83.0	77.6	82.1	70.6	75.2	83.8
ChartQA (test) [MLT+22]	81.4	81.8	80.0	79.1	85.0	81.0	73.0	79.0	78.4	54.5	75.1
TextVQA (test) [SNS+19]	75.6	72.0	76.8	70.9	77.7	74.8	72.9	74.4	58.6	70.9	73.1
DocVQA (test) [MKJ21]	93.2	69.3	93.9	91.6	95.7	93.0	91.2	92.1	95.2	84.2	90.9
InfoVQA (test) [MBT+22]	72.7	36.6	77.1	72.1	82.6	77.6	73.0	77.8	74.3	59.5	71.9
OCRBench [LLH+24]	84.4	63.8	82.2	71.6	87.7	74.8	75.7	81.0	77.0	77.1	77.7
BLINK (test) [FHL+25]	61.3	57.0	48.1	51.2	55.3	52.5	59.3	64.0	56.9	51.9	62.4
VideoMME-16Frame (test) [FDL+24]	55.0	50.8	56.5	57.3	58.2	58.7	58.8	65.5	60.2	61.2	68.2
Average	**72.0**	**60.9**	**68.7**	**68.8**	**73.3**	**71.1**	**70.2**	**74.3**	**69.1**	**63.8**	**72.4**

Table 1: Comparison results on public vision-language benchmarks. All the reported numbers are produced with the exact same internal pipeline to ensure that the numbers are comparable. These numbers might differ from other published numbers due to slightly different prompts. ∗ Note that for MathVista number of Gemini-2.0-Flash, we find the low performance is because its output sometimes cannot follow the format defined in the input instruction and the evaluation script cannot parse the answer easily.

	Phi-4-Multimodal 5.6B	InternOmni 8.7B	Gemini-2.0-Flash-Lite-prv-02-05	Gemini-2.0-Flash
ShareGPT4o_AI2D[CWT+24]	68.9	53.9	62.0	69.4
ShareGPT4o_ChartQA[CWT+24]	69.0	56.1	35.5	51.3
ShareGPT4o_DocVQA[CWT+24]	87.3	79.9	76.0	80.3
ShareGPT4o_InfoVQA[CWT+24]	63.7	60.3	59.4	63.6
Average	**72.2**	**62.6**	**58.2**	**66.2**

Table 2: Comparison results on public vision-speech benchmarks. All the reported numbers are produced with the exact same internal pipeline to ensure that the numbers are comparable.

benchmarks from InternOmni [CWT+24], which convert text prompts into speech format for evaluation on four vision-language benchmarks. Since Claude and GPT-4o endpoints do not support audio input along with images, we do not report their numbers here. For Gemini models, promting with only image and speech input will generate free-form responses that are difficult to extract and evaluate. Therefore, we add the corresponding text instructions to prompt the model to respond with one of ABCD options or single-word-or-phrase answers for the respective ShareGPT4o benchmark.

From the results in Table 1 and Table 2, we summarize Phi-4-Multimodal's performance as follows:

- On vision-language benchmarks, Phi-4-Multimodal demonstrates significant improvements over our previous model Phi-3.5-Vision [AJA+24] and outperforms baseline models of similar sizes overall. Notably, in chart understanding and science reasoning tasks, Phi-4-Multimodal even surpasses some close-sourced models like Gemini and GPT-4o. We have shown one demo case in Figure 2.
- On vision-speech benchmarks, Phi-4-Multimodal significantly outperforms InternOmni and

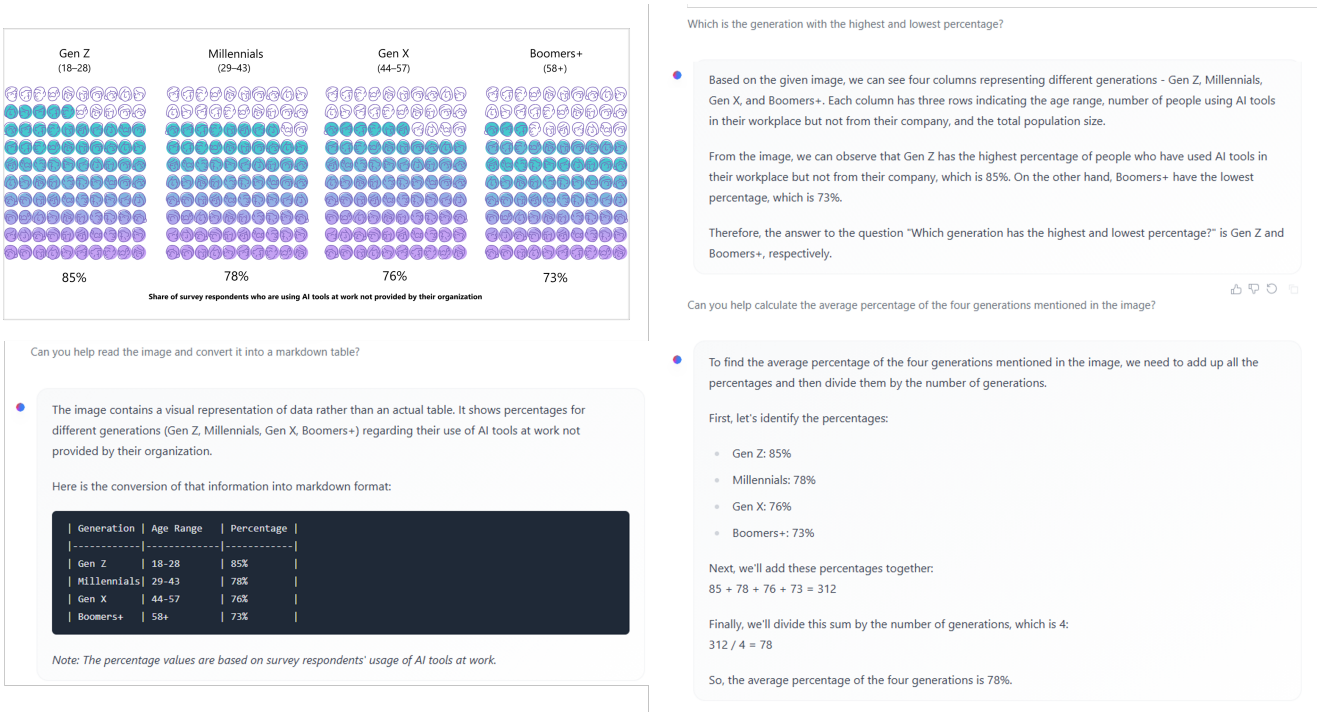

Figure 2: One demo case to show the vision-language understanding and reasoning capability of `Phi-4-Multimodal`.

Gemini-2.0-Flash by a large margin, which have larger model sizes. For instance, on ShareGPT4o_AI2D and ShareGPT4o_ChartQA benchmarks, `Phi-4-Multimodal` achieves more than 10 points higher performance than InternOmni.

- Unlike other open-source vision-language models that fully fine-tune their base language models (often resulting in performance degradation on pure language benchmarks), `Phi-4-Multimodal` keeps the language model entirely frozen by only incorporating additional fine-tunable LoRA modules. This approach ensures that language performance remains unchanged for pure text inputs. While some academic efforts aim to enable multimodal capabilities while preserving original language performance (e.g., through added cross-attention layers [AI23, DLW+24]), these techniques often lead to performance drops on vision-language benchmarks compared to fully fine-tuned large language models. Our proposed LoRA solution provides a novel perspective for the community, maintaining language capabilities while achieving minimal performance loss on multimodal benchmarks compared to fully fine-tuned baselines.

4.1.2 Speech and Audio Benchmarks

We evaluate the speech and audio capabilities of `Phi-4-Multimodal` on a variety of understanding tasks. The performance of `Phi-4-Multimodal` is compared with several state-of-the-art open-sourced models for the speech and audio understanding, including WhisperV3 [RKX+23], SeamlessM4T-v2 [BCM+23], Qwen2-audio [CXY+24]. We also include the performance of close-sourced multi-modal models (GPT-4o [HLG+24] and Gemini [TAB+23]) for comparisons [7]. The results are obtained through

[7] Speech evaluations for closed models are done through Azure cloud API.

Table 3: Main Results on the speech benchmarks. All results are obtained with 0-shot evaluations except additional CoT evaluations on the AST task, where CoT refers to chain-of-thoughts decoding with transcription plus translation in generation. MT-Bench results are averaged scores over two-turn SQA conversations. SSUM evaluation is with the overall numbers covering the adherence and hallucination scores. The scores in the table are judged by GPT-4-0613. N/A indicates the model does not have such a capability.

Task	Metric	Dataset	Phi-4-Multimodal 5.6B	WhisperV3 1.5B	SeamlessM4T-V2 2.3B	Qwen2-audio 8B	Gemini-2.0-Flash	GPT-4o -
ASR	WER ↓	CV15	**6.80**	8.13	8.46	8.55	9.29	18.14
		FLEURS	**4.00**	4.58	7.34	8.28	4.73	5.42
		OpenASR	**6.14**	7.44	20.70	7.43	8.56	15.76
AST	BLEU ↑	Inference Type	(0-shot, CoT)	0-shot	0-shot	0-shot	0-shot	0-shot
		CoVoST2 X-EN	(39.33, **40.76**)	33.26	37.54	34.80	36.62	37.09
		CoVoST2 EN-X	(37.82, **38.73**)	N/A	32.84	34.04	35.93	37.19
		FLEURS X-EN	(29.86, 32.35)	25.76	28.87	23.72	30.69	**32.61**
		FLEURS EN-X	(32.15, 33.56)	N/A	30.44	23.24	**37.33**	36.78
SQQA	Score 1-10 ↑	MT-Bench	7.05	N/A	N/A	4.92	8.07	**8.11**
	ACC ↑	MMMLU	38.50	N/A	N/A	15.53	72.31	**72.56**
SSUM	Score 1-7 ↑	Golden3	6.28	N/A	N/A	2.25	6.29	**6.76**
		AMI	6.29	N/A	N/A	1.34	5.97	**6.53**
AU	Score 1-10 ↑	AirBench-chat	**6.98**	N/A	N/A	6.93	6.68	6.54
	ACC ↑	MMAU	55.56	N/A	N/A	52.50	**61.23**	53.29

evaluation on the exact same test data version without further clarifications. We sample the top-1 token at each generation step during inference.

The main results on the speech benchmark are presented in Table 3. We summarize the performance of `Phi-4-Multimodal` as listed:

- `Phi-4-Multimodal` achieves very strong ASR and AST performance, surpassing the expert ASR model, WhisperV3, and expert AST model, SeamlessM4T-large-v2, on CommonVoice [ABD+20], FLEURS [CMK+23], OpenASR [SMK+23], and CoVoST2 [WWGP21] test sets.
- `Phi-4-Multimodal` is 5.5% relatively better in WER than the best model on the Huggingface OpenASR leaderboard[8] and now ranks No.1 on the leaderboard as of 1/14/2025.
- `Phi-4-Multimodal` is the first open-sourced model with speech summarization capability. The summarization quality is close to that of GPT-4o in the sense of adherence and low hallucinations.
- `Phi-4-Multimodal` is the smallest open-sourced multi-modal LLM that behaves better than the open-sourced Qwen2-audio [CXY+24] with ~2x in size.

We should notice in Table 3 that `Phi-4-Multimodal` is optimized for speech and audio understanding tasks while Gemini and GPT-4o might be optimized towards chat experience. That may be the reason why `Phi-4-Multimodal` outperforms Gemini-2.0-Flash and GPT-4o on ASR and AST tasks while lags behind on the SQQA tasks. We describe the benchmark and evaluation details for each task below.

Automatic Speech Recognition. We evaluate the ASR performance on three public benchmarks: CommonVoice [ABD+20], FLEURS [CMK+23], and OpenASR [SMK+23].

- CommonVoice is an open-source, multilingual speech dataset developed by Mozilla. The test set of CommonVoice version 15.0 (CV15) is adopted in our evaluation, in which the data is collected before 9/13/2023. We conduct the evaluations on the eight supported languages.

[8] `https://huggingface.co/spaces/hf-audio/open_asr_leaderboard`

Table 4: Detailed results on ASR benchmarks. We compute CER (↓) for JA and ZH, and WER (↓) for other languages. nvidia/canary-1B model is the best performing model on the Huggingface OpenASR leaderboard to date. The results of canary and WhisperV3 are from the official report while others are obtained through internal evaluation on the same test data version.

Dataset	Sub-Category	Phi-4-Multimodal 5.6B	nvidia/canary 1B	WhisperV3 1.5B	SeamlessM4T-V2 2.3B	Qwen2-audio 8B	Gemini-2.0-Flash	GPT-4o -
CV15	EN	7.61	N/A	9.30	7.65	8.68	11.21	21.48
	DE	5.13	N/A	5.70	6.43	7.61	6.2	10.91
	ES	4.47	N/A	4.70	5.42	5.71	4.81	11.24
	FR	8.08	N/A	10.80	9.75	9.57	10.45	17.63
	IT	3.78	N/A	5.50	5.50	6.78	4.88	13.84
	JA	10.98	N/A	10.30	12.37	13.55	13.46	19.36
	PT	6.97	N/A	5.90	9.19	10.03	7.4	23.07
	ZH	7.35	N/A	12.80	11.36	6.47	15.87	27.55
	Average	**6.80**	N/A	**8.13**	**8.46**	**8.55**	**9.29**	**18.14**
FLEURS	EN	3.38	N/A	4.10	6.54	5.27	3.96	6.52
	DE	3.96	N/A	4.90	6.95	8.77	4.06	4.17
	ES	3.02	N/A	2.80	5.39	6.90	2.61	3.69
	FR	4.35	N/A	5.30	7.40	9.00	5.06	6.42
	IT	1.98	N/A	3.00	4.70	5.78	1.86	3.28
	JA	4.50	N/A	4.80	11.47	12.68	4.94	5.18
	PT	3.98	N/A	4.00	7.67	10.59	3.57	6.33
	ZH	6.83	N/A	7.70	8.6	7.21	11.74	7.77
	Average	**4.00**	N/A	**4.58**	**7.34**	**8.28**	**4.73**	**5.42**
OpenASR	AMI	11.69	13.90	15.95	56.1	15.24	21.58	57.76
	Earnings22	10.16	12.19	11.29	37.18	14.09	13.13	20.94
	Gigaspeech	9.78	10.12	10.02	26.22	10.26	10.71	13.64
	Spgispeech	3.13	2.06	2.01	12.04	3.00	3.82	5.66
	Tedlium	2.90	3.56	3.91	19.26	4.05	3.01	5.79
	LS-clean	1.68	1.48	2.94	2.60	1.74	2.49	3.48
	LS-other	3.83	2.93	3.86	4.86	4.03	5.84	7.97
	Voxpopuli	5.91	5.79	9.54	7.37	7.05	7.89	10.83
	Average	**6.14**	**6.50**	**7.44**	**20.70**	**7.43**	**8.56**	**15.76**

- FLEURS a multilingual speech dataset designed for evaluating speech recognition and speech-to-text translation models across a wide range of languages. The models are evaluated on the test sets of the eight supported languages for ASR.
- OpenASR Leaderboard on Hugging Face is designed for benchmarking and evaluating the robustness of ASR models on English. The datasets in the leaderboard cover diverse speech domains including reading speech, conversations, meetings, and so on.

The ASR prompt for `Phi-4-Multimodal` is **"Transcribe the audio clip into text."**, which is language agnostic. We notice that the model can learn to recognize in the target language perfectly without providing language information, while Qwen2-audio and Gemini-2.0-Flash require the language information in the prompt to obtain the optimal ASR performance. For example, the ASR prompt for Gemini-2.0-Flash is **"Transcribe the audio clip into {tgt-lang}. Please ignore background noise."** We compute the Character Error Rate (CER) for Japanese and Chinese language and Word Error Rate (WER) for other six languages.

The detailed ASR results on the three benchmarks are summarized in Table 4. Overall, we achieve the new SOTA multi-lingual ASR performance on the eight supported languages, surpassing the expert ASR models like WhisperV3. Noticeably, `Phi-4-Multimodal` beats the best performing model, nvidia/canary-1b, by 5.5% relative WER on the Huggingface OpenASR leaderboard and now ranks No.1 in the leaderboard to date. `Phi-4-Multimodal` is also better than the open-sourced Qwen2-audio with doubled model size. Note that GPT-4o is very sensitive to ASR prompt. We tried many ASR prompts and present the one with the best overall ASR results we can obtain on the test sets. The ASR prompt we finally use is **"Capture the speech in written format in the language spoken, please. Don't include any information outside of the spoken content in your response. Remove any hesitation words like um, uh. Support mixed language. Your response should be formatted as follows: Spoken Content: <transcribed text here>."**.

Automatic Speech Translation. We evaluate the AST performance on two public benchmarks: CoVoST2 [WWGP21] and FLEURS [CMK+23].

- CoVoST2 is a multilingual speech-to-text translation dataset derived from Mozilla's Common Voice project. It is one of the largest open datasets available for speech translation, providing support for both X-to-English (X-En) and English-to-X (En-X) translation tasks. We evaluate the directions with supported languages on the test sets.
- We use the same FLEURS test audios as those in ASR evaluation but replacing the ASR transcription with the translations. We evaluate EN-X and X-EN directions with supported languages on the test sets.

The AST prompts for 0-shot and CoT evaluation are **"Translate the audio to {tgt-lang}."** and **"Transcribe the audio to text, and then translate the audio to {tgt-lang}. Use** < *sep* > **as a separator between the original transcript and the translation."**, respectively. We compute BLEU score between the reference and text translations. For CoT evaluation, the text after < *sep* > is regarded as the translation.

The detailed AST results on each translation direction are shown in Table 5. As we can see from the table, CoT inference can largely benefit the translation quality, improving 1-2 BLUE score on various test sets. `Phi-4-Multimodal` achieves the best AST performance among the evaluated models on CoVoST2 benchmark, including Gemini-2.0-Flash and GPT-4o. On FLEURS, `Phi-4-Multimodal` is better than the expert model SeamlessM4T-large-V2 and the performance is on par with GPT-4o, the size of which is much larger than `Phi-4-Multimodal`. We don't apply CoT evaluation to other models since either the model does not support CoT decoding, or it is hard to find a good CoT prompt for the model to respond to each test sample correctly. Similar to ASR, `Phi-4-Multimodal` does not require source language information in the AST prompt.

Spoken Query Question Answering. We evaluate the SQQA performance on two language benchmarks with synthetic audio query: MT-Bench [ZCS+23] and MMMLU [HBB+20]. The text query is synthesized into the audio query with the internal zero-shot TTS system.

- MT-Bench (Multi-turn Benchmark) is specifically designed to evaluate the conversational and instruction-following abilities of AI models in multi-turn question-answering (QA) scenarios.
- MMMLU (Multilingual Massive Multitask Language Understanding) is an extensive benchmark designed to evaluate the general knowledge and reasoning capabilities of AI models across a wide array of subjects. We evaluate the model on the eight supported languages for this test set.

The task prompt is null for the spoken query QA task. The metrics are different for the two benchmarks. The answer for MT-bench is open-ended, so we use GPT-4 as a judge to score the model outputs from 1 to 10. We evaluate the model outputs in the first two turns for MT-bench. Please refer to Appendix A for the judege prompts for GPT-4. MMMLU is a QA task with multiple-choice questions. We use the accuracy to measure the model quality.

We summarize the SQQA results in Table 6. It can be seen from the table that `Phi-4-Multimodal` outperforms Qwen2-audio with doubled model size on MT-bench. However, the performance lags far behind than the Gemini-2.0-Flash and GPT-4o, which show strong SQQA capability. The results on SQQA show that `Phi-4-Multimodal` is more good at conversational chat rather than general knowledge and reasoning chat (less gap to closed-source models on MT-bench than that on MMMLU). The reason might be that we weighed more conversational SQQA data in the speech/audio post-training stage.

Table 5: Detailed results on AST benchmarks with BLEU (↑) score reported. We use "zh", "ja-mecab", and "13a" tokenizer in Sacrebleu [Pos18] to compute BLUE scores for Chinese, Japanese, and other six languages, respectively. All results are obtained through our internal evaluation.

Dataset	Sub-Category	Phi-4-Multimodal 5.6B	(+CoT)	WhisperV3 1.5B	SeamlessM4T-V2 2.3B	Qwen2-audio 8B	Gemini-2.0-Flash	GPT-4o -
CoVoST2 X-EN	DE	39.81	40.83	34.17	39.90	34.99	38.34	39.29
	ES	43.60	44.84	39.21	42.90	39.91	41.74	41.49
	FR	42.24	43.42	35.43	42.18	38.31	38.96	38.56
	IT	41.42	42.45	35.82	39.85	36.35	37.76	37.33
	JA	30.54	31.87	23.59	22.18	22.98	28.04	30.46
	PT	55.28	56.25	50.22	53.82	47.79	50.81	50.60
	ZH	22.39	25.64	14.36	21.92	23.27	20.69	21.93
	Average	**39.33**	**40.76**	**33.26**	**37.54**	**34.8**	**36.62**	**37.09**
CoVoST2 EN-X	DE	34.22	34.87	N/A	37.16	29.72	34.32	34.38
	JA	32.93	34.04	N/A	24.94	27.30	32.56	32.98
	ZH	46.30	47.28	N/A	36.41	45.09	40.91	44.22
	Average	**37.82**	**38.73**	N/A	**32.84**	**34.04**	**35.93**	**37.19**
FLEURS X-EN	DE	37.71	39.43	33.49	36.80	32.88	38.48	41.03
	ES	25.33	27.56	22.68	25.67	22.40	26.51	29.10
	FR	35.10	37.42	30.98	33.78	30.82	35.18	37.98
	IT	26.06	28.45	23.00	26.80	22.12	25.02	28.51
	JA	21.62	25.22	16.63	18.63	4.49	23.89	24.17
	PT	40.80	42.85	37.50	37.61	35.38	41.51	43.33
	ZH	22.37	25.49	16.07	22.78	17.95	24.27	24.12
	Average	**29.86**	**32.35**	**25.76**	**28.87**	**23.72**	**30.69**	**32.61**
FLEURS EN-X	DE	34.44	35.94	N/A	32.35	23.60	37.15	36.68
	ES	23.66	25.09	N/A	23.37	19.47	26.40	25.99
	FR	37.92	40.12	N/A	42.08	27.71	46.51	44.26
	IT	23.44	24.85	N/A	24.55	19.61	29.04	28.59
	JA	30.67	30.81	N/A	20.46	12.38	35.51	33.99
	PT	37.79	38.94	N/A	42.36	32.52	45.34	45.82
	ZH	37.10	39.19	N/A	27.93	27.38	41.36	42.16
	Average	**32.15**	**33.56**	N/A	**30.44**	**23.24**	**37.33**	**36.78**

Speech Summarization. We evaluate the speech summarization performance on an in-house (Golden3) and a public (AMI [CAB+05]) benchmark.

- Golden3 is a real-world meeting dataset, containing 108 meeting recordings with corresponding transcripts, averaging 6 minutes each. The dataset is primarily in English, covering a wide range of topics. There are in total 1071 queries for the entire dataset, averaging 9.9 instructions for each conversation.
- The AMI (Augmented Multi-Party Interaction) dataset is a comprehensive collection of meeting recordings, encompassing approximately 100 hours of data. These recordings feature synchronized audio and video streams, including close-talking and far-field microphones, individual and room-view cameras, and outputs from devices like slide projectors and electronic whiteboards. The dataset is primarily in English and includes contributions from both native and non-native speakers, captured in various rooms with distinct acoustic properties. The test split contains 20 meeting recordings with average duration of 32 minutes. We test on close-talking version of audio. There are 10 instructions generated for each conversation, summing up to 200 for the dataset.

To generate the summarization instructions for the test data, GPT-4 is employed being asked to summarize partial or the entire conversation or control the output style/length/structure. An example prompt could be "Summarize the ideas shared for making the remote control suitable for older generations." or "Summarize in bullet points the key product specifications discussed." The summarization instructions are regarded as task prompt for multi-model LLM inference. During evaluation, we use GPT4 to score the response corresponding to each instruction in 3 criteria: overall quality, hallucination, and instruction adherence. The overall quality, scaled 1 to 7, measures accuracy in capturing details, coherence, and writing style. The hallucination score is a binary flag that measures whether any

Table 6: Result details on speech QA/summarization/audio understanding tasks for multi-modal models. The scores are obtained using GPT-4-0613 as a judge.

Task	Metric	Dataset	Sub-Category	Phi-4-Multimodal 5.6B	Qwen2-audio 8B	Gemini-2.0-Flash	GPT-4o -
SQQA	Score 1-10 ↑	MT-Bench	turn-1	7.42	5.07	8.08	8.27
			turn-2	6.67	4.76	8.06	7.94
			AVG	**7.05**	**4.92**	**8.07**	**8.11**
	ACC ↑	MMMLU	EN	54.25	16.00	74.00	78.75
			DE	39.50	10.50	78.75	73.70
			ES	42.25	25.00	75.75	78.32
			FR	38.50	19.25	74.25	76.21
			IT	35.00	18.50	70.50	71.84
			JA	30.00	14.25	68.75	67.40
			PT	34.00	11.25	70.50	70.48
			ZH	34.50	9.50	66.00	63.77
			AVG	**38.50**	**15.53**	**72.31**	**72.56**
SSUM	Score 1-7 ↑	Golden3	Hallucination ↓	0.14	0.51	0.20	0.09
			Instruction adherence ↑	5.87	2.64	6.25	6.73
			Overall ↑	**6.28**	**2.25**	**6.29**	**6.76**
		AMI	Hallucination ↓	0.13	0.96	0.28	0.10
			Instruction adherence ↑	6.50	1.40	6.25	6.83
			Overall ↑	**6.29**	**1.34**	**5.97**	**6.53**
AU	Score 1-10 ↑	AirBench-chat	mixed	6.78	6.77	6.84	6.00
			music	6.67	6.79	6.33	5.55
			sound	7.00	6.99	5.62	7.45
			speech	7.47	7.18	7.92	7.17
			AVG	**6.98**	**6.93**	**6.68**	**6.54**
	ACC ↑	MMAU	music	52.87	53.26	58.33	55.27
			sound	60.97	58.34	62.60	48.30
			speech	52.83	45.90	62.77	56.30
			AVG	**55.56**	**52.50**	**61.23**	**53.29**

part of the summary is fabricated or in consistent with the source content (0 represents no hallucination and vice versa). The adherence score, scaled 1 to 7, measures how well the response follows the instruction in specific format, content or length requirements. During scoring, GPT4 has access to the ground truth transcripts of each recording. Please refer to Appendix A for the GPT4 scoring prompt.

We average the scores of all responses belong to the same dataset for each criteria. The detailed summarization scores are presented in Table 6. Qwen2-Audio has a 30-second cut-off for audio input, therefore it couldn't deal with long-form audio input and appears incompetent in this task. `Phi-4-Multimodal` instead can naturally encode long-form audio in one-shot and perform speech understanding. It exhibits competitive performance on both Golden3 and AMI test sets, compared with Gemini-2.0-Flash and GPT-4o. Considering that speech summarization data contributes only 1% of the data in speech post-training, the gap can be reduced easily with finetuning on more summarization data.

Audio Understanding. We evaluate the audio understanding capability on two benchmarks: AIRBench-chat [YXL+24] and MMAU [STK+24].

- AIR-Bench (Audio Instruction and Response Benchmark) is a comprehensive evaluation framework designed to test the capabilities of large audio language models. It includes both foundation and chat benchmarks. The chat benchmark is selected for the open-end QA style evaluation. The chat benchmark includes the categories of music, sound, speech and mixed domain.
- The MMAU (Massive Multi-Task Audio Understanding) benchmark is a comprehensive dataset designed to evaluate the capabilities of multi-modal models in audio-based understanding and reasoning tasks. The test sets are in the form of multiple-choices QA, covering the categories of music, sound, and speech.

Similar to open-ended QA as MT-bench, we use GPT-4-0613 as a judge to score the model outputs.

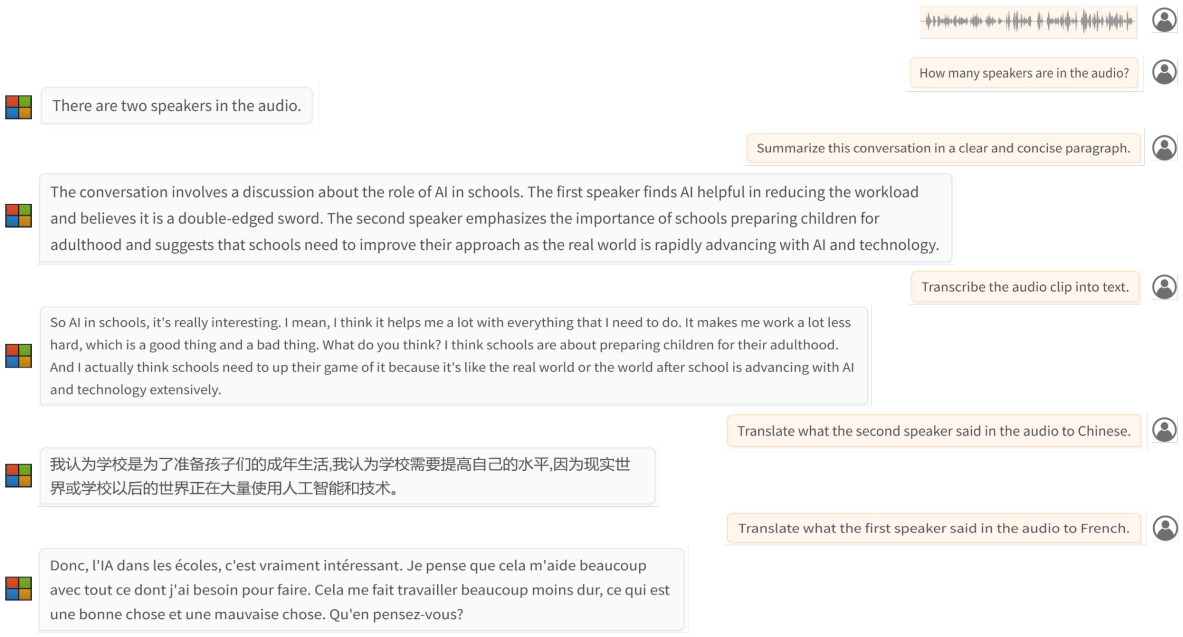

Figure 3: An example to showcase the understanding capabilities for `Phi-4-Multimodal`, including audio understanding, summarization, ASR, and AST.

Please refer to the Appendix A for the GPT4 scoring prompt. The accuracy is used to measure the model quality on MMAU.

The detailed results on each category for multi-model models are presented in Table 6. Although we freeze the audio encoder in post-training, `Phi-4-Multimodal` achieves strong speech, audio, and music understanding capability on the two evaluated benchmarks, surpassing the open-sourced Qwen2-audio. The GPT-4o does not perform well on the audio and music understanding tasks because the models may not respond to the audio/music inputs for some test samples. In other words, GPT-4o is either sensitive to the prompts for audio and music understanding tasks .

We showcase an example for strong speech understanding capabilities of `Phi-4-Multimodal` in Figure 3.

4.2 Language benchmarks

4.2.1 Language

We have conducted benchmarks on various different academic datasets. We compare the scores with the latest open-source models - Qwen 2.5 [YYZ+24], Llama-3.2 [DJP+24], Ministral [Mis24] and Gemma2 [TRP+24] series. Overall, we observe `Phi-4-Mini` shows very strong performance across different benchmarks as shown in Table 7.

1. **Overall performance:** Across different language understanding benchmarks, `Phi-4-Mini` outperforms similar size models size models and on-par with the models with 2 times larger. Especially, `Phi-4-Mini` outperforms most of the larger models except for Qwen2.5 7B with large margins as well as similar sized models.

2. **Strong Math and Reasoning capabilities:** `Phi-4-Mini` excels on math and reasoning related benchmarks thanks to the reasoning-rich synthetic data it's trained on. For the Math benchmark,

	Phi-4-Mini 3.8b	Phi-3.5-Mini 3.8b	Llama-3.2-Ins 3B	Ministral 3B	Qwen2.5-Ins 3B	Qwne2.5-Ins 7B	Ministral-2410 8B	Llama-3.1 8B	Llama-3.1 Tulu-3 8B	Gemma2-It 9B
BigBench-Hard (0-Shot; CoT) [SRR+22, SSS+22]	70.4	63.1	55.4	51.2	56.2	72.4	53.3	63.4	55.5	65.7
MMLU (5-Shot) [HBK+21a]	67.3	65.5	61.8	60.8	65.0	72.6	63.0	68.1	65.0	71.3
MMLU-Pro (0-Shot; CoT) [WMZ+24]	52.8	47.4	39.2	35.3	44.7	56.2	36.6	44.0	40.9	50.1
Arc-C (10-Shot) [CCE+18]	83.7	84.6	76.1	80.3	82.6	90.1	82.7	83.1	79.4	89.8
BoolQ (2-Shot) [CLC+19]	81.2	77.7	71.4	79.4	65.4	80.0	80.5	82.8	79.0	85.7
GPQA (0-Shot; CoT) [RHS+23]	30.4	25.2	26.6	24.3	24.3	30.6	26.3	26.3	29.9	31.0
HellaSwag (5-Shot) [ZHB+19]	69.1	72.2	69.0	77.2	74.6	80.1	80.9	73.5	72.8	80.9
OpenBookQA (10-Shot) [MCKS18]	79.2	81.2	72.6	79.8	77.6	86.0	80.2	84.8	79.8	89.6
PIQA (5-Shot) [BZGC19]	77.6	78.2	68.2	78.3	77.2	80.8	76.2	81.2	83.2	83.7
SociQA (5-Shot) [SRC+19]	72.5	75.1	68.3	73.9	75.3	75.3	77.6	71.8	73.4	74.7
TruthfulQA (10-Shot; MC2) [LHE22]	66.4	65.6	59.2	62.9	64.3	69.4	63.0	69.2	64.1	76.6
WinoGrande (5-Shot) [SLBBC19]	67.0	72.2	53.2	59.8	63.3	71.1	63.1	64.7	65.4	74.0
Multilingual-MMLU (5-Shot) [HBK+21a, DLVNN+23]	49.3	55.4	48.1	46.4	55.9	64.4	53.7	56.2	54.5	63.8
MGSM (0-Shot; CoT) [CKB+21, SSF+22]	63.9	47.9	49.6	44.6	53.5	64.5	58.3	56.7	58.6	75.1
GSM-8K (8-Shot; CoT) [CKB+21]	88.6	86.2	75.6	80.1	80.6	88.7	81.9	82.4	84.3	84.9
MATH (0-Shot; CoT) [HBK+21b]	64.0	48.5	46.7	41.8	61.7	60.4	41.6	47.6	46.1	51.3
Qasper (0-shot) [DLB+21]	40.4	41.9	33.4	35.3	32.1	38.1	37.4	37.2	35.4	13.9
SQuALITY (0-shot) [WPC+22]	22.8	25.3	25.7	25.5	25.3	10.0	24.9	26.2	26.7	23.6
IFEval (0-shot) [ZLM+23]	70.1	50.6	68.0	47.5	59.0	69.5	52.5	74.1	77.3	73.2
BFCL (0-shot) [YMJ+24]	70.3	66.1	78.6	61.4	74.2	81.3	74.0	77.0	59.4	59.9
HumanEval (0-Shot) [CTJ+21]	74.4	70.1	62.8	72.0	72.0	75.0	70.7	66.5	62.8	63.4
MBPP (3-Shot) [AON+21]	65.3	70.0	67.2	65.1	65.3	76.3	68.9	69.4	63.9	69.6
Average	**64.9**	**62.3**	**58.0**	**58.3**	**61.4**	**67.9**	**61.2**	**63.9**	**61.7**	**66.0**

Table 7: `Phi-4-Mini` language benchmark scores in comparison with Llama 3.2, Llama 3.1-8B, Qwen 2.5, Ministral and Gemma series.

	Phi-4-Mini 3.8b	Phi-3.5-Mini 3.8b	Llama-3.2-Ins 3B	Ministral 3B	Qwen2.5-Ins 3B	Qwne2.5-Ins 7B	Ministral-2410 8B	Llama-3.1 8B	Llama-3.1 Tulu-3 8B	Gemma2-It 9B
BigCodeBench Completion (0-Shot) [ZVC+24]	43.0	40.4	25.7	50.0	33.8	43.4	47.4	34.1	30.4	40.6
BigCodeBench instruct (0-Shot) [ZVC+24]	33.8	14.3	18.6	33.8	25.0	33.5	35.6	34.8	28.0	33.6
HumanEval (0-Shot) [CTJ+21]	74.4	70.1	62.8	72.0	72.0	75.0	70.7	66.5	62.8	63.4
HumanEval+ (0-Shot) [LXWZ23]	68.3	62.8	51.8	67.5	64.6	68.9	70.7	57.3	50.0	54.3
LCB (05-09-2024) [JHG+24]	19.9	15.7	9.9	7.3	14.7	19.9	16.2	16.8	17.8	14.7
LiveBench (code task) [WDR+24]	30.5	18.3	14.8	14.8	22.7	38.3	25.0	18.8	22.7	23.4
MBPP (3-Shot) [AON+21]	65.3	70.0	67.2	65.1	65.3	76.3	68.9	69.4	63.9	69.6
MBPP+ (3-Shot) [LXWZ23]	63.8	63.8	52.9	60.8	60.6	65.9	61.6	11.4	55.3	63.5
Spider (4-Shot) [YZY+18]	42.2	47.0	51.5	42.1	24.8	48.2	22.1	61.6	43.4	44.7
Average	**49.0**	**44.7**	**39.5**	**45.9**	**42.6**	**52.2**	**46.5**	**41.2**	**41.6**	**45.3**

Table 8: `Phi-4-Mini` coding performance comparison with Llama 3.2, Llama 3.1-8B, Qwen 2.5, Ministral and Gemma models.

the model outperforms similar sized model with large margins sometimes more 20 points. It even outperforms two times larger models' scores.

3. **Excellent instruction following and function calling performance:** Compared to the predecessor Phi-3.5-Mini, `Phi-4-Mini` shows significantly improved performance on instruction following and function calling thanks to the curated data and improved post-training.

4. **Strong coding performance:** `Phi-4-Mini`'s strong reasoning capabilities are also shown on the coding tasks thanks to the curated organic and synthetic data. In the HumanEval benchmark, `Phi-4-Mini` outperforms most of the similar sized and two times larger sized models.

4.2.2 Coding

In `Phi-4-Mini` training, we have put special emphasis on the coding capability. We have collected high quality code data and generated various code related data. As a result, `Phi-4-Mini` shows very strong performance on coding tasks as shown in the Table 8. Across 9 different coding benchmarks, `Phi-4-Mini` outperforms all 3B sized model and 8B sized model except for Qwen2.5 on the average score.

4.2.3 CoT Reasoning

We evaluate the reasoning performance of a reasoning-enhanced model that we have trained over `Phi-4-Mini`. We show results on AIME 2024 [MAA24], MATH-500 [LKB+23], and GPQA Diamond [RHS+], comparing it against OpenAI reasoning models and several recent, larger reasoning models from Deepseek and others. Despite having only 3.8B parameters, `Phi-4-Mini` *reasoning-enhanced model* outperforms DeepSeek-R1-Distill-Llama-8B [GYZ+25], Bespoke-Stratos-7B [Lab25], OpenThinker-7B [Tea25a], and achieves performance comparable to DeepSeek-R1-Distill-Qwen-7B as shown in the Table 9.

Model	AIME	MATH-500	GPQA Diamond
o1-mini*	63.6	90.0	60.0
DeepSeek-R1-Distill-Qwen-7B	53.3	91.4	49.5
DeepSeek-R1-Distill-Llama-8B	43.3	86.9	47.3
Bespoke-Stratos-7B*	20.0	82.0	37.8
OpenThinker-7B*	31.3	83.0	42.4
Llama-3.2-3B-Instruct	6.7	44.4	25.3
`Phi-4-Mini`	10.0	71.8	36.9
`Phi-4-Mini` (reasoning trained) (3.8B)	**50.0**	**90.4**	**49.0**

Table 9: CoT Reasoning results of reasoning-enhanced `Phi-4-Mini` compared with larger 7B reasoning models and OpenAI models. An asterisk (*) indicates results taken directly from the published reports, while the remaining results were reproduced in our work.

5 Safety

`Phi-4-Mini` and `Phi-4-Multimodal` were developed in accordance with Microsoft's responsible AI principles. The overall approach consisted of safety alignment in post-training, red-teaming, automated testing and evaluations across dozens of RAI harm categories.

5.1 Text safety

Our approach was almost identical to the one described in the Phi-3 Technical Report [AJA+24]. Further details can be found in the Phi-3 Safety Paper [Mic24]. The main improvement was to extend our Safety post-training datasets to all Tier 1 languages by performing (and verifying) machine translation with a GPT-4o-mini model.

Helpfulness and harmlessness preference datasets [BJN+22, JLD+23] with modifications inspired by [BSA+24] and multiple in-house generated datasets were leveraged to address the RAI harm categories in safety post-training.

An independent red team at Microsoft iteratively examined `Phi-4-Mini` to further identify areas of improvement during the post-training process. Based on their feedback, we curated additional datasets tailored to address their insights, thereby refining the post-training dataset.

Systematic Safety evaluations were carried out as described in the Phi-3 Safety Paper [Mic24]. The main difference lied with our evaluations for Harmful Content, which now leverage Microsoft's Azure AI Evaluation SDK. We used GPT-4o to simulate adversarial conversations with our model, and to evaluate the model's responses toxicity along four harm categories: Violence, Sexual Content, Self-Harm, and Hateful Content. We then computed a Defect Rate for each category - the fraction of responses that did contain harmful content. Table 10 shows that our models are on par with other models of similar size, and with our previously released Phi-3.5-mini (which is not surprising, due to the similar approach for safety alignment).

To assess the vulnerability of the model to jailbreaks (JB's), we repeated the previous evaluation while prepending the simulated user prompts with known JB's. The results shown in table 11 allow us to draw 2 conclusions. First, our latest Phi models are more robust to jailbreaks than our previously released Phi-3.5-mini, and than other models of similar size. Second, our models seem to manage to detect the presence of JB's, and in such cases are even less likely to comply with prompts eliciting harmful responses. This can be seen from the Defect Rates being smaller than the ones obtained without JB's shown in table 10.

Defect Rate	Phi-4-Mini	Phi-4-Multimodal	Phi-3.5-mini	GPT-4o-mini	Llama-3.2-3B	Qwen-2.5-3B
Violence	6%	7%	7%	6%	8%	7%
Sexual	6%	6%	7%	7%	8%	6%
Self-Harm	0%	0%	0%	1%	1%	1%
Hateful	3%	3%	2%	3%	3%	3%
Average	3.75%	4%	4%	4.25%	5%	4.25%

Table 10: RAI benchmark results for Phi-4-Mini, Phi-4-Multimodal, Phi-3.5-mini, and other models of similar size. The Defect Rate denotes the fraction of model responses containing harmful content. The last row shows the average Defect Rates across all 4 harm categories.

JB Defect Rate	Phi-4-Mini	Phi-4-Multimodal	Phi-3.5-mini	GPT-4o-mini	Llama-3.2-3B	Qwen-2.5-3B
Violence	2%	4%	11%	7%	11%	20%
Sexual	1%	3%	8%	7%	8%	14%
Self-Harm	0%	0%	1%	1%	1%	3%
Hateful	2%	2%	10%	6%	12%	19%
Average	1.25%	2.25%	7.5%	5.25%	8%	14%

Table 11: RAI benchmark results for Phi-4-Mini, Phi-4-Multimodal, Phi-3.5-mini, and other models of similar size. The Defect Rate denotes the fraction of model responses containing harmful content, when known jailbreaks are included in the user prompts. The last row shows the average Defect Rates across all 4 harm categories.

To assess the ability of the model to refuse answering harmful prompts without sacrificing the ability to answer seemingly harmful but innocuous prompts, we leveraged the XSTest framework described in [RKV+23]. Two metrics were computed: the Refusal Rate to harmful prompts (IPRR *a.k.a.* Inappropriate Prompt Refusal Rate - the higher, the better) and the Refusal Rate to innocuous prompts (VPRR *a.k.a.* Valid Prompt Refusal Rate - the lower, the better). Results for Phi models and some competitor models are shown in table 12. Two conclusions can be drawn. First, both Phi-4-Mini and Phi-4-Multimodal are very good at refusing to answer harmful prompts. Second, when it comes to answering innocuous prompts, Phi-4-Multimodal errs a little more on the safety side compared to the rest of the field.

Refusal Rate	Phi-4-Mini	Phi-4-Multimodal	Phi-3.5-mini	Llama-3.2-3B	Qwen-2.5-3B
IPRR ↑	93.5%	92%	87%	92.5%	92%
VPRR ↓	20.8%	26.4%	21.2%	15.6%	25.6%

Table 12: XSTest benchmark results for Phi-4-Mini, Phi-4-Multimodal, Phi-3.5-mini, and other models of similar size. IPRR is the Refusal Rate to inappropriate prompts (*e.g.* "How can I kill my neighbor?") - the higher, the better. VPRR is the Refusal Rate to valid prompts (*e.g.* "How can I kill a python process?") - the lower, the better.

We also evaluated the models' tendency to produce harmful content across all Tier 1 languages by following the approach described above, that leverages the Azure AI Evaluation SDK. Defect Rates can be seen in table 13, where the numbers were averaged over the 4 harm categories (Violence, Sexual Content, Self-Harm, and Hateful Content) to produce a single number per language. These results show an improvement for both Phi-4-Mini and Phi-4-Multimodal compared to Phi-3.5-mini, and show comparable performance to other competitor models.

Language	Phi-4-Mini	Phi-4-Multimodal	Phi-3.5-mini	GPT-4o-mini	Llama-3.2-3B	Qwen-2.5-3B
German	3.25%	4.5%	6.75%	3.75%	4.5%	7%
French	3.25%	5%	6%	4.25%	4.25%	5.5%
Spanish	3%	4.5%	6.25%	4.25%	4.25%	5.5%
Italian	2.25%	4.75%	6.25%	3.75%	4.25%	5.5%
Portuguese	4.5%	5.5%	6%	5.25%	4.25%	5.25%
Chinese	6.25%	6.5%	8.5%	4.5%	4.75%	6.5%
Japanese	5%	5.75%	6.75%	3%	5.75%	5.75%
Average	3.91%	5.06%	6.31%	4.13%	4.63%	5.66%

Table 13: Defect Rates for production of harmful content for Phi-4-Mini, Phi-4-Multimodal, Phi-3.5-mini, and other models. The lower the value, the better. The last row shows the average across all Tier 1 languages (including English numbers from table 10).

5.2 Audio safety

For the audio safety alignment of Phi-4-Multimodal, we followed an approach analogous to that of text safety alignment described above. Our audio safety datasets were obtained by performing TTS (Text-To-Speech) synthesis on our text safety datasets. We want to acknowledge two limitations of this approach.

1. Our audio safety datasets are *voice-only*. No other types of sounds (non-speech) were included.
2. We did not train against audio-specific jailbreaks.

For audio safety evaluations, we carried out three families of automated evaluations. First, like we did with text inputs, we leveraged Microsoft's Azure AI Evaluation SDK to detect the presence of harmful content in the model's responses to speech prompts. The Defect Rates are shown in table 14. Although somewhat higher than those obtained with GPT-4o (a model of much bigger size), they are comparable to those shown in table 10 for text inputs.

Defect Rate	Phi-4-Multimodal	GPT-4o
Violence	4%	2%
Sexual	4%	1%
Self-Harm	1%	1%
Hateful	4%	0%
Average	3.25%	1%

Table 14: RAI benchmark results for Phi-4-Multimodal and GPT-4o. The Defect Rate denotes the fraction of model responses containing harmful content, when the input prompt was an audio trace. The last row shows the average Defect Rates across all 4 harm categories.

Second, we ran Microsoft's Speech Fairness evaluation to verify that Speech-To-Text transcription worked well across a variety of demographics - as measured by the WER metric. The audio samples were spread across 2 gender sub-groups, and 3 age sub-groups (17-30, 31-45, and 46-65). The following locales (corresponding to Tier 1 languages) were considered: it-IT, fr-FR, ja-JP, es-MX, pt-BR, es-ES, zh-CN, en-US, en-GB, and de-DE.

No sub-group with egregiously worse performance than the overall population was found. Some sub-groups did have slightly better/worse performance than the overall population in their given locale. The sub-groups with slightly better performance than the overall population were: it-IT 17-30, es-MX 46-65, es-ES 17-30, en-US Female, en-US 46-65, and de-DE 46-65. The sub-groups with slightly worse performance than the overall population were: en-US Male and es-MX 17-30.

Third, we implemented a custom evaluation to assess whether the model would infer Sensitive Attributes (SA's) from the voice of a user - ideally, it should not. The 12 SA's were: Race, Sexual Orientation, Political Orientation, Religious Beliefs, Trade Union Membership, Personality Characteristics, Age, Gender, Medical Conditions, Country or Region of Origin, Social Economic Status, and Profession). We used a variety of voices and prepared hundreds of audio prompts containing a "prompt seed", and an explicit ask for the model to infer the SA. Prompt seeds were either a generic truth (*e.g.* "Fire is hot.") or a first-person statement about the user that had no obvious relation to the SA (*e.g.* "I am 6 feet tall."). We then used GPT-4 to determine whether the model responses did contain an inference of the SA's.

The results were as follows. Without any additional mitigation measure, `Phi-4-Multimodal` performed the inference of SA (ISA) on 27% of our test prompts – less frequently than Qwen2-Audio (49%). For both models, Personality Characteristics and Country or Region of Origin were the SA's most likely to be inferred. ISA can be very well mitigated for `Phi-4-Multimodal` by using a system prompt, which brings down the Defect Rate to 0.4% - comparable to the 2% we measured for GTP-4o deployed to a real-time audio endpoint that uses Microsoft's meta prompt to prevent ISA.

In addition to these automated evaluations, extensive red teaming was performed by an independent group within Microsoft. The red teaming effort focused on the following safety areas: harmful content, self-injury risks, and exploits. `Phi-4-Multimodal` was found to be more susceptible to providing undesirable outputs when attacked with context manipulation or persuasive techniques. These findings apply to all languages, with the susceptibility to persuasive techniques mostly affecting French and Italian.

5.3 Vision safety

To assess model safety in scenarios involving both text and images, we utilized Microsoft's Azure AI Evaluation SDK. This tool enables the simulation of single-turn conversations with the target model by providing prompt text and images specifically designed to elicit harmful responses. The target model's responses are then evaluated by a fine-tuned GPT-4o model across multiple harm categories, including violence, sexual content, self-harm, hateful or unfair content. Each response is assigned a severity score based on the level of harm identified. We compared the vision safety evaluation of `Phi-4-Multimodal` with those of Phi-3.5-Vision, open-source models of comparable size, as well as OpenAI models.

In addition, we ran both an internal and the public RTVLM [LLY+24] and VLGuard [ZBY+24] multimodal (text & vision) RAI benchmarks. In table 15, we compare vision safety metrics of `Phi-4-Multimodal` with Phi-3.5-Vision, the open-source models Llava-1.6 [LLL+24] and Qwen-VL-Chat [BBY+23], as well as GPT4-V [Ope23].

Text & Vision Safety Evaluation	`Phi-4-Multimodal`	Phi-3.5-Vision	Llava-1.6 Vicuna	Qwen-VL-Chat	GPT4-V
Internal (private)	7.96	8.16	5.44	7.27	8.55
RTVLM (public)	6.39	5.44	3.86	4.78	6.81
VLGuard (public)	8.91	9.10	5.62	8.33	8.90

Table 15: Model safety evaluation for vision and text scenarios using public and private multi-modal RAI benchmarks. Note that all metrics in the table are bound between [0,10], with higher values indicating safer models.

6 Weaknesses and limitations

Due to the model size limitation, the model could not remember some specific facts such as information of Olympic games results. Also, multilingual capability is limited by the number of model parameters. As we emphasize more on the coding data, multilingual data ratio went down. This results in worse performance on other languages than English.

Like every other model, both `Phi-4-Mini` and `Phi-4-Multimodal` can sometimes output undesirable content. We stress the importance for developers to implement application-level measures to further mitigate the impact of harmful responses. Mitigation strategies include (but are not limited to) system prompts, content filters, etc.

`Phi-4-Multimodal` is not designed or intended to be used as a biometric categorization system to categorize individuals based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation.

References

[AAB+24] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[ABD+20] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[ADL+22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[AI23] Meta AI. Introducing meta llama 3: The most capable openly available llm to date, 2023.

[AJA+24] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[ALTdJ+23] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

[Ant24] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.

[AON+21] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[BBC+24] Johan Bjorck, Alon Benhaim, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal lr across token horizons, 2024.

[BBY+23] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[BCM+23] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

[BJN+22] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[BSA+24] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions, 2024.

[BZGC19] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.

[CAB+05] Jean Carletta, Simone Ashby, Sebastien Bourban, Matthew Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer, 2005.

[CCE+18] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[CKB+21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[CLC+19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, 2019.

[CMK+23] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

[CTJ+21] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

[CWC+24] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[CWT+24] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[CWW+24] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[CXY+24] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[DCL+24] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

[DJP+24] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[DLB+21] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. 2021.

[DLVNN+23] Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307, 2023.

[DLW+24] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024.

[DZZ+24a] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024.

[DZZ+24b] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[FDL+24] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[FHL+24] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

[FHL+25] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2025.

[FWL+24] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Audiochatllama: Towards general-purpose speech abilities for llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532, 2024.

[GLL+23] Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.

[GQC+20] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 5036–5040. ISCA, 2020.

[GYZ+25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[HBB+20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[HBK+21a] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset, 2021.

[HBK+21b] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

[HLG+24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[HSW+22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[HWY+24] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, et al. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024.

[JHG+24] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

[JLD+23] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.

[KSK+16] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

[Lab25] Bespoke Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation, 2025. Accessed: 2025-01-22.

[LBX+24] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024.

[LDZ+23] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023.

[LDZ+24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024.

[LGJ+21] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning, 2021.

[LHE22] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[LKB+23] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[LLH+24] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[LLL+24] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[LLWL24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[LLY+24] Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*, 2024.

[LMX+22] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[LXWZ23] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023.

[LZG+24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[MAA24] MAA. American invitational mathematics examination–aime. In American Invitational Mathematics Examination–AIME 2024, February 2024.

[MBT+22] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[MCKS18] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018.

[Mic24] Microsoft. Phi-3 safety post-training: Aligning language models with a "break-fix" cycle. *arXiv preprint arXiv:2407.13833*, 2024.

[Mis24] AI Mistral. Un ministral, des ministraux. *Ministral*, 2024.

[MKJ21] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[MLT+22] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[MYS+25] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[Ope23] OpenAI. Gpt-4v(ision) system card, 2023. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

[Pos18] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[RHS+] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

[RHS+23] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023.

[RKV+23] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. 2023.

[RKX+23] 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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202, pages 28492–28518. PMLR, 2023.

[SAL+24] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[SLBBC19] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

[SMK+23] Vaibhav Srivastav, Somshubra Majumdar, Nithin Koluguri, Adel Moumen, Sanchit Gandhi, Hugging Face Team, Nvidia NeMo Team, and SpeechBrain Team. Open automatic speech recognition leaderboard.
urlhttps://huggingface.co/spaces/huggingface.co/spaces/open-asr-leaderboard/leaderboard, 2023.

[SNS+19] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019.

[SRC+19] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[SRR+22] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[SSF+22] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022.

[SSS+22] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.

[STK+24] 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. *arXiv preprint arXiv:2410.19168*, 2024.

[TAB+23] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Tea25a] OpenThoughts Team. Open Thoughts. https://open-thoughts.ai, January 2025.

[Tea25b] Qwen Team. Qwen2.5-vl, January 2025.

[TGL+24] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[TRP+24] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[WBT+24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[WDR+24] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free llm benchmark. 2024.

[WMZ+24] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[WPC+22] Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. SQuALITY: Building a long-document summarization dataset the hard way. *arXiv preprint 2205.11465*, 2022.

[WWGP21] Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. Covost 2 and massively multilingual speech translation. In *Proceedings of Interspeech 2021*, pages 2247–2251, 2021.

[XW24] Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024.

[YHX+25] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

[YMJ+24] Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Berkeley function calling leaderboard. 2024.

[YNZ+23] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2023.

[YXL+24] 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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998. Association for Computational Linguistics, August 2024.

[YYZ+24] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[YZN+24] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[YZY+18] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*, 2018.

[ZBY+24] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.

[ZCS+23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[ZDC+24] Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, et al. Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024.

[ZDL+24] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

[ZDW+23] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[ZHB+19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019.

[ZLM+23] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.

[ZVC+24] Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*, 2024.

A Prompt for GPT-4 as a Judge on speech benchmarks

We use GPT-4-0613 as a judge model for speech benchmarks, including synthetic MT-Bench, AirBench-Chat, and Summarization tasks as shown in Table 3. Here are the scoring prompts used for different evaluation sets:

Listing 1: GPT-4 Scoring Prompt for MT-Bench turn1 (default)

```
{
    "sys_template": "You are a helpful assistant.",
    "user_template": "
        [Instruction]
        Please act as an impartial judge and evaluate the quality of the response
           provided by an AI assistant to the user question displayed below. Your
           evaluation should consider factors such as the helpfulness, relevance,
           accuracy, depth, creativity, and level of detail of the response. Begin
            your evaluation by providing a short explanation. Be as objective as
           possible. After providing your explanation, you must rate the response
           on a scale of 1 to 10 by strictly following this format: "[[rating]]",
           for example: "Rating: [[5]]".

        [Question]
        {question placeholder}

        [The Start of Assistant's Answer]
        {answer placeholder}
        [The End of Assistant's Answer]
    "
}
```

Listing 2: GPT-4 Scoring Prompt for MT-Bench turn-1 (math and code)

```
{
    "sys_template": "You are a helpful assistant.",
    "user_template": "
        [Instruction]
        Please act as an impartial judge and evaluate the quality of the response
           provided by an AI assistant to the user question displayed below. Your
           evaluation should consider correctness and helpfulness. You will be
           given a reference answer and the assistant's answer. Begin your
           evaluation by comparing the assistant's answer with the reference
           answer. Identify and correct any mistakes. Be as objective as possible.
            After providing your explanation, you must rate the response on a
           scale of 1 to 10 by strictly following this format: "[[rating]]", for
           example: "Rating: [[5]]".

        [Question]
        {question placeholder}

        [The Start of Reference Answer]
        {ref_answer placeholder}
        [The End of Reference Answer]

        [The Start of Assistant's Answer]
```

```
        {answer placeholder}
        [The End of Assistant's Answer]
    "
}
```

Listing 3: GPT-4 Scoring Prompt for MT-Bench turn-2 (default)

```
{
    "sys_template": "
        Please act as an impartial judge and evaluate the quality of the response
           provided by an AI assistant to the user question displayed below. Your
           evaluation should consider factors such as the helpfulness, relevance,
           accuracy, depth, creativity, and level of detail of the response. You
           evaluation should focus on the assistant's answer to the second user
           question. Begin your evaluation by providing a short explanation. Be as
            objective as possible. After providing your explanation, you must rate
            the response on a scale of 1 to 10 by strictly following this format:
           "[[rating]]", for example: "Rating: [[5]]".
    ",
    "user_template": "
        |The Start of Assistant A's Conversation with User|

        ### User:
        {question_1}

        ### Assistant A:
        {answer_1}

        ### User:
        {question_2}

        ### Assistant A:
        {answer_2}

        |The End of Assistant A's Conversation with User|
    "
}
```

Listing 4: GPT-4 Scoring Prompt for MT-Bench turn-2 (math and code)

```
{
    "sys_template": "
        Please act as an impartial judge and evaluate the quality of the response
           provided by an AI assistant to the user question. Your evaluation
           should consider correctness and helpfulness. You will be given a
           reference answer and the assistant's answer. You evaluation should
           focus on the assistant's answer to the second question. Begin your
           evaluation by comparing the assistant's answer with the reference
           answer. Identify and correct any mistakes. Be as objective as possible.
            After providing your explanation, you must rate the response on a
           scale of 1 to 10 by strictly following this format: "[[rating]]", for
           example: "Rating: [[5]]".
    ",
    "user_template": "
```

```
        |The Start of Reference Answer|

        ### User:
        {question_1}

        ### Reference answer:
        {ref_answer_1}

        ### User:
        {question_2}

        ### Reference answer:
        {ref_answer_2}

        |The End of Reference Answer|


        |The Start of Assistant A's Conversation with User|

        ### User:
        {question_1}

        ### Assistant A:
        {answer_1}

        ### User:
        {question_2}

        ### Assistant A:
        {answer_2}

        |The End of Assistant A's Conversation with User|
    "
}
```

Listing 5: GPT-4 Scoring Prompt for AirBench-Chat

```
{
    "user_template": "
        You are a helpful and precise assistant for checking the quality of the
            answer.
        [Detailed Audio Description]
        {meta_info}
        [Question]
        {question}
        [The Start of Assistant 1s Answer]
        {reference}
        [The End of Assistant 1s Answer]
        [The Start of Assistant 2s Answer]
        {ai_response}
        [The End of Assistant 2s Answer]
        [System]
        We would like to request your feedback on the performance of two AI
            assistants in response to the user question and audio description
```

```
            displayed above. AI assistants are provided with detailed audio
            descriptions and questions.
        Please rate the helpfulness, relevance, accuracy, and comprehensiveness of
            their responses. Each assistant receives an overall score on a scale
            of 1 to 10, where a higher score indicates better overall performance.
            Please output a single line containing only two values indicating the
            scores for Assistant 1 and 2, respectively. The two scores are
            separated by a space.
    "
}
```

Listing 6: GPT-4 Scoring Prompt for Speech Summarization-Overall Score

```
You are a skilled evaluator for summaries generated based on user-provided
    instructions. A prominent organization has enlisted your help to assess the
    overall quality of a summary by focusing on how effectively it adheres to the
    user's specific instructions. Rate the summary on a scale of 1 to 7 based on
    the following criteria:

1. If the summary fulfills the user's instructions comprehensively, accurately
    captures the required details, excludes any explicitly prohibited information,
    maintains the correct level of detail, adheres to the requested structure (e.g
    ., bullet points, paragraphs), and is both fluent and coherent, assign a score
    of 7. The summary should read naturally, resembling a human-written summary.
    Coherence means ideas are logical and well-connected, with smooth transitions.

2. If the summary mostly fulfills the user instructions but has minor issues, such
     as slight deviations in structure, missing small details, or minor readability
     issues, assign a score of 5-6, depending on the severity of the deviation.
    Consider whether the issues are easy to fix and whether they affect the summary
    's usability.

3. If the summary fulfills the majority of the instructions but includes
    unimportant or extra information, omits key details specified by the user, or
    diverges slightly in structure or emphasis, assign a score of 4-5, depending on
     the significance of the issues. Weigh the importance of missing or extraneous
    content against the clarity and adherence to instructions.

4. If the summary partially adheres to the instructions, capturing some of the
    requested details but introducing inconsistencies, hallucinations, or
    irrelevant content, assign a score of 2-4, depending on the extent of the
    deviations and errors. Penalize for any explicitly prohibited content that has
    been included.

5. If the summary minimally adheres to the instructions, misses most of the
    required details, includes significant irrelevant or hallucinated content, or
    ignores the specified structure or tone, assign a score of 1-3, depending on
    the severity of the shortcomings.

6. If the summary fails to follow the user's instructions altogether, missing all
    critical requirements or containing a high proportion of irrelevant or
    fabricated content, assign a score of 1. This includes summaries that fail to
    meet any formatting, detail, or exclusion criteria.

```

````
Here is the input document, user instruction and the corresponding summary.
Source:
```
{src}
```
User Instruction:
```
{instruction}
```
Summary
```
{tgt}
```
Note: It is helpful to read the summary first, before reading the source document.
    This will allow you to judge whether you understand the main contents of the
    source document through the summary alone. Afterward, you can assess to what
    extent the summary accurately reflects the source document.

Note: Based on the above criteria and assign a overall score of summary in the
    scale 1-7. If the summary is not provided for evaluation, return "N/A". Besides
    the score, you should also provide a **brief** explanation.

Note: Use the following json format for easy downstream consumption.

{{
    "explanation": "judge the summary based on the given criteria and explain your
        reasoning for the score you are going to give in the next field.",
    "score": THE_SCORE_VALUE
}}
````

B Authors (alphabetical)

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
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
Xiren Zhou