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

Anonymous authors
Paper under double-blind review

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 (LLMs) 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, we show that DiVA better meets user preferences, achieving a 72% win rate compared with state-of-the-art models like Qwen 2 Audio, despite using >100x less training compute.
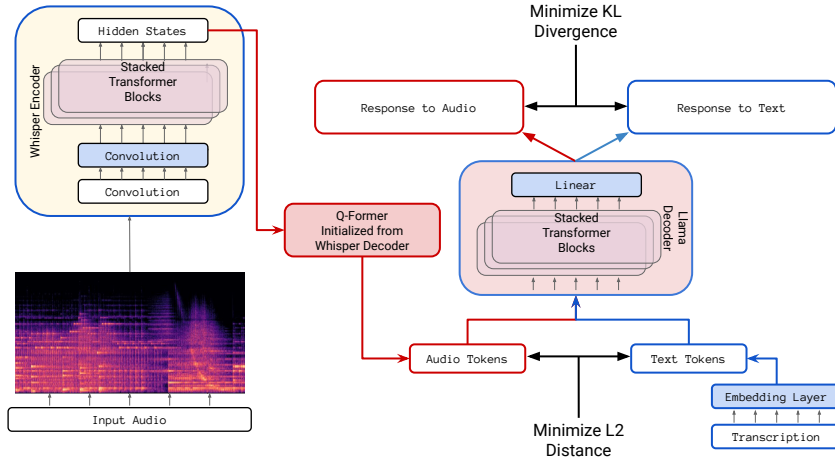

Figure 1: Training Pipeline for Distilled Voice Assistant (DiVA), Red indicates trainable components while Blue indicates frozen pretrained modules. DiVA modifies a text-only LLM into a general purpose Speech LLM by using the model's own responses to transcribed speech as self-supervision.

1 Introduction

As Large Language Models (LLMs) capabilities increase, 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 interaction surface for language technology (Murad et al., 2019), offering measurable efficiency gains for users (Ruan et al., 2018). One straightforward method of integrating speech with LLMs is to feed audio to an Automatic Speech Recognition (ASR) model and produce a text transcription for the LLM to use. However, this process loses meaningful information carried through tone, pacing, and accent (Upadhyay et al., 2023) regardless of the transcription accuracy. Furthermore, finetuning these pipelined systems requires supervision for both transcription and response generation, increasing annotation complexity and costs.

Table 1: High-Level comparison with state-of-the-art open-access Speech & Audio LLMs which we compare to. DiVA offers an entirely different approach to training using context distillation.

Model	Base LLM	Training Method	# Tasks	# Hours
SALMONN	Alpaca	SFT	12	4400
Qwen Audio Chat	Qwen Chat	SFT	31	∼50k
Qwen 2 Audio Instruct	Qwen2 Instruct	SFT, DPO	Unreported	>370k
DiVA (Ours)	Llama 3	Distillation	N/A	3.5k

As such, LLMs that interface with speech directly 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., 2023a), many of which utilize the same well-established approach: large-scale multi-task supervised finetuning (SFT).

Models using SFT face several challenges. First, without a large degree of task diversity in their training data, they 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". In order to generalize well, SFT must be trained with labeled data from a wide range of tasks and domains, with minimal imbalance between tasks. However, broad annotated speech instruction training data does not currently exist.

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, however, we argue that these "limitations" of existing data are artificially imposed by SFT. The speech community has already 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 data in just one task: Automatic Speech Recognition (ASR). This means that models trained with SFT cannot make use of the entirety of this data without "forgetting" non-ASR capabilities.

We solve the "forgetting" problem by training a model that generalizes well despite using *only* ASR data. 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, the CommonVoice, consisting of speech and transcriptions contributed by volunteers around the world and recorded on their own devices (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 trains more efficiently and generalizes better *without* requiring investment in 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.

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. For these features to be usable by the LLM, the dimensionality must be reduced, either by stacking consecutive features (Wu et al., 2023a; 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.

How can speech LLMs be trained to achieve instruction following abilities using existing data? Prior work has explored two routes for creating instruction data without major financial investment. The first approach is to transform many existing datasets into an instruction-following format (Dai et al., 2023; Chu et al., 2023; Tang et al., 2023; Dai et al., 2023; Liu et al., 2023a). In this case, limitations are often set by datasets that are not aligned with intended LLM usage and imbalances across tasks. The second approach is to train on synthetic responses to text representations of the new modality from commercial models (Liu et al., 2023b; Gong et al., 2023).

Our approach is most related to the latter. Rather than generating data with an external model, we capitalize on the idea that instruction-tuned language models can provide valuable learning signals as long as the input is within the textual modality. Using the output distribution in response to text transcripts, we can more strongly guarantee the transfer of existing capabilities using context distillation (Snell et al., 2022; Mu et al., 2024).

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 virtual assistant 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 reports no training 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. Beyond our novel method, we believe this too 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. Similar to prior works, we initialize the audio encoder from the 1.5B parameter Whisper-Large-v3 model. Unlike prior works, we use all components of Whisper: not only reusing the encoder but also initializing a Q-Former from the decoder. We train this architecture using distillation loss on the input and output distribution of the text-only 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 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).

Audio-Text Feature Alignment While the Whisper encoder extracts meaningful audio features, they are encoded at high granularity, with one token for every 40 milliseconds of input audio. By comparison, humans speak around one syllable every 150 milliseconds on average 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 granulatity 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 $\sigma(\frac{\boldsymbol{Q}(\boldsymbol{KA}^\mathsf{T})}{\sqrt{d_k}})(\boldsymbol{VA})$ 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 for ASR: mapping audio embeddings to discrete text tokens. 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 a set of $\{\boldsymbol{t}_q^{audio} \in \mathbb{R}^{H \times |Q|}\}$ output tokens representing the audio.

Text Decoding For language processing and instruction following capabilities, we use Llama 3 (Dubey et al., 2024)[1] 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 prior work for Vision-LMS (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 CROSS-MODAL TOKEN ALIGNMENT LOSS

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 audio recording where $|Q| > N$. We align these representations by minimizing the L_2 distance between the text embeddings and the final i audio embeddings:

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

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.

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 appears to enable reasoning between text and audio tokens, vastly improving text instruction adherence.

---

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

3.2.2 DISTILLATION FROM OUTPUT EMBEDDING DISTANCE

Voice assistant models should give coherent, helpful, and harmless responses to user speech. Thankfully, many openly accessible text-only LLMs have 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. While, in theory, input token alignment could achieve this, even minor differences in input embeddings can significantly affect model behavior in practice (Cai et al., 2022).

Distillation loss, on the other hand, directly optimizes for the similarity of the output distribution (Hinton et al., 2015). 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[2]. (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 read 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[3]. The first factor means that the resulting DiVA models we release can be adopted for use broadly, while the latter two help make the training data more representative of real users.

---

[2]We explore this with an isolated small-scale experiment in Appendix A.2

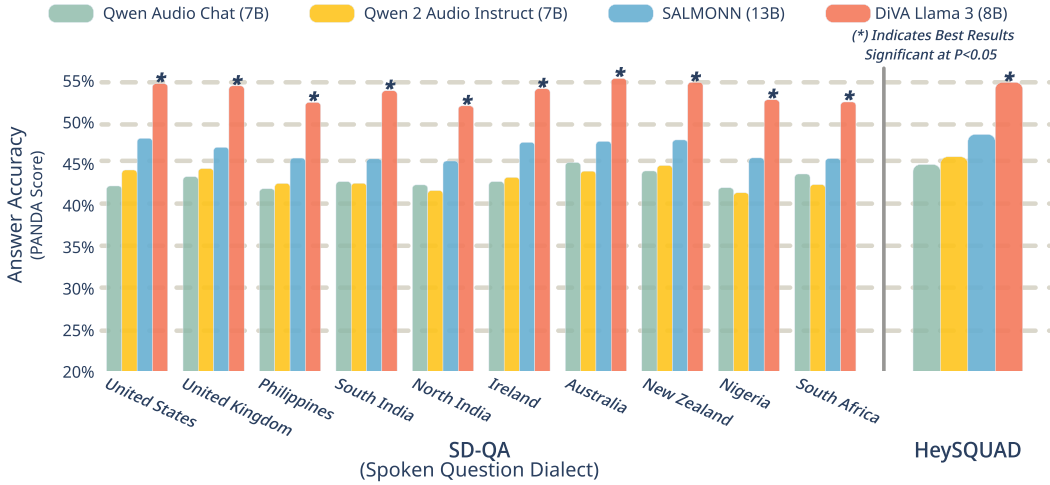[3]Statistics drawn from the official CommonVoice tracker

Figure 2: Results across our two Question Answering benchmarks covering both standard evaluation and robustness to regional accents. Model correctness is assessed using the PANDA metric, which is tuned for strong correlation with human judgments of correctness (Li et al., 2024), and significance is from a paired bootstrap test (Dror et al., 2018).

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 learning rate schedule which decays the learning rate to 0 over the course of the training run. The training run completes in approximately 12 hours on on a TPU v4-256 pod.

5 QUANTITATIVE AND QUALITATIVE EVALUATIONS

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

However, these benchmarks were all designed to test single task systems focused on each individual task. It is unclear whether these benchmarks capture the capabilities users expect from virtual assistants that speech LLMs are now powering commercially. To assess this, we run a further side-by-side comparison of DiVA with the best performing model on the benchmark evaluation Qwen 2 Audio (Chu et al., 2024).

Baselines We compare our results to three openly available Speech Language Models: SALMONN, Qwen Audio Chat, and Qwen 2 Audio Instruct. Notably, all the baseline models utilize SFT covering these benchmark tasks. This makes them strong baselines: they all use similar scale base LLMs to DiVA, all make use of 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

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 IEMO-CAP (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. Finally, speech translation is tested on CoVoST 2, translating
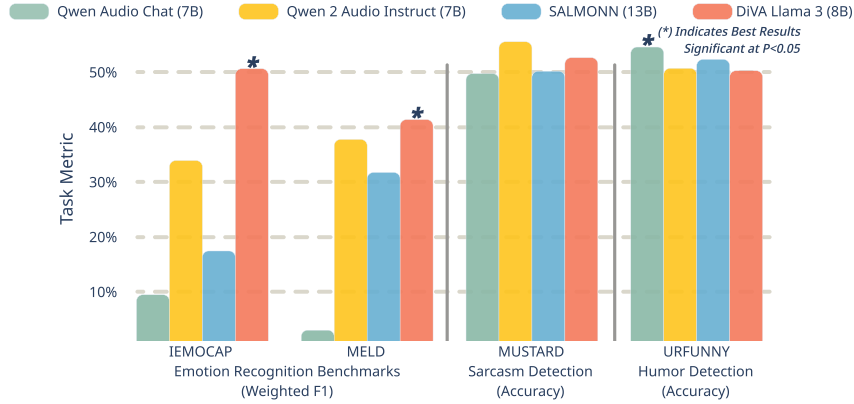
Figure 3: Results across Emotion, Humor, and Sarcasm classification tasks. We measure class-weighted F1 for multi-class classification and accuracy for binary classification. Significance computed using a paired bootstrap test.

15,500 English examples into seven commonly-tested typologically diverse languages (Clark et al., 2020). These datasets cover a wide range of traditionally benchmarked speech tasks drawn from prior work, which we cover in greater depth in Appendix A.3.

5.1.1 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 and the prompt: *You are a helpful assistant. Give answers as a simple single sentence*. 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 a case where we expect DiVA to perform particularly well despite never having been explicitly trained on spoken questions.

Empirically, this expectation is met as shown in Figure 2. DiVA significantly (P<0.05) over the baselines by at least 10% (+5 PANDA) across both benchmarks and all accents.

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.

5.1.2 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 may enable the audio encoder to transfer some amount of this tone information based on weak supervision from text. 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 For emotion classification, we prompt each model with the instructions *Respond in a single word what emotion the input exhibits. If there is no clear emotion, respond*
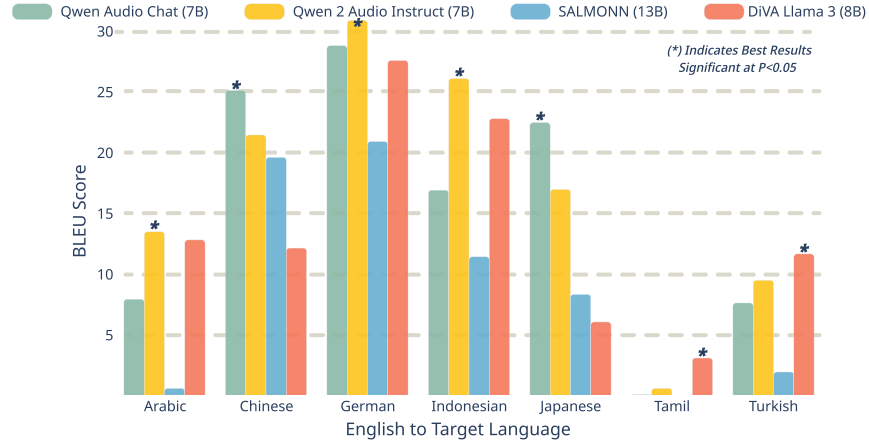
Figure 4: Results for Speech Translation across 7 typologically diverse languages. We evaluate using SacreBLEU and compute confidence intervals using a Paired Bootstrap.

'Neutral'. To use each model as a classifier, we follow prior work (Hendrycks et al., 2021) and use the log-probabilities assigned to each possible label as classifier scores.

DiVA performs significantly better than both baselines on both the MELD benchmark, sourced from television, and IEMOCAPS, which operates over recorded conversations. In comparison to DiVA, both baseline models 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 behaves similar with Neutral predictions.

These results are quite surprising given that DiVA is trained without explicit emotion supervision. However, many examples in both IEMOCAPS and MELD communicate emotion through both text and audio signals which may confound how well these evaluations capture true sociophonetic signal.

Sarcasm & Humor Detection We also evaluate on two tasks where communicative intent is expressed largely through tone. For Sarcasm Detection, we prompt each model to *Respond whether the input is sarcastic. Answer with a simple yes or no* for sarcasm detection and *Respond whether the input is intended to be humorous. Answer with a simple yes or no*. In both cases, we compare the log-probability assigned to either the token "Yes" or the token "No.

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.

These tasks also highlight a shortcoming of distillation – namely, that DiVA inherits even non-desirable behaviors from the base LLM. Even when asked whether obviously very serious text is humorous or sarcastic, Llama 3 will almost always find a way to argue that it is humorous. DiVA inherits this behavior, predicting the "Yes" label in both tasks over 90% of the time.

5.1.3 SPEECH TRANSLATION

Finally, we assess the speech-to-text translation capabilities of each model from English Speech to text in another language. We prompt each model with the instruction *Translate the input from {input_lang} to {output_lang}*.

Results on this benchmark are far more mixed. The original Qwen Audio performs best on Chinese and Japanese, Qwen 2 Audio performs best on Arabic, German and Indonesian, and DiVA performs best on 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 it trains on similar or increased volumes of data from CoVost2 as the original Qwen. This
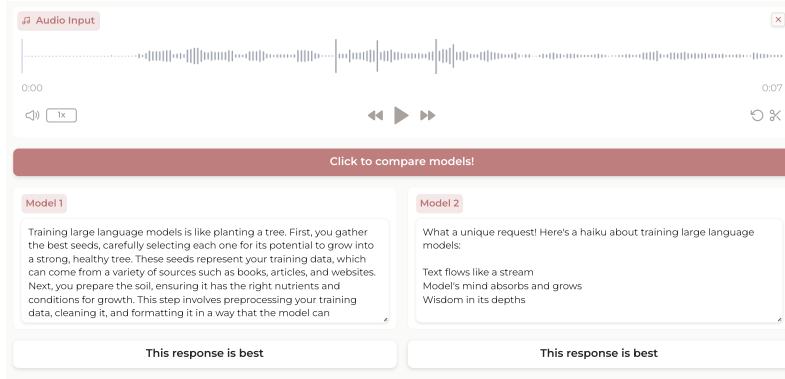
Figure 5: Example of the double-blind interface for the user study with responses (Left: Qwen 2, Right: DiVA) to the speech *Can you tell me about Large Language Models in the style of a haiku?*.
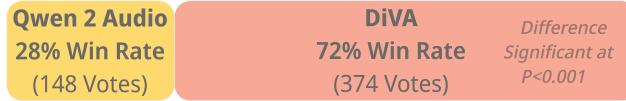

Figure 6: Win-rate between models in our 522 preferences from 53 Prolific users.

highlights the data and compute efficient transfer of the DiVA approach, as both of these models trained on more translation specific data than DiVA used for it's entire training.

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 Romanji) rather than the expected native script. This leads to especially poor results in these languages.

5.2 QUALITATIVE USER STUDY

Finally, to get a sense of how well suited the resulting models are to user preferences, which may or may not be well captured in existing benchmarks, 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.5.

We pre-screened for users who report familiarity with existing LLM chatbots and virtual assistants (e.g. ChatGPT, Gemini, Claude and others). In order to avoid biasing our participants, we prompt them without reference to specific tasks to *Record something you'd say to an AI Assistant! Think about what you usually use Siri, Google Assistant, or ChatGPT for.* Users were then shown responses from each model, without knowledge of which model was which. To avoid any positional bias, we shuffle the order which users were shown model responses for each recording submitted.

5.2.2 RESULTS

Despite no clear winner on all benchmarks for 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. Beyond showing that DiVA improves preference alignment, this indicates that benchmarks may not correlate with practical usage.

6 LOSS ABLATION

To better understand each component of our distillation loss, we investigate the influence of each loss component independently. In Figure 7, 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 classification tasks, the token-alignment only model never performs significantly better than random guessing.
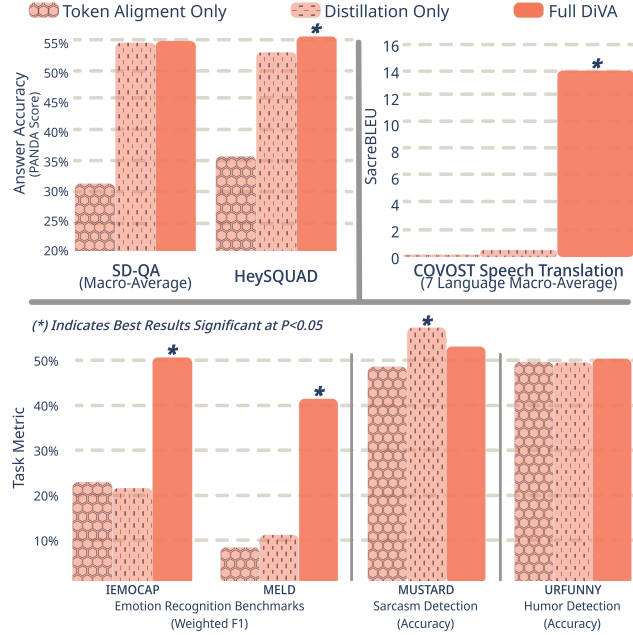

Figure 7: Ablation of the loss components from Section 3. Distillation leads to a capable audio-only model, but token-alignment improves instruction adherence.

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 since using the KL Divergence loss alone already 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[4].

This indicates that the token-alignment loss is key to achieving a model using distillation that can follow both audio instructions and text instructions such as a system prompt.

7 CONCLUSION

In summary, we release DiVA, an end-to-end Voice Assistant model capable of processing text and audio 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 large investments in new training datasets[5].

---

[4]We include LID results for all models in Appendix A.4

[5]We include anonymized links to training and evaluation code in Appendix A.1

REFERENCES

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

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

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

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

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

Xiangrui Cai, Haidong Xu, Sihan Xu, Ying Zhang, and Xiaojie Yuan. Badprompt: Backdoor attacks on continuous prompts, 2022. URL https://arxiv.org/abs/2211.14719.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1455. URL https://aclanthology.org/P19-1455.

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

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

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

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. Speechnet: A universal modularized model for speech processing tasks. *arXiv preprint arXiv:2105.03070*, 2021b.

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. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

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

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio Technical Report, 2024. URL https://arxiv.org/abs/2407.10759.

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

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

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

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

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL `https://aclanthology.org/P18-1128`.

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, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur,

Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

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

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

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

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

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

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

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

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2046–2056, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1211. URL https://aclanthology.org/D19-1211.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. Emotion-Lines: An Emotion Corpus of Multi-Party Conversations. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1252.

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

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2068.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1147–1154. IEEE, 2021.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023a. URL `https://arxiv.org/abs/2301.12597`.

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

Zongxia Li, Ishani Mondal, Yijun Liang, Huy Nghiem, and Jordan Lee Boyd-Graber. Panda (pedantic answer-correctness determination and adjudication):improving automatic evaluation for question answering and text generation, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, 2023a.

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

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. "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, 2021.

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

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens, 2024.

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

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

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

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

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

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL `https://aclanthology.org/P19-1050`.

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

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. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

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

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

Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. 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, 2018.

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

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context, 2022.

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

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

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

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

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

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. HeySQuAD: A Spoken Question Answering Dataset. *arXiv preprint arXiv:2304.13689*, 2023b.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. Heysquad: A spoken question answering dataset, 2024. URL https://arxiv.org/abs/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, Shinji Watanabe, and Hung yi Lee. A Large-Scale Evaluation of Speech Foundation Models, 2024. URL https://arxiv.org/abs/2404.09385.

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

A APPENDIX

A.1 REPRODUCIBILITY STATEMENT

We release our training code, as well as evaluation code, demo code & raw outputs at Anonymized links for the purpose of review. All dataset processing details are included in Appendix A.3.

A.2 TOY EXPERIMENT ON KL DIVERGENCE VERSUS HIDDEN STATE ALIGNMENT

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.
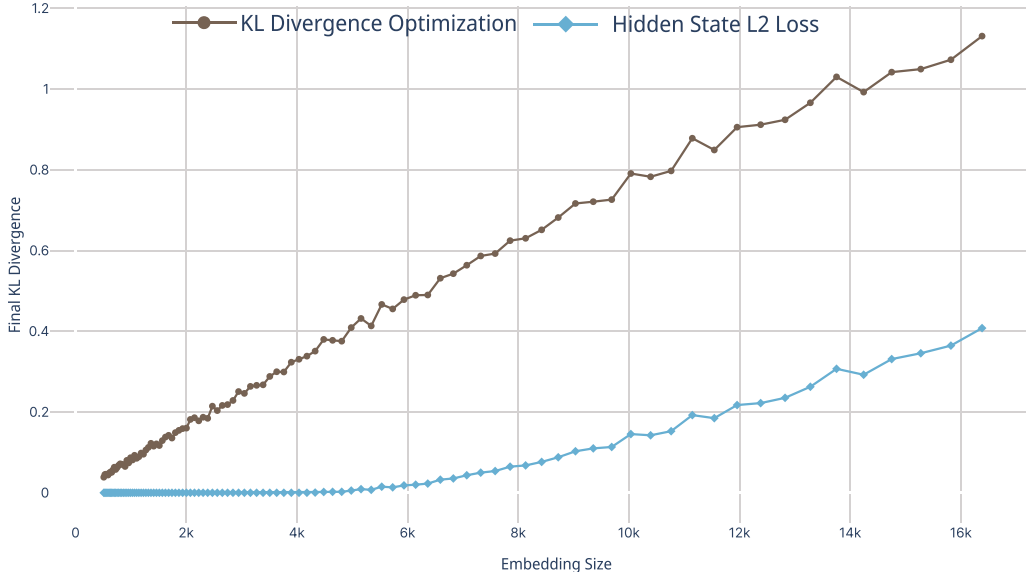

Figure 8: Empirical Comparison of the KL Divergence with our Proxy L_2 loss in a toy experimental setup. Optimizing the KL Divergence directly leads to *worse* KL Divergence than optimizing the L_2 loss. This gap increases as the hidden dimension becomes larger.

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.

17

A.3 In-Depth Evaluation Description

A.3.1 Spoken Question Answering

HeySquad HeySquad Wu et al. (2023b) 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.3.2 Speech Classification

Emotion Recognition Interactive Emotional Dyadic Motion Capture (IEMOCAP) IEMO-CAP 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) MUS-TARD 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. URFun-nyV2 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.3.3 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.4 LANGUAGE ID OUTPUTS FOR ALL MODELS

Table 2: Percentage of outputs for which Language ID matches the target language

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%

A.5 PROLIFIC USER DEMOGRAPHICS

Table 3: Aggregate Metrics for Age, Gender Identity, and High-Level Ethnicity Information 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.

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%