Distilling an End-to-End Voice Assistant Without Instruction Training Data

William Held[γ,σ*] **Yanzhe Zhang**[γ,σ] **Minzhi Li**[ν]
Weiyan Shi[ϵ] **Michael J Ryan**[σ] **Diyi Yang**[σ]
[γ]Georgia Institute of Technology, [ϵ]Northeastern University
[ν]National University of Singapore, [σ]Stanford University

Abstract

Voice assistants, such as Siri and Google Assistant, typically model audio and text separately, resulting in lost speech information and increased complexity. Recent efforts to address this with end-to-end Speech Large Language Models (speech-in, text-out) trained with supervised finetuning (SFT) have led to models "forgetting" capabilities from text-only LLMs. Our work proposes an alternative paradigm for training Speech LLMs without instruction data, using the response of a text-only LLM to transcripts as self-supervision. Importantly, this process can be performed without annotated responses. We show that our Distilled Voice Assistant (DiVA) generalizes to Spoken Question Answering, Classification, and Translation. Furthermore, DiVA better matches user preferences, achieving a 72% win rate compared with state-of-the-art models like Qwen 2 Audio, despite using >100x less training compute.

1 Introduction

As Large Language Model (LLMs) capabilities improve, so does the value of bringing these capabilities to new modalities, including audio and speech (Shu et al., 2023; Wang et al., 2023; Gong et al., 2023). Speech is a natural interface for language technology (Murad et al., 2019), offering large communication speedups (Ruan et al., 2018).

One straightforward method of enabling speech inputs to LLMs is to feed audio to an Automatic Speech Recognition (ASR) model and produce a text transcription for the LLM to use. However, these pipelined systems cannot capture paralinguistic information such as tone or pace (Upadhyay et al., 2023) and require supervision for both transcription and response generation to be finetuned.

As such, LLMs that directly process speech have the potential to accelerate inference, reduce annotation costs, and capture the rich information inevitably lost by ASR. In this pursuit, a variety of works have trained audio encoders on top of LLMs (Ao et al., 2021; Chen et al., 2021b; Deshmukh et al., 2023; Chu et al., 2023; Wu et al., 2023), many of which utilize the same well-established approach: large-scale multi-task supervised finetuning (SFT). However, models using SFT face two key challenges both stemming from the limited available speech instruction data.

First, SFT-trained Speech LLMs often fail to generalize capabilities from the text-only LLM to speech. As observed in Tang et al. (2023), freezing the weights of the text-only LLM is insufficient to prevent this "forgetting". For text LLMs, this is solved with instruction data covering different tasks and domains. However, broad annotated speech instruction training data does not currently exist.

Secondly, the limited instruction data that does exist is often collected from a small pool of speakers (Kim et al., 2021; Tomasello et al., 2023) or intended for evaluation rather than training (Faisal et al., 2021; Eisenstein et al., 2023). This lack of representation of speech from the wider population is likely to exacerbate biases in speech processing (Koenecke et al., 2020; Mengesha et al., 2021; Chan et al., 2022; Javed et al., 2023; Brewer et al., 2023). At present, Speech LLMs appear fundamentally limited by existing instruction data.

In this work, we argue that these "limitations" of existing data are artificially imposed by SFT. The speech community has invested in large-scale data collection from the internet (Radford et al., 2023; Chen et al., 2021a; Li et al., 2023b), audiobooks (Panayotov et al., 2015; Pratap et al., 2020), and public archives (Galvez et al., 2021). Furthermore, several datasets have been explicitly gathered to represent diverse demographics (Porgali et al., 2023; Garg et al., 2023). However, these large-scale and diverse datasets are dominated by one task: Automatic Speech Recognition (ASR). Adding ASR data into SFT will weaken non-ASR

*Contact: held@stanford.edu, diyiy@stanford.edu.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7876–7891

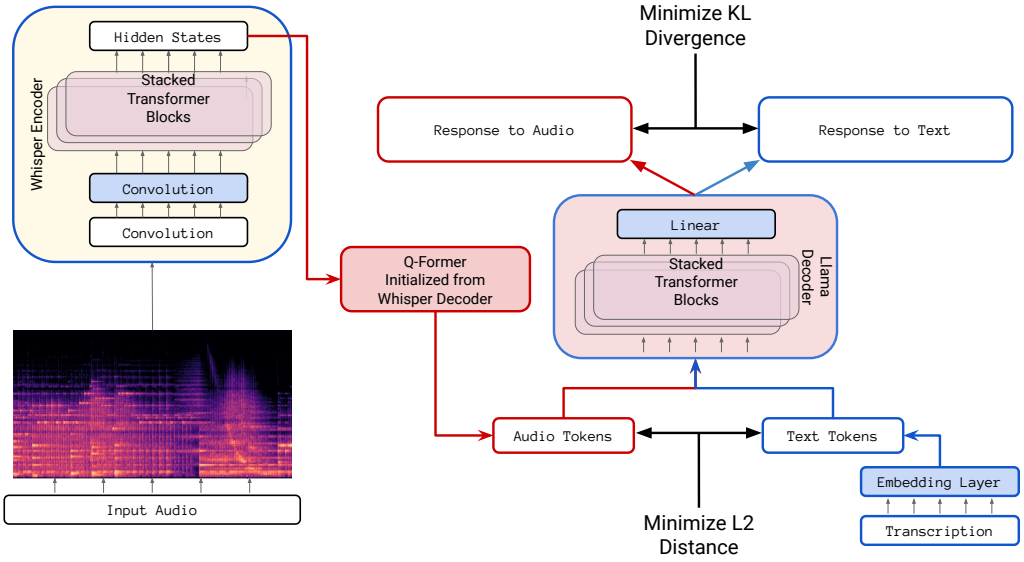

Figure 1: Training pipeline for Distilled Voice Assistant (DiVA). Red indicates trainable components, while Blue indicates frozen pretrained modules. DiVA modifies a text-only LLM into a general-purpose Speech LLM by using the model's own responses to transcribed speech as self-supervision.

capabilities due to imbalanced distribution.

We train a model that generalizes well despite using *only* ASR data[1]. Rather than relying on external labels, our **Di**stilled **V**oice **A**ssistant (**DiVA**) self-supervises learning using the output distribution of an LLM in response to transcripts as a target, a cross-modal form of context distillation (Snell et al., 2022; Mu et al., 2024). We test our approach by training on just a single corpus, CommonVoice, consisting of speech and transcriptions from volunteers around the world (Ardila et al., 2019).

Despite this data simplicity, DiVA generalizes to Spoken Question Answering, Classification, and Translation. Furthermore, DiVA is preferred by users to our most competitive baseline Qwen 2 Audio in 72% of trials despite DiVA using over 100x less training compute. Beyond contributing a new Speech LLM, DiVA creates a new approach to Speech LLMs that improves generalization *without* requiring new speech instruction data.

2 Related Work

LLMs have been extended to both audio and image inputs using cross-modal encoders. For example, LLaVA (Liu et al., 2023b) enables image understanding by connecting CLIP (Radford et al., 2021) to Vicuna (Chiang et al., 2023) through an MLP layer. Several recent works (Zhang et al., 2023; Gong et al., 2023; Tang et al., 2023; Chu et al., 2023, 2024) have connected audio-encoders (Gong et al., 2021; Hsu et al., 2021) to LLMs. There are two critical questions in this space.

Model	Base LLM	Training Method	# Hours
BLSP	Llama 2	Continuation Writing	∼20k
SALMONN	Alpaca 7B	SFT	4.4k
Qwen Audio	Qwen 7B	SFT	∼50k
Qwen 2 Audio	Qwen 7B	SFT, DPO	>370k
UltraVox	Llama 3 8B	Output Distillation	∼10k
DiVA (Ours)	Llama 3 8B	Input & Output Distillation	3.5k

Table 1: High-Level comparison with state-of-the-art open-access Speech & Audio LLMs which we compare to. DiVA offers an new form of context distillation which improves generalization.

How can audio features be transformed into a reasonable number of LLM input embeddings? Audio comes at high sample rates, and therefore, audio encoders often have a large number of outputs. To use these features for LLMs, the dimensionality must be reduced, either by stacking consecutive features (Wu et al., 2023; Fathullah et al., 2024) or learning an adapter-module, such as an MLP (Liu et al., 2023b; Gong et al., 2023), or Q-Former (Dai et al., 2023; Tang et al., 2023).

While learned approaches are more flexible, allowing for an adaptive reduction, they generally require learning a cross-attention mechanism, which generally requires significant training (Li et al., 2023a). In this work, we find the best of both worlds by leveraging the Whisper decoder (Radford et al., 2023) to initialize the text-audio cross-attention mechanism of a Q-Former.

[1]We open-source our models, code, and data in A.2

How can Speech LLMs be trained to achieve instruction following abilities using existing data? Prior work has explored two main routes for creating instruction data without major financial investment. The first approach transforms existing datasets into instruction-following formats (Dai et al., 2023; Chu et al., 2023; Tang et al., 2023; Liu et al., 2023a). While this leverages available resources, it often faces limitations from dataset-task misalignment and imbalanced coverage across different capabilities. The second approach generates synthetic responses by having commercial models process text representations of new modalities (Liu et al., 2023b; Gong et al., 2023; Wang et al., 2024).

Within the synthetic generation paradigm, several approaches use conceptually similar ideas of utilizing supervision from text. Some methods employ hard distillation, where discrete outputs are sampled from teacher models and used as training targets (Wang et al., 2024). This approach treats the teacher's outputs as ground truth labels, optimizing cross-entropy loss on these discrete tokens. In contrast, soft distillation methods attempt to match the continuous probability distributions of the teacher model, which has shown to be more efficient due to rich supervision across the entire distribution from the teacher model Hinton et al. (2015a).

Among soft distillation approaches, implementations vary significantly in both methodology and computational efficiency. UltraVox[2] and other recent models employ direct KL divergence distillation as a natural strategy, though this can be computationally expensive as shown in Snell et al. (2022). Computing full KL divergence over the vocabulary requires $O(V \cdot d)$ operations, where V represents vocabulary size and d the hidden dimension—a significant cost given typical vocabulary sizes of most LLMs.

A key limitation of existing distillation methods is their exclusive focus on output distributions. However, the input representations from text-backbone models also contain valuable supervisory signals that could enhance knowledge transfer. By incorporating distillation losses from both input and output distributions, more comprehensive alignment between student and teacher models becomes possible. Additionally, we introduce an alternative formulation of the KL divergence objective which reduces computational complexity from $O(V \cdot d)$ to $O(d)$.

How can we train foundation models for speech using open and permissively licensed data? Recently, frontier LLMs have begun integrating native speech capabilities. Unlike prior speech foundation models (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Kim et al., 2021; Peng et al., 2023, 2024), these models offer instruction following capabilities rather than self-supervised audio representations or transcriptions. It is unclear to what degree these results are dependent on internal datasets, especially since even the state-of-the-art *open-access* Speech LLM with such capabilities do not report data details other than size (Chu et al., 2024).

Similar to the Open Whisper-style Speech Model (OWSM) initiative (Peng et al., 2024), we use only open and permissively-licensed data. Furthermore, unlike the baselines we compare to in Table 1, we release the training code, rather than just the inference code, which can help reproduce a DiVA-style model easily. In addition to our novel method, we believe this broadens the ability to train and understand Speech LLMs.

3 Method

DiVA is an end-to-end voice and text assistant, trained using the process shown in Figure 1. We focus heavily on effectively using pretrained models in each domain (Section 3.1). Similar to prior works, we initialize the audio encoder from the 1.5B parameter Whisper-Large-v3 model. Unlike previous works, we further reuse the Whisper decoder as the initialization of the Q-Former between the audio encoder and the text-only LLM. We train our model using distillation loss on the input and output distribution of the LLM, which we discuss in Section 3.2.

3.1 Model Initialization

When adding multimodal capabilities to an existing language model, the new modality must be represented as embeddings that can used in place of text token embeddings. Achieving this goal has two steps. First, meaningful features must be extracted from the input modality. Second, these features must be aggregated to be in-distribution for the downstream language model.

Audio Feature Extraction We follow prior works (Chu et al., 2023; Tang et al., 2023) and use

[2]While no paper or technical report exists for UltraVox, we analyze the available training source code from Fixie AI.

the Whisper encoder (Radford et al., 2023). Whisper first transforms the raw audio signal into a 128-channel time-frequency domain Mel-spectrogram. This is then passed through two 1D convolutions and used as embeddings fed to an unmodified Transformer architecture (Vaswani et al., 2023).

All baselines, and indeed almost all recent Speech LLMs, in Table 1 use Whisper in some way. While we follow this choice for consistency with prior work, this means limitations in Whisper itself are unlikely to be measured by our experiments. While Whisper provides strong ASR capabilities, it was primarily trained for transcription tasks and may have limitations in capturing paralinguistic information such as tone, emotion, and prosody.

Audio-Text Feature Alignment While the Whisper encoder extracts meaningful audio features, they are encoded at high granularity, with one token for every 20 milliseconds of input audio. By comparison, humans speak on average around one syllable every 150 milliseconds across languages (Coupé et al., 2019), and most tokens in an LLM vocabulary are made up of several syllables. This creates a mismatch between the granularity between the Whisper encoder outputs and the downstream LLMs input distribution.

Prior work (Tang et al., 2023) addresses this using a Querying Transformer (Q-Former, Li et al. 2023a), which learns static query embeddings with cross-attention to keys and values features from another modality. Given audio embeddings $\boldsymbol{A}$, the Q-Former learns a transformer with a cross attention mechanism $\sigma(\frac{\boldsymbol{Q}(\boldsymbol{K}\boldsymbol{A}^\intercal)}{\sqrt{d_k}})(\boldsymbol{V}\boldsymbol{A})$ where $\boldsymbol{Q}$ is a static set of query vectors, while $\boldsymbol{K}$ and $\boldsymbol{V}$ are projection matrices for the audio tokens. Conceptually, this cross-attention mechanism learns to dynamically aggregate information from the audio tokens into text-like tokens. This comes at the cost of significant training required to train the transformer from scratch.

The Whisper decoder, which prior work discards, is trained with a similar goal: mapping audio embeddings to discrete text tokens for ASR. Therefore, rather than learning Q-Former parameters from scratch, we initialize K and V from Whisper's cross-attention mechanism. We adapt the model to a Q-Former by replacing the inputs with static query tokens Q. Finally, we project the output from the hidden dimension h of Whisper to the hidden dimension H of the LLM. This results in $\{\boldsymbol{t}_q^{audio} \in \mathbb{R}^{H\times|Q|}\}$ tokens representing the audio.

Text Decoding For language processing and instruction following capabilities, we use the original Llama 3 8B Instruct model (Dubey et al., 2024)[3] and leave its weight frozen throughout training.

3.2 Distillation Losses

We optimize two loss functions based on audio recordings and corresponding text transcripts from ASR data. First, we minimize the distance between embeddings of audio and text on the *input* side of the LLM, similar to Radford et al. (2021); Li et al. (2023a). Then, we minimize the KL Divergence between the *output* distribution in response to audio and text as a form of cross-modal context distillation (Mu et al., 2024; Snell et al., 2022).

3.2.1 Input Token Alignment

To capture the mutual information between recordings and text transcripts, for a given ASR example (a text transcript and an audio recording), we align speech and text tokens as follows: The text transcript is embedded as N text tokens $\boldsymbol{t}_i^{text} \in \mathbb{R}^{H\times N}$. The model produces $|Q|$ tokens from the recording where $|Q| > N$. We align these representations by minimizing the L_2 distance between the text embeddings and the final i audio embeddings:

$$L_{con} = \sum_{n=0}^{N} |\boldsymbol{t}_n^{text} - \boldsymbol{t}_{Q-N+n}^{audio}|_2 \quad (1)$$

We use the final N tokens of the audio embedding rather than the initial N tokens due to the causal attention in Whisper's decoder. Since the final tokens can attend to all preceding tokens, aligning the representations of the final tokens backpropagates signal to every token in the sequence. On the other hand, the additional $Q - N$ tokens provide information bandwidth for other information, such as sociophonetic cues, to be passed to the LLM.[4]

Empirically, as we explore in Section 6, training with only token alignment leads to poor model quality, even when low loss is achieved. However, token alignment empirically enables reasoning between text and audio tokens, vastly improving text instruction adherence.

3.2.2 Output Embedding Distillation

Voice assistant models should give coherent, helpful, and harmless responses to user speech. Thankfully, many openly accessible text-only LLMs have

[3]Training was performed before the release of Llama 3.1.

[4]We provide empirical validation of this claim through token statistics analysis in Appendix A.3.

been extensively refined for these objectives. As such, our challenge is not to learn these behaviors but instead to transfer them to the audio modality. In theory, input token alignment could achieve this. However, minor differences in input embeddings can largely affect model behavior (Cai et al., 2022).

Distillation loss, on the other hand, directly optimizes for the similarity of the output distribution (Hinton et al., 2015b). Rather than distilling a large model into a smaller model, recent work has applied to distilling useful context into model weights, a process termed context distillation (Snell et al., 2022; Mu et al., 2024). Here, we apply context distillation across modalities, aiming to distill a text context into the audio modality under the assumption that the model should respond similarly to audio and text for most inputs.

In prior context distillation works, the full Kullback–Leibler (KL) Divergence has been shown to be prohibitively expensive at training time due to the large vocabulary of modern LLMs. Therefore, the KL Divergence is instead approximated by sampling random tokens (Snell et al., 2022). In our case, where the output embedding matrix is frozen, we show that there is an objective function easier to optimize:

Lemma 1. *Given the probability P_t from a teacher model and the probability P_s from a student model, the KL Divergence is defined as $KL(P_t, P_s) = P_t \cdot (\log P_t - \log P_s)$. For a transformer language model, $P_s = \sigma(O_s h_s)$ where h_s is the final hidden state, O_s is the output embedding matrix, and σ is the softmax function. Let θ_s be the student weights which we are trying to train to minimize the KL Divergence, then*

$$\arg_{\theta_s} \min ||\boldsymbol{h}_s - \boldsymbol{h}_t||_2 \subset \arg_{\theta_s} \min KL(P_t, P_s)$$

Proof. The KL divergence is minimized when $P_s = P_t$. Based on our definition of LM probability, this is equivalent to achieving $\sigma(O_s h_s) = \sigma(O_t h_t)$. In the special case we consider, where the teacher and student are initialized from the same weights, and O_s is held constant, we know that $O_s = O_t$. Thus, a non-unique global minimum will be achieved when $h_s = h_t$, where the non-uniqueness comes from the softmax function σ, which is not injective. □

More importantly, we find that: (1) The gradient for L2 loss is much smoother than minimizing the KL divergence empirically[5]. (2) Since the vocabulary size of most modern LLMs is far larger than the hidden dimension, the distance between hidden states can be computed using far fewer operations than the KL divergence. In practice, we optimize the similarity of only the first predicted next token (after all I text tokens/all Q audio tokens) for efficiency, as Morris et al. (2023) has shown that just a single token probability encodes significant information, both for prior and future tokens.

Notably, training with this loss only guarantees that the output distribution is well aligned in response to audio. However, our intuition is that this loss alone is likely to be less robust to input distribution shift without our token alignment loss, which we explore in Section 6.

4 Experimental Setup

4.1 Training Data

We utilize the English subsection of CommonVoice 17 (Ardila et al., 2019) as the dataset for all DiVA training runs. The dataset comprises just over 3.5 thousand hours of reading text that has been crowdsourced and validated on the CommonVoice website. We select the CommonVoice for three reasons. Firstly, it is permissively licensed for commercial and research use. Secondly, it contains speech recorded in realistic settings on an individual's device rather than in a professional studio. Finally, it includes speech from 93,725 speakers from a global pool of volunteers[6]. The first factor means that the resulting DiVA models we release can be adapted for use broadly, while the latter two make the training data more representative of real users.

4.2 Training Hyperparameters

We train for 4300 steps and a batch size of 512 using the AdamW Optimizer, a learning rate of $5E^{-5}$, and a weight decay of 0.1. This amounts to roughly two epochs over the data. We linearly warm up the learning rate for the first 1% of steps and then follow a cosine learning rate schedule, which decays the learning rate to 0 throughout the training run. The training run completes in ∼12 hours on a TPU v4-256 pod.[7]

[5] We empirically validate the utility of our approximation in an isolated, small-scale experiment in Appendix A.4

[6] Statistics drawn from the official CommonVoice tracker

[7] All training configurations can be found on Github

5 Quantitative and Qualitative Evaluations

We first assess how DiVA compares to baseline models for various spoken language benchmarks that SFT models target. We evaluate benchmarks for spoken question answering, speech classification, and speech translation. This provides a quantitative validation of DiVA's generalization.

As these benchmarks were all designed to test single-task systems focused on each task, it is unclear whether they capture the capabilities users expect from virtual assistants that Speech LLMs are now powering commercially. To assess this, we compare DiVA with the best-performing model on the benchmark evaluation (Qwen 2 Audio) in a side-by-side user study.

Baselines We compare our results to five openly available Speech Language Models: BLSP (Wang et al., 2024), UltraVox, SALMONN (Tang et al., 2023), Qwen Audio Chat (Chu et al., 2023), and Qwen 2 Audio Instruct (Chu et al., 2024). Both the Qwen models and SALMONN train on SFT mixtures that covers these benchmark tasks. This makes them strong baselines: they all use similar scale base LLMs to DiVA, all use the Whisper encoder, and all have received direct supervision on the evaluated tasks. For our user study, we compare with Qwen 2 Audio, which reports state-of-the-art numbers and achieves the best average performance in our benchmarks.

5.1 Benchmarking

Speech translation is tested on CoVoST 2, translating 15,500 English examples into seven commonly tested typologically diverse languages (Clark et al., 2020). For question answering, we use HeySquad (Wu et al., 2024) and SDQA (Faisal et al., 2021), testing on 4,000 and 494 question-answer pairs, respectively. Classification is broken down into emotion recognition, sarcasm detection, and humor recognition. Emotion recognition is assessed on IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019), with 1,241 and 2,608 utterances. We evaluate sarcasm detection on MUSTARD's 690 clips (Castro et al., 2019) and humor recognition on URFunnyV2's 7,614 examples. These datasets cover a wide range of traditionally benchmarked speech tasks drawn from prior work, which we cover in greater depth in Appendix A.6.

5.1.1 Speech Translation

First, we assess the speech-to-text translation capabilities of each model from English Speech to text in another language.

On this benchmark, which SALMONN and Qwen Audio Chat were trained for, the results are mixed across languages. The original Qwen Audio performs best in Chinese and Japanese, Qwen 2 Audio performs best in Arabic, German, and Indonesian, and DiVA performs best in Tamil and Turkish. Notably, the original Qwen trains with more than 3700 hours of speech-to-text translation data from CoVost2. While Qwen 2 does not report which tasks it trains on, it is likely trained on similar or increased volumes of data from CoVost2 as the original Qwen. This highlights the data efficiency and the transferability of the DiVA approach, as both of these models were trained on more translation-specific data than the DiVA used for its entire training.

The other distillation-based approaches show notably poor translation performance. BLSP, which uses hard distillation, achieves very low BLEU scores across all languages (averaging 5.05), while UltraVox's soft output distillation performs similarly poorly, with particularly low scores on Japanese (0.20) and Tamil (0.17). These results suggest that output distillation alone—whether hard or soft—has shortcomings for translation.

DiVA's most notable underperformance is in Chinese and Japanese, where it underperforms both other models. Inspecting DiVA's outputs and comparing them to translations from Llama 3 in response to text, we again find that our distillation loss leads us to preserve a negative behavior — for both Chinese and Japanese, Llama 3 has a strong bias towards generating translations in the Latin alphabet (Pinyin and Romaji) rather than the expected native script. This leads to especially poor results in these languages. Notably, this shortcoming also impacts performance for UltraVox which is derived from Llama 3 as well.

5.1.2 Spoken Question Answering

We evaluate all models on zero-shot spoken question answering by prompting them with recorded audio of a speaker asking a question. The underlying LLMs for all baseline models are capable of question-answering, meaning that the audio encoder only needs to learn to map audio to the correct corresponding text to achieve strong results. This is where we expect DiVA to perform partic-

Model	Arabic	Chinese	German	Indonesian	Japanese	Tamil	Turkish
BLSP	2.95	0.10	16.30	13.30	0.01	0.44	2.22
UltraVox Llama 3	6.59	2.64	14.31	10.55	0.20	0.17	5.22
SALMONN	0.73	19.63	20.93	11.50	8.42	0.04	2.08
Qwen Audio	8.03	**25.11***	28.78	16.95	**22.48***	0.23	7.72
Qwen 2 Audio	**13.55**	21.47	**30.85***	**26.08***	17.02	0.74	9.58
DiVA Llama 3	12.88	12.22	27.56	22.80	6.17	**3.22***	**11.74***

Table 2: Results for Speech Translation across 7 typologically diverse languages. We evaluate using SacreBLEU (Post, 2018). * indicates significant ($P < 0.05$) improvement over other models using a paired bootstrap test.

Model	Spoken-Dialect QA										HeySquad
	USA	GBR	PHL	IND-S	IND-N	IRL	AUS	NZL	NGA	ZAF	
BLSP	44.8%	45.3%	42.9%	43.4%	43.3%	44.7%	45.8%	44.2%	42.6%	43.4%	46.8%
UltraVox Llama 3	42.6%	43.5%	39.3%	41.0%	40.4%	43.4%	42.9%	41.3%	39.9%	43.0%	43.2%
SALMONN	48.0%	46.9%	45.6%	45.5%	45.3%	47.5%	47.6%	47.8%	45.7%	45.6%	48.9%
Qwen Audio	42.3%	43.4%	41.9%	42.8%	42.4%	42.8%	45.1%	44.1%	42.1%	43.7%	45.3%
Qwen 2 Audio	44.2%	44.4%	42.6%	42.6%	41.7%	43.3%	44.1%	44.7%	41.5%	42.4%	46.2%
DiVA Llama 3	**54.6%**	**54.3%**	**52.3%**	**53.7%**	**51.9%**	**54.0%**	**55.2%**	**54.8%**	**52.6%**	**52.4%**	**55.2%**

Table 3: Results across our two Question Answering benchmarks covering both standard evaluation and robustness to regional accents. Accuracy is assessed using the PEDANTS metric, which is tuned for strong correlation with human judgments of reference-based correctness (Li et al., 2024). All improvements are significant ($P < 0.05$).

ularly well despite never having been explicitly trained on spoken questions.

Empirically, this expectation is met as shown in Table 3. DiVA significantly (P<0.05) over the baselines by at least 10% (+5 PANDA) across both benchmarks and all accents.[8]

However, it's unclear whether lower accuracy can be directly attributable to "forgetting". We qualitatively explore this question by labeling a sample of 50 responses from the HeySQUAD dataset for whether the responses include even an attempted answer relevant to the task. Qwen Audio shows signs of severe forgetting, with 30% of responses ignoring the prompt instructions entirely and instead transcribing the question e.g. *"The citation for the Pearson v. Society of Sisters case is What is the citation for the Pearson v. Society of Sisters case?"*. By comparison, SALMONN, which takes inference time interventions to reduce overfitting by partially ablating the LoRA modules learned for the base LLM, sees reduced overfitting with only 8% of model responses ignoring the prompt and instead transcribing. Qwen 2 Audio sees further reduced overfitting, likely due to its DPO process using unreleased data, with only 4% instances where the instruction is ignored. DiVA, despite being trained only on transcription data, is the only model adheres to the instruction consistently.

[8]We isolate the contribution of our methodology from base model effects through cascaded baseline analysis in Appendix A.5.

Model	IEMOCAP	MELD	MUSTARD	URFUNNY
	Weighted F1		Accuracy	
BLSP	42.9	40.1	48.4	49.9
UltraVox	32.2	33.5	56.1	50.3
SALMONN	17.4	31.7	50.1	52.3
Qwen	9.4	3.0	49.7	**54.5***
Qwen 2	33.4	37.7	**55.5**	50.6
DiVA	**50.6***	**41.3***	52.6	50.2

Table 4: Results across Emotion, Humor, and Sarcasm classification tasks. * indicates significant ($P < 0.05$) improvements computed using a paired bootstrap test.

5.1.3 Speech Classification

One possible downside of our distillation approach is that the loss function contains minimal supervision for tasks where the audio of speech itself contains rich information through tone. However, tone is frequently correlated with the semantics of the text itself. We hypothesize this mutual information provides signal for sociophonetic understanding. To assess this, we evaluate on speech classification tasks where tone is likely to play a major role: Sarcasm Detection, Humor Detection, and Emotion Recognition.

Emotion Recognition DiVA performs significantly better than all baselines on both the MELD

benchmark and IEMOCAPS benchmark. Alternative distillation-based approaches show more promise than SFT models: BLSP is somewhat competitive with DiVA (42.9 F1 on IEMOCAP), while UltraVox shows intermediate results (32.2 F1 on IEMOCAP, 33.5 F1 on MELD). In contrast, all models trained with SFT seem to struggle to predict a diverse array of labels. Qwen Audio predicts the emotion as Sadness for greater than 90% of inputs for both MELD and IEMOCAPS, while SALMONN and Qwen 2 Audio behave similarly with Neutral predictions.

While cross-modal distillation proves surprisingly effective for these tasks, this may be because IEMOCAP and MELD retain examples where emotion is detectable from text alone, limiting their ability to measure true sociophonetic understanding.

Sarcasm & Humor Detection We also evaluate on two tasks where communicative intent is expressed largely through tone. No model performs particularly well in these tasks. None of the evaluated models perform significantly ($P > 0.05$) better than chance on sarcasm detection and only Qwen Audio Chat performs better than chance on Humor Detection. This suggests there is significant progress to be made in enabling speech-oriented language models to understand more complex social signals in speech.

5.2 Qualitative User Study

Finally, to get a sense of how well the resulting models match user preferences, we recruit participants to compare DiVA to the top performing baseline, Qwen 2 Audio.

5.2.1 Recruitment & Study Design

We recruit 53 participants on the Prolific platform to provide preference ratings. Each user was allowed to contribute a maximum of 10 ratings, but able to opt-out at any time, resulting in 522 preference ratings comparing the models. We paid users 2.50$ per 10 ratings, which took fewer than 10 minutes of active time for all annotators involved, for an effective pay rate of 15$ per hour. We report annotator demographics in Appendix A.8.

We pre-screened for users who report familiarity with existing LLM chatbots and virtual assistants (e.g. ChatGPT, Gemini, Claude and others). Users were then shown responses from each model, without knowledge of which model was which.

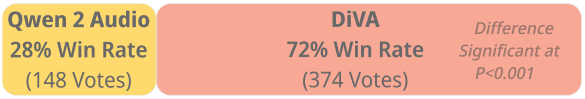

Figure 2: Win rate between models in our 522 preferences from 53 Prolific users.

Model	QA	Classification	Translation
DiVA	**55.17**	**48.70**	**13.80**
Loss Ablations			
Input Only	35.43	32.21	0.01
Output Only	52.55	34.69	0.53
Architecture Ablations			
No Init.	47.63	47.27	0.00
MLP	29.48	33.85	0.00

Table 5: Model ablations. QA shows mean scores between SD-QA and HeySquad. Classification shows mean performance between IEMOCAPS, MELD, MUSTARD, and URFUNNY. Translation shows mean BLEU scores across 7 COVOST2 languages.

To avoid any positional bias, we shuffle the order which users were shown model responses for each recording submitted.

5.2.2 Results

While there is no consistent winner across benchmarks between Qwen 2 and DiVA, DiVA generally is strongly preferred by users, with a 72% win rate at the preference level. At the user level, 41/53 (77%) of users preferred DiVA for the majority of their inputs. This is extremely promising as it indicates that the data scale reportedly used for Qwen 2 (Chu et al., 2023) may not be required for effective speech-in text-out models.

6 Architecture and Loss Ablations

To better understand each component of our model, we investigate the influence of each loss component along with our architecture decisions. In Table 5, we compare results between the complete DiVA method, using just the output distillation loss, and using just the input token alignment loss.

Impacts of KL Divergence Loss The most clear necessity for DiVA is the KL Divergence loss on the output distribution. Using token-alignment only does not simply lead to marginally worse results, it causes generations to be often incoherent. For generative tasks, the model often outputs sentences which are only vaguely semantically related to the input or unrelated markdown headers. In classifica-

tion tasks, the token-alignment only model never performs significantly better than random guessing.

Impacts of Token Alignment Loss This might raise the question: why use the token-alignment loss if it performs so poorly? In evaluations on question answering, this is certainly reasonable as the KL Divergence loss alone leads to stronger performance than the SFT baselines.

However, for translation and emotion recognition tasks, we see near-zero results from KL Divergence loss alone. Qualitatively, we observe that the distillation only model replies directly to the speech regardless of the text instructions.

We quantify this failure to adhere to instructions for the translation task using FastText Language ID (Joulin et al., 2017) on the outputs, under the assumption that outputs which are not in the correct target language are the result of ignored instructions. DiVA outputs the correct language 74% of the time while the distillation only model outputs the correct language only 1.4% of the time[9].

Impacts of Architecture Choices DiVA utilizes a Q-Former initialized from the weights of the Whisper Decoder (see Section 3). This decision results in two architecture decisions which we ablate. First is the selection of the Q-Former, rather than a simpler intervention such as a more simple projection of concatenated audio tokens, as done in the Qwen models Chu et al. (2023, 2024). Second, is the initialization of the Q-Former from pretrained weights, rather than training it from scratch.

Our ablation studies quantify the impact of both Q-Former design choices. First, removing the pretrained initialization (`No Init.`) leads to a 13.7% drop in QA performance (from 55.17 to 47.63) and a 2.9% drop in classification accuracy. Second, replacing the Q-Former architecture entirely a simple MLP further degrades performance substantially, with an additional 38.1% decrease in QA (to 29.48) and 28.1% drop in classification (to 33.85), demonstrating that both DiVA's architecture and pretrained initialization improve results.

7 Conclusion

In summary, we release DiVA, an end-to-end Voice Assistant model capable of processing text and speech natively. Our cross-modal distillation loss from text to speech showcases a promising direction for cost-effective capabilities transfer from one modality to another. Our Distilled Voice Assistant generalizes to Spoken Question Answering, Classification, and Translation despite only being trained on transcription data. Furthermore, DiVA is preferred by users to our most competitive baseline Qwen 2 Audio in 72% of instances despite DiVA taking over 100x less training compute. Together, these contributions highlight a path forward for rapid adaptation of LLMs to Speech, without significant investments in new training datasets.

8 Acknowledgements

The authors would like to thank Prof. Larry Heck, Prof. Karen Livescu, Ryan Li, Chenglei Si, Yijia Shao, and Rose Wang for their comments on this work and on audio model evaluation at various stages in this project. We also are very grateful for the code and system design review from David Hall when adding audio support into Levanter. This research is supported in part by grants from ONR grant N000142412532, and NSF grant IIS-2247357. Computing resources used for this work were funded through a Stanford HAI-GCP Cloud Credit Grant, as well through the Google TPU Research Cloud.

9 Limitations

By using weak supervision from text, DiVA is inherently only capable of learning speech signals which have some mutual information with the text transcripts. This means that DiVA is limited in the paralinguistic information it can capture at this stage. While we show that DiVA outperforms our compared Speech LLM baselines even on emotion recognition tasks, which might be seen as requiring paralinguistic signals, this limitation is unavoidable in our loss design. On the other hand, unlike pipelined models, DiVA is end-to-end finetuneable which means that it can be used to learn paralinguistic information through finetuning when it is relevant to a downstream task.

Our evaluation focuses exclusively on single-turn interactions, yet real voice assistants must handle multi-turn conversations where context accumulates across exchanges. While preliminary experiments suggest DiVA can process multi-turn interactions by concatenating hidden states across turns, we have not systematically evaluated this capability.

While we show the initialization methods we use for the modality connector offer significant

[9]We include LID results for all models in Appendix A.7

benefits, this decision also heavily constrains the architecture choice. Architecture design for multimodal adapters is an active area of research as we explore in our related work. As this field advances, the architectural insights from DiVA are likely to offer diminishing value compared to initializing a more optimal architecture from scratch.

Finally, DiVA's training approach, while computationally efficient, also relies heavily on the quality of the base LLM's responses to transcribed text. This creates a potential bottleneck where biases or limitations in the text model are transferred to the speech domain. Given the rapid rate of advances at present in text LLMs, this trade-off seems reasonable but this does limit the ability of the DiVA approach to *improve* capabilities offered by text LLMs by exploiting meaningful training data and information that is unique to speech.

10 Ethics Statement

Collecting speech data raise privacy concerns as human speech is inherently personally identifiable. While this paper focuses on technical capabilities, deployment of such systems require careful consideration of user consent, data handling, and potential misuse for surveillance or unauthorized voice processing. For our user study, users had to opt-in to microphone use before beginning the study, voice data was only stored and processed on our own servers, and recordings were erased immediately after responses were generated. Furthermore, we stored only user votes without generated responses in order to avoid risks that generated responses themselves may contain PII.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, and 1 others. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. CommonVoice: A Massively-Multilingual Speech Corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in neural information processing systems*, 33:12449–12460.

Robin N Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning equitable speech technologies for black older adults. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 379–388.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Xiangrui Cai, Haidong Xu, Sihan Xu, Ying Zhang, and Xiaojie Yuan. 2022. Badprompt: Backdoor attacks on continuous prompts. *Preprint*, arXiv:2211.14719.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole R Holliday. 2022. Training and typological bias in asr performance for world englishes. In *INTERSPEECH*, pages 1273–1277.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021a. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. WavLM: Large-scale Self-supervised Pre-training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yi-Chen Chen, Po-Han Chi, Shu-wen Yang, Kai-Wei Chang, Jheng-hao Lin, Sung-Feng Huang, Da-Rong Liu, Chi-Liang Liu, Cheng-Kuang Lee, and Hung-yi Lee. 2021b. Speechnet: A universal modularized model for speech processing tasks. *arXiv preprint arXiv:2105.03070*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TydiAa: A Benchmark for Information-seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*, 5(9):eaaw2594.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *Preprint*, arXiv:2305.06500.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. Md3: The multi-dialect dataset of dialogues. *arXiv preprint arXiv:2305.11355*.

Fahim Faisal, Sharlina Keshava, Antonios Anastasopoulos, and 1 others. 2021. Sd-qa: Spoken dialectal question answering for the real world. *arXiv preprint arXiv:2109.12072*.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, and 1 others. 2024. Prompting Large Language Models with Speech Recognition Abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage. *arXiv preprint arXiv:2111.09344*.

Shefali Garg, Zhouyuan Huo, Khe Chai Sim, Suzan Schwartz, Mason Chua, Alëna Aksënova, Tsendsuren Munkhdalai, Levi King, Darryl Wright, Zion Mengesha, and 1 others. 2023. Improving speech recognition for african american english with audio classification. *arXiv preprint arXiv:2309.09996*.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. *arXiv preprint arXiv:2104.01778*.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015a. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015b. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M Khapra. 2023. Svarah: Evaluating english asr systems on indian accents. *arXiv preprint arXiv:2305.15760*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia,

and Dilek Hakkani-Tür. 2021. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154. IEEE.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023b. YODAS: Youtube-Oriented Dataset for Audio and Speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Zongxia Li, Ishani Mondal, Yijun Liang, Huy Nghiem, and Jordan Lee Boyd-Graber. 2024. Pedants: Cheap but effective and interpretable answer equivalence. *Preprint*, arXiv:2402.11161.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. In *NeurIPS*.

Potsawee Manakul, Guangzhi Sun, Warit Sirichotedumrong, Kasima Tharnpipitchai, and Kunat Pipatanakul. 2024. Enhancing low-resource language and instruction following capabilities of audio language models. *Preprint*, arXiv:2409.10999.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4:725911.

John X. Morris, Wenting Zhao, Justin T. Chiu, Vitaly Shmatikov, and Alexander M. Rush. 2023. Language Model Inversion. *Preprint*, arXiv:2311.13647.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. 2024. Learning to compress prompts with gist tokens. *Preprint*, arXiv:2304.08467.

Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or evolution? speech interaction and hci design guidelines. *IEEE Pervasive Computing*, 18(2):33–45.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR Corpus Based on Public Domain Audio Books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, and 1 others. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, and 1 others. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. 2023. The casual conversations v2 dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10–17.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. *InterSpeech*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. 2018. Comparing speech and keyboard text entry for short messages in two

languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–23.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. 2023. Llasm: Large language and speech model. *arXiv preprint arXiv:2308.15930*.

Charlie Snell, Dan Klein, and Ruiqi Zhong. 2022. Learning by distilling context. *Preprint*, arXiv:2209.15189.

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

Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po-Chun Hsu, Duc Le, Adithya Sagar, Ali Elkahky, Jade Copet, Wei-Ning Hsu, Yossi Adi, and 1 others. 2023. Stop: A dataset for spoken task oriented semantic parsing. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 991–998. IEEE.

Pooja Upadhyay, Sharon Heung, Shiri Azenkot, and Robin N Brewer. 2023. Studying exploration & long-term use of voice assistants by older adults. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–11.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Changhan Wang, Anne Wu, and Juan Pino. 2020. CoVoST 2: A Massively Multilingual Speech-to-Text Translation Corpus. *Preprint*, arXiv:2007.10310.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2024. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *Preprint*, arXiv:2309.00916.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K Rubenstein, and 1 others. 2023. Slm: Bridge the thin gap between speech and text foundation models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and 1 others. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. 2024. Heysquad: A spoken question answering dataset. *Preprint*, arXiv:2304.13689.

Shuwen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Shang-Wen Li, Abdelrahman Mohamed, and 2 others. 2024. A Large-Scale Evaluation of Speech Foundation Models. *Preprint*, arXiv:2404.09385.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. *Preprint*, arXiv:2305.11000.

A Appendix

A.1 Contributions

Will and Diyi led the project, scoped the goals, and planned the overall experimental procedure. Will implemented and trained DiVA, as well as the inference code to serve interactive evaluations. Yanzhe helped Will design and validate the DiVA architecture and loss. Ella, Weiyan, and Michael helped format, integrate, and evaluate models on existing static benchmarks. All authors helped review, draft, and edit the writing of this work.

A.2 Reproducibility Statement

We release our training code, as well as evaluation code, demo code & raw outputs. All dataset processing details are included in Appendix A.6. We release all model weights, as well as inference code, for both ablations and the main model on HuggingFace, where they have been downloaded >150,000 times externally since our public model release on July 26th, 2024 including for extensive external evaluations in English and Thai which concluded that *DiVA is the only model that performs well on the Speech [Instruction Following] task, but it experiences a notable drop when tested on Thai.* (Manakul et al., 2024).

A.3 Allocation Between Free and Text Aligned Tokens

The alignment loss in Equation 1 operates on two dimensions: the number of text tokens N and the number of audio tokens Q. For DiVA, we fix $Q = 448$ while N varies per utterance based on the training data. This design ensures $Q - N$ tokens remain available for encoding non-textual information such as paralinguistic cues.

To validate that $N \ll Q$ in practice, we analyze the token distribution across the CommonVoice evaluation set (16,411 utterances):

Statistic	Number of Tokens
Mean	15.22
Standard Deviation	4.08
Maximum	39
75th Percentile	18
Median	15
25th Percentile	12
Minimum	5

Table A.1: Distribution of text tokens (N) across CommonVoice evaluation utterances. With $Q = 448$ audio tokens, over 90% remain unconstrained by text alignment.

These statistics demonstrate that at most 39 tokens (8.7% of audio tokens) and typically only 15 tokens (3.3%) are directly aligned with text content. Consequently, the vast majority of audio tokens (>90%) remain unconstrained by the alignment loss, preserving capacity for paralinguistic information. This empirical validation supports our design choice and addresses concerns about information bottlenecks in the alignment mechanism.

A.4 Toy Experiment on KL Divergence versus Hidden State Alignment

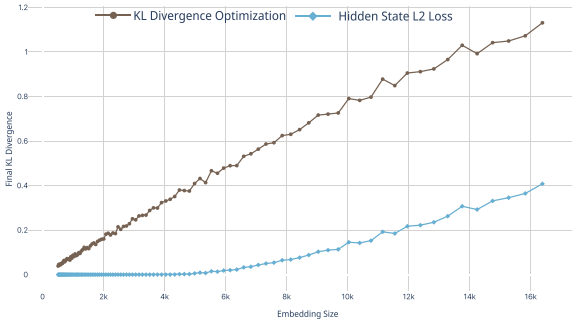

Figure A.1: Empirical Comparison of the KL Divergence with our Proxy L_2 loss in a toy experimental setup. Optimizing the KL Divergence directly leads to *worse* KL Divergence than optimizing the L_2 loss. This gap increases as the hidden dimension becomes larger.

Beyond being a valid and efficient approximate of the KL Divergence, the L_2 loss should offer a more stable gradient, especially early in training when the output distributions are extremely different. When P_t is positive and P_s is near zero, the KL divergence explodes to extremely large values which can make optimization difficult and subject to significant numerical error.

In order to test this intuition, we set up a toy experiment where the student model outputs a single hidden state h_s and the teacher model outputs a single hidden state h_t. In this highly simplified space, each model is fully parameterized by the these hidden states. We initialize and output vocabulary from the normal distribution with $32,000$ vocabulary items. Then, we optimize h_s based on either the L_2 distance with h_t or the KL Divergence with the output probabilities. Finally, for both procedures, we optimize for 100 steps with stochastic gradient descent, running the experiment 100 times at logarithmically increasing embedding dimensions, and plot the final KL divergence achieved under each loss function.

We see that, as the embedding dimension grows, optimizing the L_2 loss actually achieves *lower* KL divergence in this setup than optimizing the KL Divergence directly. To some extent, this makes sense as the L_2 loss is an incredibly simple convex function to optimize in this setting, while the KL divergence introduces significant additional complexity and a much sharper loss landscape early in optimization. We used this setup early in model design phases to help validate the choice of this approximation empirically, without training full scale models.

A.5 Isolating Methodological Contributions from Base Model Effects

SDQA	DiVA Improvement over			Llama 3 Improvement over	
	BLSP	Qwen 2 Audio	Qwen Audio	vs. Qwen	vs. Llama 2
Australia	+9.44	+11.15	+10.11	+5.16	+3.29
New Zealand	+10.56	+10.01	+10.67	+5.21	+3.55
Great Britain	+9.03	+9.96	+10.93	+5.11	+3.03
United States	+9.80	+10.39	+12.31	+5.54	+3.80
Ireland	+9.21	+10.66	+11.17	+5.12	+2.96
South India	+10.31	+11.12	+10.94	+4.70	+3.09
South Africa	+9.02	+9.95	+8.67	+5.10	+3.09
Philippines	+9.47	+9.75	+10.40	+4.88	+3.05
Nigeria	+10.07	+11.19	+10.59	+3.81	+1.72
North India	+8.58	+10.19	+9.51	+4.81	+2.30

Table A.2: DiVA's improvements over end-to-end Speech LLMs compared to base model differences in cascaded systems. The cascaded results show the performance gap when using Whisper + different LLMs on the same task.

To isolate the contribution of our distillation methodology from the choice of base LLM, we compare DiVA's performance gains against the inherent differences between base models. Specifically, we evaluate cascaded ASR + LLM pipelines where different LLMs respond to Whisper transcriptions, providing an upper bound on base model contributions:

The analysis reveals that DiVA's improvements (9-11 percentage points) substantially exceed the performance differences attributable to base model selection alone (3-5 percentage points in cascaded systems). This demonstrates that our cross-modal distillation methodology contributes significant value beyond the choice of Llama 3 8B as the base model.

A.6 In-Depth Evaluation Description

A.6.1 Spoken Question Answering

HeySquad HeySquad (Wu et al., 2024) is a spoken question answering (QA) dataset that aims to measure the QA ability of digital agents. It is based on the SQuAD dataset Rajpurkar et al. (2016) with 76K human-spoken and 97K machine-generated questions, and the corresponding answers. We evaluate the models on the open-source validation set with around 4K QA pairs.

Spoken Dialect Question Answering (SDQA) SDQA (Faisal et al., 2021) assesses the robustness of Spoken Language Understanding to global phonological variation in English. The dataset is made up of the same 1000 questions spoken and recorded by speakers in 10 accent regions where English is frequently spoken. We evaluate on the 494 of these questions which contain ground truth answers.

A.6.2 Speech Classification

Emotion Recognition **Interactive Emotional Dyadic Motion Capture (IEMOCAP)** IEMOCAP (Busso et al., 2008) is a dataset of ∼12 hours of videos, audio, motion capture, and transcripts of actors performing both improvised and scripted scenes. The seven professional and three student actors perform emotionally expressive scenes. Each conversation turn in each scene was labeled by six evaluators as demonstrating "happiness," "sadness," "anger," "surprise," "fear," "disgust," "frustration," "excitement," "neutral state," or "other." We follow Yang et al. (2024) and remove unbalanced class labels, resulting in 1241 audio utterances in the fifth fold used by Tang et al. (2023).

Multimodal EmotionLines Dataset (MELD) MELD (Poria et al., 2019) contains 13,708 utterances labeled by emotion and collected from the sitcom *Friends*. MELD builds on EmotionLines (Hsu et al., 2018); however, the authors of MELD ask annotators to watch the videos instead of simply reading the transcripts to produce labels. Three graduate student annotators labeled all utterances for emotions: "anger," "disgust," "fear," "joy," "neutral," "sadness," and "surprise," as well as for sentiments "positive," "negative," "neutral." We evaluate on the test set of 2608 utterances.

Communicative Intent Recognition Multimodal Sarcasm Dataset (MUSTARD) MUSTARD (Castro et al., 2019) is a collection of 690 clips from the TV shows Friends, The Golden Girls, The Big Bang Theory, and Sarcasmaholics Anonymous, labeled as sarcastic or non-sarcastic by three annotators. The clips were collected primarily from YouTube using keywords like *Chandler sarcasm, Friends sarcasm, etc.* and sampled from MELD (Poria et al., 2019). The final dataset was filtered to have an even number of labels of sarcastic and non-sarcastic clips. We evaluate on all 690 clips to test the models' capability in understanding intended sarcasm.

URFunny URFunny (Hasan et al., 2019) is a multimodal humor recognition benchmark constructed from 90.23 hours of TED talk recordings, spanning 1741 speakers and 417 topics. TED produces transcripts for the talks, which contain "[laughter]" markers that show when the audience laughs. The authors sampled the context and punchline before laughter markers for 8257 positive examples and random parts of the transcript without laughter markers for 8257 negative examples. URFunnyV2 filters out noise and reduces overlap in examples. We evaluate 7614 examples from the train split of URFunnyV2 to evaluate the models' ability to understand speakers' humorous intents.

A.6.3 Speech Translation

CoVoST 2 CoVoST 2 (Wang et al., 2020) is a speech-to-text translation benchmark to and from English. The speech inputs are sourced from the CommonVoice and professional translators are hired to translate the recording into a target language. The test dataset is large, made up of 15,500 examples translated from English to each target language. We evaluate on 7 target languages selected for their typological diversity in prior work (Clark et al., 2020).

A.7 Language ID Outputs For All Models

Language	DiVA	KL Only	Token Alignment	Qwen	Qwen 2	SALMONN
Arabic	84%	2%	0%	95%	90%	19%
German	90%	1%	0%	99%	98%	77%
Indonesian	85%	1%	0%	97%	97%	77%
Japanese	28%	2%	0%	100%	99%	67%
Tamil	96%	1%	0%	60%	79%	8%
Turkish	74%	1%	0%	93%	92%	28%
Mandarin	60%	2%	0%	91%	83%	93%

Table A.3: Percentage of outputs for which Language ID matches the target language.

A.8 Prolific User Demographics

Age		Gender Identity		Ethnicity (Simplified)	
Median Age	34	Man	50.9%	White	59.2%
Max Age	69	Woman	49.1%	Black	16.3%
Minimum Age	19	Other	0%	Asian	12.2%
				Mixed	4.1%
				Other	8.2%

Table A.4: Aggregate metrics for age, gender identity, and ethnicity from our user study. Our participants cover a wide range of ages, are gender balanced, and have a similar distribution of ethnicities as reported in the United States Census.