

Train multi-modal LLM to understand diverse speech paralinguistics by distilling from teacher with meta-information prompt

Jeremy H. M. Wong, Muhammad Huzaifah, Hardik B. Sailor, Shuo Sun, Kye Min Tan, Bin Wang, Qiongqiong Wang, Wenyu Zhang, Xunlong Zou, Nancy F. Chen, and Ai Ti Aw

Institute for Infocomm Research (I²R), A*STAR, Singapore
jeremy_wong@a-star.edu.sg

Abstract

A Large Language Model (LLM) can be extended for input audio, by expressing the audio embeddings as an interpretable prompt to the LLM. The adaptor that computes these audio embeddings is often trained using multi-modal Instruction Fine-Tuning (IFT) data. It is labour intensive to scale up the creation of such data to many datasets, tasks, and audio information types. The labour can be reduced with Knowledge Distillation (KD), by prompting an external teacher LLM with meta-information from the audio dataset, and using the teacher’s output as a reference to train a student that is now prompted with the audio. Prior KD work has only used a few datasets, and does not present experiment comparisons against fair choices of IFT models. This paper scales up KD to a larger collection of datasets, comprising a wider variety of meta-information types. Fair experiment comparisons on Dynamic-SUPERB and AudioBench show that KD and IFT are complementary, but KD alone may not outperform IFT.

Introduction

Large Language Models (LLM) exhibit good generalisation across text tasks that they were not explicitly trained to do, through zero-shot prompting (Radford et al. 2019). In-context learning (Brown et al. 2020) and retrieval-augmented generation (Lewis et al. 2020) can further adapt the LLM to the task, without computationally expensive fine-tuning. A prerequisite for these behaviours to manifest is arguably that the LLM should have a sufficient degree of understanding of natural language text.

Extending the capabilities of LLMs to other modalities may benefit a wider variety of applications. Multi-modal LLM works text in (Ao et al. 2022; Tang et al. 2024; Hu et al. 2024; Chu et al. 2023; Pan et al. 2024; Wang et al. 2024; Kang et al. 2024; Lu et al. 2024, 2025; He et al. 2025) use audio as an input modality. It seems reasonable to assume that behaviours such as good generalisation with zero-shot prompting, in-context learning, and retrieval-augmented generation may analogously only manifest if the multi-modal LLM has a sufficient degree of understanding of information expressed within the audio. Furthermore, a diverse understanding of different types of audio and speech information may facilitate the LLM to generate responses

that are more contextualised toward the user’s situation, and thereby improve user experience. Audio and speech have many forms of information that may not often be expressed in text data. These include speaker characteristics like accent, age, gender, fluency, and emotion. It also includes background noise, overlapping speakers, sounds from non-speech events, music, reverberation, and distortions due to the recording setup.

For an LLM to acquire this understanding, it needs to be ensured that the embeddings extracted from the audio should be aligned with embeddings of the text prompts in the way that information is expressed. With an aligned expression, the audio embeddings can then be used together with the text prompt embeddings, as inputs to the LLM. This interpretation alignment is often facilitated by an adaptor neural network, which is used to transform embeddings extracted from an audio encoder into audio prompt embeddings, which are more similar to the text prompt embeddings. The adaptor can be fine-tuned on multi-modal Instruction Fine-Tuning (IFT) data. For the audio and speech modality, each data point comprises a prompt text instruction describing the task, an audio input, and a reference text output response. Many audio and speech datasets exist, covering a wide diversity of tasks. However, they are often not formatted in the prompt-response format required for IFT, and thus need to be converted to this format. Two tasks with abundantly available data and where format conversion is simple are Automatic Speech Recognition (ASR) and speech translation. The reference output is simply copied as is from the dataset to the response, and only the prompt instruction needs to be designed. Both the ASR and speech translation tasks benefit from well expressed lexical information of the words that are said, but not so much from paralinguistic and non-speech information. As such, the audio prompt embeddings computed from an adaptor fine-tuned on these tasks can be expected to express primarily the lexical content of the speech well.

The generalisation of the information expressed in the audio prompt embeddings can be improved by fine-tuning over speech and audio tasks that require utilisation of more diverse information. Audio and speech datasets are often annotated with a variety of meta-information types, such as the transcript, speaker gender, emotion, noise condition, and caption. To use such data for IFT, the annotated meta-information needs to be converted to prompt instructions and

reference output responses in natural language. This conversion can be done manually, as in (Tang et al. 2024; He et al. 2025), but is labour intensive. Each dataset alone may already be annotated with multiple meta-information types, and the labour cost is exacerbated when aiming to use all meta-information types in each dataset. Different prompt and response templates need to be devised for each different type of meta-information. The labour cost limits the diversity of meta-information types in the IFT data, which then limits the types of information that the multi-modal LLM learns to understand. It is possible to increase the diversity of IFT data, by using an LLM to generate multiple paraphrasings of an initial set of prompt and response templates. However, doing so still requires the labour of verifying the generated paraphrasings for each meta-information type.

A less labour intensive approach to create fine-tuning data is to use an external LLM as a teacher to generate the reference text response. The student multi-modal LLM is then fine-tuned toward this output reference through Knowledge Distillation (KD) (Bucilă, Caruana, and Niculescu-Mizil 2006; Li et al. 2014; Hinton, Vinyals, and Dean 2014). The teacher LLM is prompted with the dataset’s meta-information expressed in text format. IFT data may comprise prompts that instruct many different types of tasks. Rather than accommodating a wide diversity of possible prompt-response tasks like IFT, KD instead often focuses on a single type of task. This may limit the multi-modal LLM student’s ability to follow diverse instructions, but it simplifies the fine-tuning data creation for diverse audio and speech information types. It is possible to use both KD and IFT, either at different fine-tuning stages or interleaved within the same stage, to yield diverse information understanding and diverse instruction following.

Prior works in (Chu et al. 2023; Pan et al. 2024; Wang et al. 2024; Kang et al. 2024; Lu et al. 2024, 2025) have investigated such KD on only a few isolated speech and audio meta-information types. Work in (Chu et al. 2023) expresses the KD data as a question answering task, by instructing an external LLM to generate text question and answer pairs from text format meta-information, while (Pan et al. 2024) prompts the teacher with only the transcript. Focusing on emotion recognition, (Wang et al. 2024) expresses the KD data as a sentence continuation task, by instructing the teacher LLM to continue the sentence when prompted with the transcript and emotion in natural language. Sentence continuation is also used in (Kang et al. 2024), where the teacher’s prompt is the transcript and an emotion flag. Work in (Lu et al. 2024) expresses the KD data as a captioning task, by instructing the LLM teacher to generate a description when prompted with the transcript, emotion, and speaker gender, expressed in natural language. Work in (Lu et al. 2025) extends the KD approach further, by considering 12 meta-information types from across 6 speech datasets. In sentence continuation, it is difficult to ensure that the teacher LLM expresses a wide coverage of the provided prompt information when generating the response. In a captioning task, the teacher can be explicitly instructed to consider all types of meta-information in the prompt when generating its response. Expressing the KD data as a captioning task

therefore makes it easier to ensure wide coverage of multiple meta-information types, thereby easing the inclusion of diverse meta-information types.

This paper expands the KD captioning framework to an even larger scale, by sourcing a broader diversity of paralinguistic and non-speech audio information across many datasets. This aims to empower the multi-modal LLM to understand an even wider diversity of information, while avoiding the expensive labour cost of the creation of IFT data. Expanding further than (Lu et al. 2025), this paper considers 29 types of meta-information across 13 open-source datasets, with not just speech data like in (Lu et al. 2025), but also non-speech audio data. The experiments in (Lu et al. 2025) compare various KD and IFT approaches, but each model in the comparison uses different fine-tuning datasets, model architectures, and text LLM models. This is understandably so, because of the difficulty of preparing the data and doing large-scale training for each separate fine-tuning approach. It is thus difficult to ascertain whether the observed improvements are due to the KD approach, the larger collection of fine-tuning datasets, or the different model architectures. In this paper, KD is compared to IFT across the same diverse collection of fine-tuning datasets and common model architectures, for a fairer experiment comparison. For each dataset, multiple forms of information are extracted from the dataset’s meta-information. All extracted forms of meta-information are used together and expressed in text format when prompting the teacher LLM to generate the reference text. When fine-tuning the student multi-modal LLM, the same prompt is modified by replacing the part that expresses the textually represented meta-information with the audio prompt embeddings. Fine-tuning the student toward the teacher’s output thus encourages the audio prompt embeddings to express the diverse meta-information. Through the fairer comparison between KD and IFT in this paper, the experiments show that this form of meta-information KD is complementary to IFT in empowering the multi-modal LLM to perform well across a broad range of speech and audio tasks.

In summary, this paper proposes to:

- Scale up captioning-style KD to more audio and speech information types, across more open-source datasets, and to use all meta-information types present in each dataset, to allow the multi-modal LLM to learn a more generalised understanding of diverse audio and speech information.
- Present a fair experimental comparison between KD and IFT.

Instruction fine-tuning

For the speech-text tasks of ASR and speech translation, training data is fairly abundant. These comprise speech as the input and text as the output. It requires minimal human labour to convert these into a form that is conducive to fine-tune a multi-modal LLM. Often, all that is done is to include an additional input text prompt that instructs the multi-modal LLM to transcribe or translate the input speech. Audio captioning datasets can also be used with minimal hu-

Table 1: Datasets used for IFT, and the tasks for which human-written prompts and responses were created for each dataset

Dataset	Prompt-response task
AIShell (Bu et al. 2017)	ASR
AudioCaps (Kim et al. 2019)	AC, ASQA
Common Voice (Ardila et al. 2020)	ASR, SQA
GigaSpeech (Chen et al. 2021)	ASR, SQA
IEMOCAP (Busso et al. 2008)	ER
Librispeech (Panayotov et al. 2015)	ASR, SQA
MELD (Poria et al. 2019)	ER, SR
NSC (Koh et al. 2019) parts 1 & 2	ASR
NSC parts 3, 4, & 5	ASR, SQA, DS, AR, GR
NSC part 6	ASR, SQA, DS
People’s Speech (Galvez et al. 2021)	ASR, SQA
SLUE phase-2 (Shon et al. 2023)	SQA
Spoken SQuAD (Li et al. 2018)	SQA
Voxceleb1 (Nagrani et al. 2020)	GR, NR
WavCaps (Mei et al. 2024)	AC, ASQA

man labour. These comprise audio as input, which may or may not include speech, and a text description of the audio as the output. To use these to fine-tune a multi-modal LLM, an additional text prompt needs to be included that instructs the model to describe the audio.

Beyond these, there is a wide diversity of many other audio and speech tasks. Many of these tasks rely on information that is extrinsic to the words spoken. Fine-tuning a multi-modal LLM on these tasks may empower the model to understand the information in the audio that is needed to accomplish each of these tasks. These include tasks for recognising emotion, sentiment, intent, sarcasm, disfluency, language, and physiological conditions. Regression tasks include measuring distance to the microphone and speech evaluation (pronunciation accuracy, fluency, prosody, intonation). Speaker-related tasks include identification, verification, and diarisation, recognition of accent, gender, and age, and counting the number of speakers. Non-speech tasks also include sound event detection, and descriptive captioning of music and audio events. Many of these tasks can be expressed within a question answering framework. However, expressing these datasets in such a framework is labour intensive, as contextually appropriate questions and answers need to be written for each data sample. The labour cost thus limits the diversity of tasks and datasets that can be incorporated.

The IFT data used in this paper, listed in Table 1, is a subset of the data¹ used in (He et al. 2025). The following prompt-response tasks were manually created.

- **Audio Captioning (AC):** The prompt is human-written, instructing the model to describe what is heard in the audio. The output is the caption from the dataset.

¹This information is from private communication with the authors of (He et al. 2025).

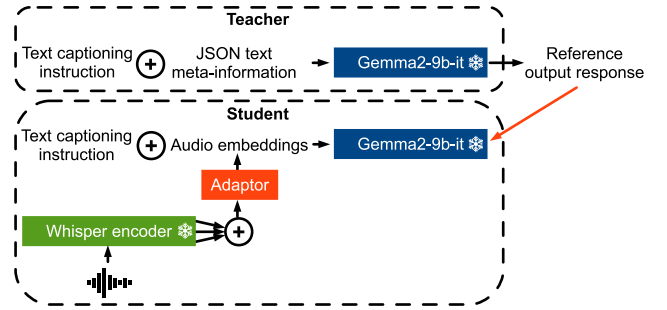


Figure 1: Teacher LLM is prompted with text format meta-information about the audio, to generate a reference output. Student is then prompted with audio prompt embeddings and is fine-tuned toward the teacher’s reference output

- **Automatic Speech recognition (ASR):** The prompt is human-written, instructing the model to transcribe the speech. The output is the transcript from the dataset, with a consistent text normalisation applied across all datasets.
- **Data Summarisation (DS):** The prompt is human-written, instructing the model to summarise the content that the speaker talks about. The output is a human-written summary.
- **Audio Scene Question Answering (ASQA):** The prompt is human-written, asking the model to identify the sounds heard within the audio. For the output, the human annotator first reads the caption in the dataset, and then writes an appropriate answer to the prompt question.
- **Speech Question Answering (SQA):** The prompt is an open-ended human-written question relating to the content that the speaker is discussing. This may comprise what, who, when, where, why, and how questions. For the output, the human annotator first reads the transcript, and then writes an appropriate answer to the prompt question.
- **x Recognition:** The prompt is a human-written question asking the model to identify x from the speech, where x can be Accent (AR), Emotion (ER), Gender (GR), Nationality (NR), or Sentiment (SR). The output is a human-written answer, based on the dataset meta-information.

The human-annotated prompts and reference responses require significant labour, time, and creativity to produce diversely and accurately. Not all of the meta-information types available from the datasets are used in the IFT data. Expanding to more meta-information types increases the labour cost. It is possible to use an LLM to aid in the diversification of the IFT data. However, significant labour is still required to verify the appropriateness of the generated data for each meta-information type and prompt-response task.

Meta-information knowledge distillation

The labour required to create the fine-tuning dataset can be reduced by using an external LLM teacher to generate the reference text that is used to fine-tune the multi-modal LLM

Table 2: Datasets and meta-information in each dataset used for KD fine-tuning

Dataset	Meta-information types
AIShell	transcript, speaking style, language, gender, nationality, number of speakers, duration
AudioCaps	caption, duration, source
Common Voice	transcript, speaking style, language, gender, age, accent, number of speakers, duration, up votes, down votes
GigaSpeech	transcript, topic, speaking style, language, number of speakers, duration, start time, end time, source
IEMOCAP	transcript, speaking style, language, gender, number of speakers, duration, start time, end time, emotion, valence, dominance, arousal
Librispeech	transcript, speaking style, language, gender, number of speakers, duration, source, noise
MELD	transcript, speaking style, language, number of speakers, duration, start time, end time, emotion, sentiment
NSC parts 1 & 2	transcript, speaking style, language, gender, ethnicity, nationality, number of speakers, duration
NSC part 3	transcript, speaking style, language, gender, age, ethnicity, nationality, education, occupation, relationship between speakers, first language, spoken languages, number of speakers, duration
NSC part 4	transcript, speaking style, language, gender, age, ethnicity, nationality, education, occupation, relationship between speakers, first language, spoken languages, dominant language, number of speakers, duration
NSC part 5	transcript, topic, speaking style, language, gender, age, ethnicity, nationality, education, occupation, relationship between speakers, first language, spoken languages, dominant language, number of speakers, duration, sentiment
NSC part 6	transcript, topic, speaking style, language, nationality, number of speakers, duration
People’s Speech	transcript, language, number of speakers, duration
SLUE phase-2	transcript, speaking style, language, number of speakers, duration
Spoken SQuAD	transcript, speaking style, language, number of speakers, duration
Voxcelebl	gender, nationality, number of speakers, duration, source
WavCaps	caption, duration

student. The labour can be further reduced, by focusing on a single prompt-response task. The task of captioning is used in this paper, because of its flexibility to accommodate diverse meta-information types and the ease of instructing the teacher to use a wide coverage of the prompted information. This is illustrated in Figure 1. The teacher LLM is prompted with an instruction as well as meta-information about the audio expressed in text format. The instruction is to describe the audio. When fine-tuning the multi-modal LLM student, the student’s prompt no longer comprises the text format meta-information, but instead uses the audio prompt embeddings. Any information that is already annotated in the dataset can be used in the meta-information prompt. For example, to generate the teacher’s output for an emotion recognition task with an emotion recognition dataset, the annotated emotion label can be extracted from the dataset and included in the teacher’s prompt. The process of extracting the meta-information from the dataset can be automated. This makes it easy to expand the framework to new meta-information types.

In (Kang et al. 2024), KD is expressed as a sentence continuation task, where the teacher’s prompt comprises the transcript of the speech and the annotated emotion. The teacher LLM then generates a continuation from this prompt. The teacher’s generated output is expected to be contextualised on both the transcript and the emotion. When the multi-modal LLM, with speech as the input, is fine-tuned toward this reference, it is hoped that it will learn to understand information about both the words and the emo-

tional expression of the speech. However, as is observed in (Lu et al. 2025), the teacher LLM may omit mentioning about the prompted meta-information, if not explicitly instructed to consider it. Unlike sentence continuation, it is easier within a descriptive captioning task to explicitly instruct the teacher to consider all prompted information when generating the output response (Wang et al. 2024).

A dataset can contain multiple types of meta-information. For manual creation of IFT data, attempting to cover multiple types of meta-information within a dataset increases the already-expensive human labour cost. In the KD framework, the teacher’s prompt can simply be modified to include multiple types of meta-information, thereby contextualising the generated output response on these meta-information types (Lu et al. 2024). Work in (Lu et al. 2024) does such KD using the diverse meta-information available from three speech datasets. These are IEMOCAP, LibriTTS, and PromptTTS. However, the types of meta-information expressed within these datasets are still limited, and most of the speech is synthesised. The number of speech datasets is further expanded to 6 in (Lu et al. 2025) comprising AccentDB, DailyTalk, IEMOCAP, PromptTTS, VCTK, and VoxCeleb, covering 12 meta-information types. Several of the meta-information types are annotated in a semi-supervised manner using an automatic model, and may thus propagate recognition errors from this model.

This paper expands the investigation of the KD captioning framework to an even larger scale, over a wider diversity of datasets and using more types of meta-information.

This may empower the multi-modal LLM to learn to understand more types of information contained within the audio. The datasets used and the meta-information extracted from each dataset are listed in Table 2. 13 open-source datasets are used, covering 29 meta-information types. The aim is to extract as many types of meta-information from each dataset as possible. *Speaking style* can be read, spontaneous, acted, or synthesised. *Source* refers to where the audio was obtained from, and includes audiobook, podcast, and YouTube. *First language* is the language that the speaker first became fluent in. *Spoken languages* are the languages that the speaker is able to converse in. *Dominant language* is the language that the speaker uses in everyday communication. *Noise* in Librispeech is either clean or noisy, depending on whether the utterance belongs to the *clean* or *other* training set. *Sentiment* is either positive or negative.

Two issues are encountered when generating the teacher’s reference output response. First, the teacher may hallucinate beyond the meta-information presented in the prompt. A student learning from this reference may develop an exacerbated hallucinative behaviour. Second, the teacher’s output may have selective coverage, and may not adequately express all meta-information from the prompt. This yields inefficient utilisation of the datasets. Multiple prompt variants were initially manually inspected for the hallucination and coverage that they induce. Finally, the following prompt format was found to yield a reasonable manually inspected balance between hallucination and coverage:

You will hear an audio segment. In natural language, describe what you hear in as much detail as possible. Do not apologize, and do not refer to the fact that you are a large language model or an AI. If the audio contains speech, then you must create a transcript of the exact words spoken by the speakers in the same language that it is spoken in. Also, give a detailed description of as many characteristics of the speakers, speaking style, environment, emotion, sentiment, topic, accent, nationality, and as many other types of characteristics as possible: {AUDIO}{JSON}{AUDIO}

This prompt is designed to be a captioning task. The meta-information within the prompt is expressed in JSON text format, with a JSON string replacing {JSON} in the prompt. Using JSON empowers the flexibility to easily include different sets of meta-information types between datasets. When fine-tuning the multi-modal LLM student, the same prompt is used, with {AUDIO}{JSON}{AUDIO} now replaced by the audio prompt embeddings. This approach is analogous to (Lu et al. 2025), which also uses a single captioning prompt to generate the teacher outputs for all datasets. Future work may consider improving the coverage of meta-information types in the output by modifying the prompt to more specifically instruct the teacher to address each type in the output.

As an example, for utterance *Ses04F_impro05_F017* from IEMOCAP, the {JSON} prompt string is

{“transcript”: “I’ve had enough. I want to talk to-no, no unacceptable I want to your supervisor. I am done

with this.”, “language”: “English”, “gender”: “female”, “emotion”: “anger”, “number of speakers”: 1, “start time”: 172.71, “end time”: 178.27, “style”: “spontaneous speech”, “duration”: 5.56, “emotion valence”: 4.5, “emotion dominance”: 4.5, “emotion arousal”: 1.5, “emotion attribute range”: [1, 5]}

The output response generated by a Gemma2-9b-it (Gemma Team 2024) teacher LLM is

*A woman speaks in a clipped, frustrated tone. She sounds angry and exasperated.
“I’ve had enough. I want to talk to-no, no unacceptable I want to your supervisor. I am done with this.”
The speech is spontaneous and lacks polish.*

During KD, the student has the opportunity to learn from the meta-information types represented in the teacher’s output. The teacher’s output has coverage over some, but not all, of the prompted meta-information. This limited coverage may in turn limit the meta-information understanding that the student can learn. However, coverage and hallucination tend to be a trade-off in the teacher’s prompt design. The prompt explicitly instructs the teacher to generate the transcript, which the teacher obediently outputs in this example, but not always. The presence of the transcript in the teacher’s output gives the student the opportunity to learn to identify the words spoken.

One aim of the KD approach is for the multi-modal LLM student to learn to compute audio prompt embeddings that express the diverse information types from the audio, in a way that is as interpretable by the LLM as the text format meta-information in the teacher’s prompt. Two forms of potential mismatch are purposely avoided here to empower such learning. First, as recommended in (Lu et al. 2025; Kuan and Lee 2025), the external teacher LLM is chosen to be the same as the student’s seed LLM. This differs from (Chu et al. 2023; Pan et al. 2024; Lu et al. 2024), which instead use different LLMs for the teacher and student. Second, the student and teacher use the same prompt, except for the segment where the meta-information is replaced by the audio prompt embeddings. In (Wang et al. 2024; Lu et al. 2024), different text prompts are used for the teacher and student. In (Kang et al. 2024), the student is prompted with audio only, without text. These choices of matched LLMs and prompts aim to preserve the LLM’s original behaviour from text instruction fine-tuning, thus avoiding catastrophic forgetting of general instruction-following capabilities across tasks outside those represented in the multi-modal fine-tuning data.

In evaluation benchmarks such as (Wang et al. 2025), the test set reference outputs are often manually written. These manual references may express linguistic characteristics that differ from the outputs that LLMs tend to generate, and may be more similar to the reference outputs in manually written IFT data. These linguistic characteristics include aspects such as sentence length, choice of vocabulary, formality, and grammatical correctness. Often, evaluation uses an external LLM to judge the similarity between the hypothesis and reference texts. Judge LLMs may exhibit artifacts in their de-

liberation, which may manifest as placing greater emphasis on linguistic similarity, than on concept correctness. As such, a multi-modal LLM fine-tuned with manually written IFT data may yield a behaviour that is favoured by evaluation benchmarks, compared to a KD-fine-tuned multi-modal LLM. This may be especially so when the reference outputs of the IFT and evaluation datasets are manually written by the same group of people, which is the case between the IFT data in (He et al. 2025) and the evaluation data in (Wang et al. 2025). To overcome this mismatch of linguistic characteristics, KD is used together with IFT, either mixed together within a single fine-tuning stage or in sequential stages. Fine-tuning with both KD and IFT together also encourages the multi-modal LLM to learn more diverse instruction following from IFT, beyond the captioning task of KD.

This paper uses a single prompt format for KD. This simplifies the verification of the teacher’s hallucination and coverage. However, it may limit the ability of the multi-modal LLM to generalise beyond this fixed prompt format. Future work may consider diversifying to a wider variety of prompt formats. Perhaps, the process to verify hallucination and coverage by the teacher may be automated to reduce human labour cost, by using external LLMs (Xie et al. 2025). Related work in (Kuan and Lee 2025) also aims to reduce hallucination in the multi-modal LLM student, by prompting the teacher to generate a response with both positive and negative contrastive outputs.

Experiments

Experiments used a multi-modal LLM architecture and IFT inspired by (He et al. 2025). The LLM was Gemma2-9b-it. Whisper-large-v3 (Radford et al. 2023) encoder was used as the audio encoder, to compute an embedding sequence from input audio. A learned layer-weighted-sum combined the embeddings across the encoder transformer layers (Yang et al. 2021). An adaptor neural network converted the audio embedding sequence into a sequence of audio prompt embeddings, which is more interpretable by the LLM. This comprised a feed-forward neural network with a single hidden layer of SiLU activations (Elfwing, Uchibe, and Doya 2018). The input of the adaptor downsampled the audio embedding sequence to have a sequence length that was more similar to that of a text token sequence. This was done using a sliding window with an equal shift and duration of 15 frames, which yields an audio prompt embedding sequence length of 100 for 30 seconds of input audio. The linear output layer converted the embedding dimension to match the prompt token embeddings. The resultant audio prompt embedding sequence was concatenated with the text prompt token embeddings as input to the LLM.

Only the parameters of the adaptor and layer-weighted-sum were updated. This is because the paper aims to investigate alignment between audio and text representations, which is primarily the job of the adaptor. The encoder was not updated, to avoid catastrophic forgetting of the ability to extract and express diverse audio information. The LLM was not updated, to avoid catastrophic forgetting of instruction-following behaviour. The setup in (Lu et al. 2025) also only

updates the adaptor. Unlike in (Yang et al. 2021), the layer-weighted-sum here is task-independent. The multi-modal LLM was fine-tuned using cross-entropy next-token prediction, over 16 GPUs with a per-GPU mini-batch size of 5 samples, using AdamW (Loshchilov and Hutter 2019) with a fixed learning rate of 5×10^{-5} . With the available computing resources, the models were able to be fine-tuned up to 240K steps, which is about a fifth of an epoch of the entire shuffled fine-tuning data collection. When more computing resources become available in the future, it may be desirable to fine-tune for longer. However, the models being compared here were all fine-tuned over the same collection of datasets, for the same number of steps, using the same model architecture, and therefore the experiment is a fair comparison. This differs from (Lu et al. 2025), which instead compares between KD and IFT models that are each fine-tuned over different datasets for differing numbers of steps, using different model architectures and text LLMs.

The models were fine-tuned on the datasets in Tables 1 and 2, which are a subset of a larger collection that is used for IFT in (He et al. 2025). The same human-annotated prompts and reference responses were taken from (He et al. 2025). The IFT prompts and responses were written to cover the tasks shown in Table 1. This only referred to a limited selection of the meta-information types available in the datasets during the annotation process, to limit the labour cost. For KD, the teacher was chosen to be the same Gemma2-9b-it LLM as used in the multi-modal LLM. It generated reference outputs using the prompt format described in the preceding section. Greedy decoding was used to generate the teacher’s output response. Future work may consider increasing the data diversity by using sampling generation, but will need to assess the worsened coverage or hallucination that may result.

The models were evaluated across the diverse speech and audio tasks in the Dynamic-SUPERB (Huang et al. 2024) and AudioBench (Wang et al. 2025) benchmarks. Dynamic-SUPERB focuses on classification tasks, with a fixed set of possible answers, while AudioBench considers ASR and tasks with open-ended responses. Each benchmark comprises datasets across multiple tasks, taxonomised into task categories. The Dynamic-SUPERB categories are audio, content, degradation, paralinguistic, semantic, and speaker tasks. The AudioBench categories are ASR, AC, AR, ASQA, GR, Speech Instruction (SI), and SQA tasks. In an SI task, the audio is of a spoken instruction command, and the model is expected to generate an output that obeys that spoken instruction. In AudioBench, SI can be viewed here as a category of unseen tasks, as the IFT data used here does not include SI. The evaluation metric for Dynamic-SUPERB is the classification accuracy. For AudioBench, ASR is evaluated using Word Error Rate (WER), while the other categories are evaluated by using an external LLM to compute a score between the hypothesis, reference, and prompt, between a worst of 0 and best of 100, as described in (Wang et al. 2025). For both benchmarks, an average was computed with equal weight across all task categories. For AudioBench, the WER for ASR tasks was first subtracted from 100%, and then averaged with the judge scores from

the other task categories. Greedy decoding was used for Dynamic-SUPERB, while the default AudioBench setting of sampling generation with a temperature of 1.0 was used. In Dynamic-SUPERB, the computation of the classification accuracy requires one of the fixed set of output classes to be inferred from the model’s output. In this paper, an external Gemma2-9b-it LLM was used to summarise the long-form hypothesis into one of the classes. The summariser LLM was prompted with

You are a helpful assistant. A speech-text model was previously given the following question about a speech utterance and then generated the following answer.

[
Question: {QUESTION},
Answer: {ANSWER}
]

Now, your job is to summarize this answer into one word of either {CHOICES} that best represents this answer, while considering this question about the speech utterance that was given to the model. Your summary must be in the same format as {CHOICES}. Do not include punctuation.

The prompt from the Dynamic-SUPERB data replaces {QUESTION}. The multi-modal LLM’s generated long-form output response replaces {ANSWER}. The set of possible classes is expressed in natural language and replaces {CHOICES}.

The aim of the experiment is to compare KD and IFT. Four fine-tuning schemes were compared:

- IFT for 240K steps.
- KD for 240K steps.
- Shuffle the IFT and KD samples together, and fine-tune with these for 240K steps. This assesses the complementarity between IFT and KD when used together.
- Start from the 120K step checkpoint of KD, then fine-tune for a further 120K steps with IFT. The aim is to first let the adaptor learn from KD about how to express diverse audio and speech information interpretably by the LLM. Then, IFT further encourages instruction-following and outputs that follow the language structure that human annotations tend to use, to better match the reference outputs in the evaluation benchmarks.

The results are shown in Table 3 for Dynamic-SUPERB and Table 4 for AudioBench. For Dynamic-SUPERB in Table 3, the best average performance is obtained when first using KD for 120K steps, then using IFT for another 120K steps, yielding a final average accuracy of 46.8%. This outperforms using either solely IFT or KD for 240K steps, with accuracies of 46.2% and 41.1% respectively. This suggests that KD and IFT may be complementary. When using KD alone, the average performance does not surpass that of using IFT alone. A closer inspection of the outputs reveals that the KD model more often than IFT complains that the task cannot be accomplished because of a lack of audio in the input. A possible explanation could be that since KD uses a fixed prompt format and single captioning task, the adaptor may not learn to represent the audio information in a manner interpretable by the LLM when used in conjunction with

other text prompts and prompt-response tasks. On the other hand, IFT uses a diversity of human-written prompts and prompt-response tasks, thereby empowering better generalisation. The complaints about missing input audio are particularly prevalent for tasks in the audio category, which are sparsely represented within the fine-tuning datasets. While these experiments used the standard evaluation prompts supplied with the Dynamic-SUPERB and AudioBench benchmarks, future work may consider mitigating this issue by modifying the prompts to explicitly instruct the model not to generate such complaints. KD yields particularly strong improvements across all of the tasks in the semantics category, of dialogue act classification, dialogue act pairing, intent classification, and sarcasm detection, with an accuracy of 74.2% at step 240K. This is surprising, as these tasks are unseen with respect to the types of meta-information used in KD fine-tuning. It may suggest that a KD fine-tuned model may generalise to a broader variety of unseen meta-information types. A closer investigation into the model outputs for these tasks after KD reveals that, unlike other task categories, the model complains less frequently that the task cannot be achieved and seems somewhat better at following the prompt instruction. The KD model also does particularly well in the content category, with an accuracy of 55.7% at step 240K. These are tasks related to recognising the words in the speech, such as speech command recognition, spoken term detection, speech text matching, language identification, and speech detection. This could be due to the frequent presence of transcript meta-information across most of the datasets, and the emphasis to generate the transcript in the KD prompt. As opposed to this, in IFT, the transcripts were only shown in the reference output for the ASR task. As such, the transcripts from several of the datasets were not used, such as IEMOCAP and MELD for which the human annotations focused on the emotion, and SLUE and Spoken SQuAD for which the human annotations focused on question answering.

For AudioBench in Table 4, the best average performance arises from using a mixture of IFT and KD data, with an average score of 57.2 at step 240K. This again suggests complementarity between IFT and KD, but perhaps in a different way than that suggested by the Dynamic-SUPERB results. Using KD alone also again does not surpass the average performance of IFT alone. As with Dynamic-SUPERB, a closer inspection of the AudioBench model outputs also reveals that with KD, the model tends to complain more often that the task cannot be accomplished because of a lack of audio input. This is especially so for the AR, ASQA, and GR tasks. One might assume that the matching of linguistic style between the model’s hypothesis and the human-annotated reference may be more crucial for AudioBench than for Dynamic-SUPERB, as AudioBench assesses long-form outputs. IFT may better align the model output’s linguistic style with that of the reference outputs, compared to KD. However, a closer inspection of the model’s outputs did not reveal any obvious bias of the scores toward the linguistic style of the IFT model outputs. Instead, for SQA, it is noticed that KD tends to yield longer model hypotheses, which may be favoured by the external LLM judge, yielding high scores of

Table 3: Dynamic-SUPERB comparison between IFT and KD

Training	Step	Classification accuracy (%) $\uparrow$						
		audio	content	degradation	paralinguistic	semantic	speaker	average
IFT	120K	43.3	50.7	43.6	31.4	53.2	51.5	45.6
	240K	42.3	47.2	43.6	32.9	60.2	51.3	46.2
KD	120K	12.1	54.5	38.4	23.1	72.0	44.8	40.8
	240K	7.2	55.7	41.6	23.4	74.2	44.3	41.1
Mix IFT & KD	120K	28.7	49.3	39.8	28.1	55.4	46.8	41.4
	240K	32.6	47.4	43.2	31.1	52.8	44.9	42.0
IFT from KD step 120K	120K	45.0	47.1	42.2	35.0	65.5	45.9	46.8

Table 4: AudioBench comparison between IFT and KD

Training	Step	WER (%) $\downarrow$	Judge score $\uparrow$							average $\uparrow$
		ASR	AC	AR	ASQA	ER	GR	SI	SQA	
IFT	120K	16.9	26.9	51.2	33.5	40.7	61.1	23.8	73.7	49.2
	240K	24.4	26.5	56.2	40.3	34.3	77.4	23.9	75.0	51.2
KD	120K	45.3	13.6	7.4	23.3	41.5	23.6	42.6	82.1	36.1
	240K	47.9	13.9	5.6	20.6	41.8	16.3	21.8	81.3	31.7
Mix IFT & KD	120K	34.7	25.4	53.7	27.9	45.7	55.4	39.6	79.4	49.0
	240K	30.6	24.3	67.7	39.2	42.8	90.8	42.7	80.8	57.2
IFT from KD step 120K	120K	36.5	27.8	51.3	28.9	34.9	83.7	7.3	75.8	46.7

82.1 and 81.3 for steps 120K and 240K respectively. For ASR, the poorer WERs of 45.3% and 47.9% for the KD model are primarily dominated by poor performance in the Earnings test sets, which have long-duration utterances. In the Earnings test sets, the KD model tends to output more formatting in the hypothesis, such as for URLs, numbers, and named entities. No text normalisation was performed in the KD teacher’s prompt meta-information. As such, the teacher may output the formatted transcript from the prompt. On the other hand, IFT does normalise the reference transcripts for the ASR task, thus suppressing the generation of formatted text in the model’s output. This matches the normalised evaluation reference. Surprisingly, even though the KD model was not fine-tuned explicitly for ASR, it is still able to consistently output only the transcription, without any descriptive captioning, when instructed to perform the ASR task. The SI category is unseen with respect to both IFT and KD. For IFT, SI is the category with the poorest judge scores of 23.8 and 23.9, suggesting a difficulty to generalise to unseen tasks. For KD, the 120K step checkpoint initially performs well on SI with a score of 42.6, but the performance is inconsistent, as it later degrades at the 240K step to a score of 21.8. Mixing IFT and KD seems to yield more consistent generalisation to the unseen SI task, with scores of 39.6 and 42.7 at steps 120K and 240K respectively.

Taken together, the results suggest that KD and IFT are complementary fine-tuning approaches. Therefore, when used together, the adaptor better learns to express diverse information from the audio in a way that is interpretable by the LLM. However, using KD alone may not be optimal. Thus, KD data can be added to existing IFT data, to increase the quantity and diversity of the data with minimal additional labour cost. The student’s frequent occurrence of

complaining that audio input is lacking may potentially be reduced by diversifying the teacher’s prompt format to improve generalisation. Future work may consider how to diversify this prompt while still ensuring wide coverage of meta-information and low hallucination in the teacher’s output response. The previous investigation in (Lu et al. 2025) did not show a degraded performance of using KD alone compared to IFT, possibly because the comparisons used different datasets, model architectures, and text LLMs for each fine-tuning approach. A fairer comparison in this paper yields more comparable trends.

Conclusion

This paper extends KD fine-tuning of a multi-modal speech-text LLM to a wider diversity of meta-information types and datasets. The experiment presents a comparison between IFT and KD in a fairer setup than prior works. The results show that IFT and KD are complementary in their ability to fine-tune the adaptor, to compute embeddings from audio that express diverse information types in a way that is interpretable by the LLM. However, KD alone may not surpass the IFT performance on average.

Acknowledgements

The computational work for this paper was performed on resources of the National Supercomputing Centre, Singapore (<https://www.nsc.sg>).

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the National Research Foundation, Singapore.

References

- Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; Wei, Z.; Qian, Y.; Li, J.; and Wei, F. 2022. SpeechT5: unified-modal encoder-decoder pre-training for spoken language processing. In *ACL*, 5723–5738. Dublin, Ireland.
- Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2020. Common Voice: a massively-multilingual speech corpus. In *LREC*, 4218–4222. Marseille, France.
- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *NeurIPS*. Vancouver, Canada.
- Bu, H.; Du, J.; Na, X.; Wu, B.; and Zheng, H. 2017. AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline. In *O-COCOSDA*, 58–62. Seoul, South Korea.
- Bucilă, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model Compression. In *KDD*, 535–541. Philadelphia, USA.
- Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335–359.
- Chen, G.; Chai, S.; Wang, G.; Du, J.; Zhang, W.-Q.; Weng, C.; Su, D.; Povey, D.; Trmal, J.; Zhang, J.; Jin, M.; Khudanpur, S.; Watanabe, S.; Zhao, S.; Zou, W.; Li, X.; Yao, X.; Wang, Y.; You, Z.; and Yan, Z. 2021. GigaSpeech: an evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech*, 3670–3674. Brno, Czechia.
- Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: advancing universal audio understanding via unified large-scale audio-language models. Technical report, Alibaba Group.
- Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107: 3–11.
- Galvez, D.; Diamos, G.; Ciro, J.; Cerón, J. F.; Achorn, K.; Gopi, A.; Kanter, D.; Lam, M.; Mazumder, M.; and Reddi, V. J. 2021. The People’s Speech: a large-scale diverse English speech recognition dataset for commercial usage. In *NeurIPS*. Sydney, Australia.
- Gemma Team. 2024. Gemma 2: improving open language models at a practical size. Technical report, Google DeepMind.
- He, Y.; Liu, Z.; Sun, S.; Wang, B.; Zhang, W.; Zou, X.; Chen, N. F.; and Aw, A. T. 2025. MERaLiON-AudioLLM: bridging audio and language with large language models. Technical report, Institute for Infocomm Research.
- Hinton, G. E.; Vinyals, O.; and Dean, J. 2014. Distilling the Knowledge in a Neural Network. In *Deep Learning and Representation Learning Workshop, NIPS*. Montréal, Canada.
- Hu, S.; Zhou, L.; Liu, S.; Chen, S.; Meng, L.; Hao, H.; Pan, J.; Liu, X.; Li, J.; Sivasankaran, S.; Liu, L.; and Wei, F. 2024. WavLLM: towards robust and adaptive speech large language model. In *EMNLP*, 4552–4572. Miami, USA.
- Huang, C.-Y.; Lu, K.-H.; Wang, S.-H.; Hsiao, C.-Y.; Kuan, C.-Y.; Wu, H.; Arora, S.; Chang, K.-W.; Shi, J.; Peng, Y.; Sharma, R.; Watanabe, S.; Ramakrishnan, B.; Shehata, S.; and Lee, H.-Y. 2024. Dynamic-SUPERB: towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP*, 12136–12140. Seoul, South Korea.
- Kang, W.; Jia, J.; Wu, C.; Zhou, W.; Lakomkin, E.; Gaur, Y.; Sari, L.; Kim, S.; Li, K.; Mahadeokar, J.; and Kalinli, O. 2024. Frozen large language models can perceive paralinguistic aspects of speech. ArXiv preprint arXiv:2410.01162.
- Kim, C. D.; Kim, D.; Lee, H.; and Kim, G. 2019. AudioCaps: generating captions for audios in the wild. In *NAACL-HLT*, 119–132. Minneapolis, USA.
- Koh, J. X.; Mislan, A.; Khoo, K.; Ang, B.; Ang, W.; Ng, C.; and Tan, Y.-Y. 2019. Building the Singapore English national speech corpus. In *Interspeech*, 321–325. Graz, Austria.
- Kuan, C.-Y.; and Lee, H.-Y. 2025. Teaching audio-aware large language models what does not hear: mitigating hallucinations through synthesized negative samples. In *Interspeech*, 2073–2077. Rotterdam, The Netherlands.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*. Vancouver, Canada.
- Li, C.-H.; Wu, S.-L.; Liu, C.-L.; and Lee, H.-Y. 2018. Spoken SQuAD: a study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech*, 3459–3463. Hyderabad, India.
- Li, J.; Zhao, R.; Huang, J.-T.; and Gong, Y. 2014. Learning Small-Size DNN with Output-Distribution-Based Criteria. In *Interspeech*, 1910–1914. Singapore.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *ICLR*. New Orleans, USA.
- Lu, K.-H.; Chen, Z.; Fu, S.-W.; Huang, H.; Ginsburg, B.; Wang, Y.-C. F.; and Lee, H.-Y. 2024. DeSTA: enhancing speech language models through descriptive speech-text alignment. In *Interspeech*, 4159–4163. Kos, Greece.
- Lu, K.-H.; Chen, Z.; Fu, S.-W.; Yang, C.-H. H.; Balam, J.; Ginsburg, B.; Wang, Y.-C. F.; and Lee, H.-Y. 2025. Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP*. Hyderabad, India.
- Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M. D.; Zou, Y.; and Wang, W. 2024. WavCaps: a ChatGPT-assisted weakly-labelled audio captioning dataset

for audio-language multimodal research. *IEEE/ACM Trans. Audio, Speech, Language Process*, 32: 3339–3354.

Nagrani, A.; Chung, J. S.; Xie, W.; and Zisserman, A. 2020. Voxceleb: large-scale speaker verification in the wild. *Computer Speech and Language*, 60: 1–15.

Pan, J.; Wu, J.; Gaur, Y.; Sivasankaran, S.; Chen, Z.; Liu, S.; and Li, J. 2024. COSMIC: data efficient instruction-tuning for speech in-context learning. In *Interspeech*, 4164–4168. Kos, Greece.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *ICASSP*, 5206–5210. Brisbane, Australia.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: a multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, 527–536. Florence, Italy.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, 28492–28518. Honolulu, USA.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Shon, S.; Arora, S.; Lin, C.-J.; Pasad, A.; Wu, F.; Sharma, R.; Wu, W.-L.; Lee, H.-Y.; Livescu, K.; and Watanabe, S. 2023. SLUE Phase-2: a benchmark suite of diverse spoken language understanding tasks. In *ACL*, 8906–8937. Toronto, Canada.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2024. SALMONN: towards generic hearing abilities for large language models. In *ICLR*. Vienna, Austria.

Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A. T.; and Chen, N. F. 2025. AudioBench: a universal benchmark for audio large language models. In *NAACL-HLT*, 4297–4316. Albuquerque, USA.

Wang, C.; Liao, M.; Huang, Z.; Wu, J.; Zong, C.; and Zhang, J. 2024. BLSP-Emo: towards empathetic large speech-language models. In *EMNLP*, 19186–19199. Miami, USA.

Xie, Z.; Lin, M.; Liu, Z.; Wu, P.; Yan, S.; and Miao, C. 2025. Audio-Reasoner: improving reasoning capability in large audio language models. ArXiv preprint arXiv:2503.02318.

Yang, S.-W.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I. J.; Lakhotia, K.; Lin, Y. Y.; Liu, A. T.; Shi, J.; Chang, X.; Lin, G.-T.; Huang, T.-H.; Tseng, W.-C.; Lee, K.-T.; Liu, D.-R.; Huang, Z.; Dong, S.; Li, S.-W.; Watanabe, S.; Mohamed, A.; and Lee, H.-Y. 2021. SUPERB: speech processing universal performance benchmark. In *Interspeech*, 1194–1198. Brno, Czechia.