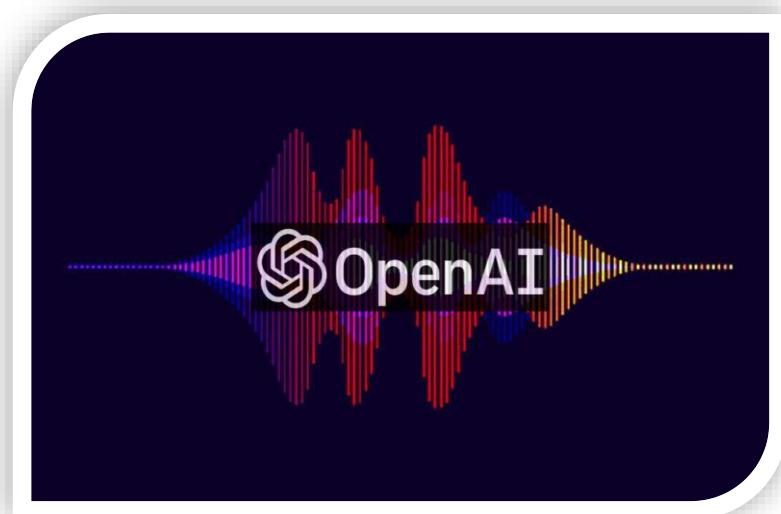


## ***Navigating Robust Speech Recognition via Large-Scale Weak Supervision*** ***Research Paper***

Traditional automatic speech recognition (ASR) systems rely heavily on large, meticulously labeled datasets for supervised learning. Curating such datasets can be a significant investment in terms of time and cost.

Whisper explores a weakly-supervised learning approach by leveraging the vast amount of audio-text data readily available online. This data, while not perfectly aligned, offers a substantial volume for training.



Whisper explores leveraging **weakly-supervised learning**??

That's right! Whisper, the speech recognition system by OpenAI, is built on the idea of weakly-supervised learning.

Here's the gist of how it works:

**Large scale, diverse data:** Whisper trains on a whopping 680,000 hours of audio data, encompassing multiple languages and various content types. This vast amount of data, though not meticulously labeled, provides a rich learning environment. Whisper tackles multiple NLP tasks concurrently, including speech recognition, language translation, and speaker accent identification.

**Focus on transcripts:** Instead of needing perfectly labeled data points, Whisper thrives on transcripts – written versions of the spoken audio. Even if these transcripts aren't 100% accurate, they offer enough clues for Whisper to learn the connection between speech and text.

Weakly-supervised approach offers Whisper some advantages:

- **Better handling of real-world audio:** Real-world speech is messy, with accents, background noise, and jargon. Whisper, trained on this very kind of data, performs well in these scenarios compared to models trained on pristine recordings.
- **Robustness across languages:** The multilingual nature of the training data makes Whisper adaptable to various languages, even those with less representation.

While Whisper excels at handling diverse speech varieties, it's important to acknowledge that it might not achieve the same level of accuracy on specific tasks compared to models trained with perfectly labeled data for those tasks. However, its strength lies in its ability to generalize effectively across a wider range of speech scenarios.

### Introduction:

One of the key advancements in speech recognition has been the development of unsupervised pre-training, like Wav2Vec 2.0. These techniques are exciting because they can leverage massive amounts of unlabeled speech data. Traditionally, supervised learning relied on meticulously labelled datasets, often limited to just a thousand hours. Unsupervised pre-training sidesteps this bottleneck by learning directly from raw audio. This

allows us to scale up training data by orders of magnitude – recent work has used up to 1 million hours! The impact is significant, particularly in low-resource settings. When fine-tuned on standard benchmarks, these models achieve state-of-the-art performance.



## Challenges of Unsupervised Pre-training for Speech Recognition

Unsupervised pre-training of audio encoders has shown promise in learning informative speech representations. However, these models inherently lack a well-performing decoder to translate these representations into actionable outputs for tasks like speech recognition. This necessitates a fine-tuning stage, which can be a complex process requiring specialized expertise.

Furthermore, fine-tuning introduces the risk of overfitting. Machine learning excels at identifying patterns within training data that improve performance on unseen data from the same source. However, some of these patterns might be spurious and fail to generalize to new datasets or distributions.

This suggests that while unsupervised pre-training has improved the quality of audio encoders dramatically, the lack of an equivalently high-quality pre-trained decoder, combined with a recommended protocol of dataset-specific finetuning, is a crucial weakness which limits their usefulness and robustness.

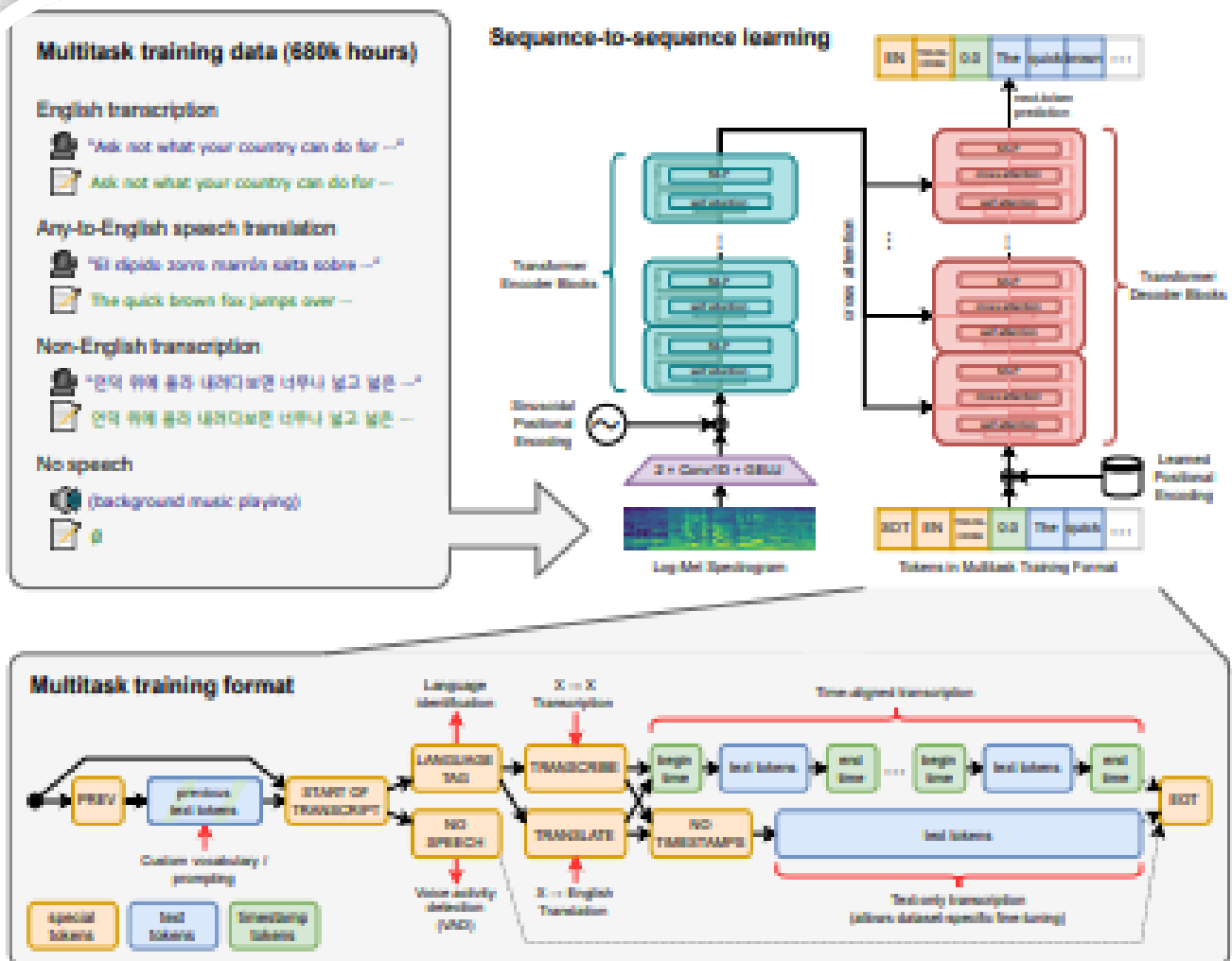
The goal of a speech recognition system should be to work reliably “out of the box” in a broad range of environments without requiring supervised fine-tuning of a decoder for every deployment distribution.

This work significantly advances weakly supervised speech recognition by scaling the training data to an unprecedented level. Whisper: Scaling Weakly Supervised Speech Recognition with Massive Multilingual Data: They introduce Whisper, a model trained on a massive dataset of 680,000 hours of weakly labeled audio data. This approach achieves state-of-the-art performance on existing benchmarks without the need for fine-tuning on specific datasets, demonstrating remarkable transferability.

Beyond scaling, Whisper pushes the boundaries of weakly supervised pre-training by incorporating multilingual and multitask learning. Their work includes 117,000 hours of data in 96 languages and 125,000 hours for speech-to-text translation. This demonstrates that large-scale weakly supervised learning can be highly effective, even surpassing the need for complex self-supervision or self-training techniques often employed in recent speech recognition research.

**MODEL:**

Since the focus of their work is on studying the capabilities of large-scale supervised pre-training for speech recognition, they use an off-the-shelf architecture to avoid confounding our findings with model improvements. They chose an encoder-decoder Transformer as this architecture has been well validated to scale reliably.



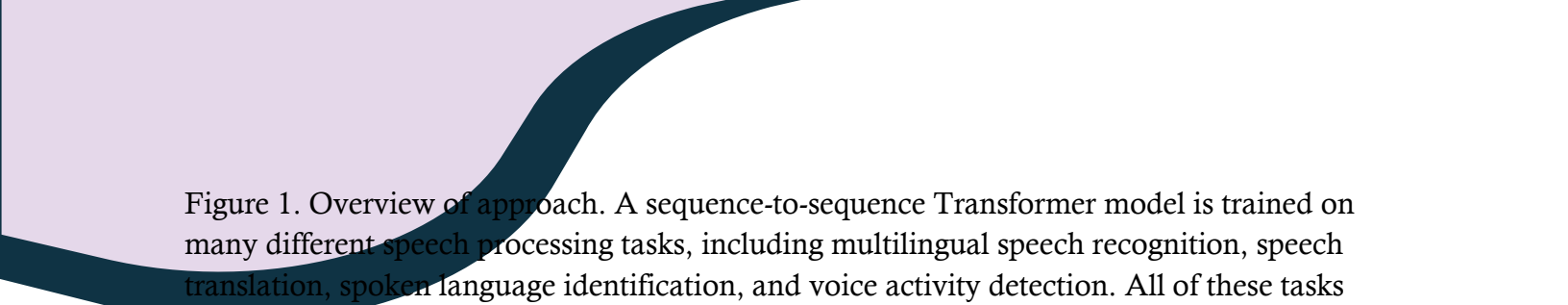


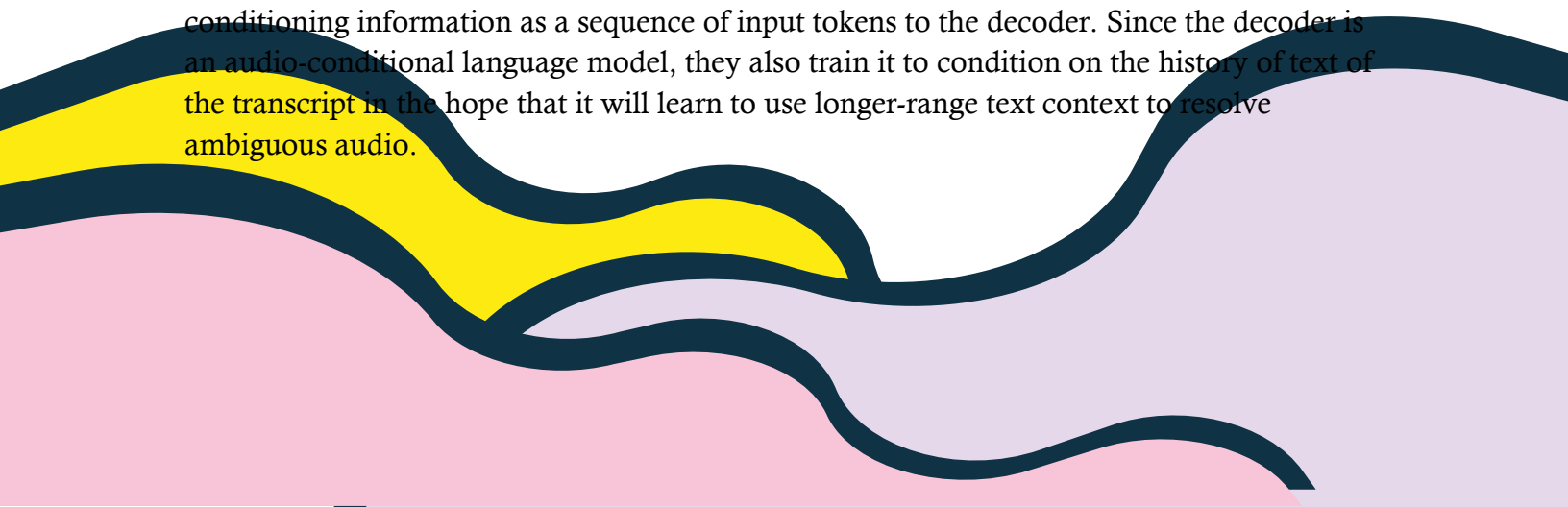
Figure 1. Overview of approach. A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline.

The encoder processes this input representation with a small stem consisting of **two convolution layers** with a filter width of 3 and the GELU activation function where the **second convolution layer** has a stride of two. Sinusoidal position embeddings are then added to the output of the stem after which the encoder Transformer blocks are applied. The transformer uses pre-activation residual, and a final layer normalization is applied to the encoder output. The decoder uses learned position embeddings and tied input-output token representations. The encoder and decoder have the same width and number of transformer blocks.

They use the same byte-level BPE text tokenizer used in GPT2 for the English only models and refit the vocabulary (but keep the same size) for the multilingual models to avoid excessive fragmentation on other languages since the GPT-2 BPE vocabulary is English only.

Predicting which words were spoken in a given audio snippet is a core part of the full speech recognition problem but it is not the only part. A fully featured speech recognition system can involve many additional components such as voice activity detection, speaker diarization, and inverse text normalization. These components are often handled separately, resulting in a relatively complex system around the core speech recognition model. To reduce this complexity, they would like to have a single model perform the entire speech processing pipeline, not just the core recognition part. An important consideration here is the interface for the model. There are many different tasks that can be performed on the same input audio signal: transcription, translation, voice activity detection, alignment, and language identification are some examples.

For this kind of one-to-many mapping to work with a single model, some form of task specification is necessary. They have used a simple format to specify all tasks and conditioning information as a sequence of input tokens to the decoder. Since the decoder is an audio-conditional language model, they also train it to condition on the history of text of the transcript in the hope that it will learn to use longer-range text context to resolve ambiguous audio.



“Specifically, with some probability they add the transcript text preceding the current audio segment to the decoder’s context. They indicate the beginning of prediction with a token.

First, predict the language being spoken which is represented by a unique token for each language in our training set (99 total). These

language targets are sourced from the aforementioned VoxLingua107 model. In the case where there is no speech in an audio segment, the model is trained to predict a token indicating this. The next token specifies the task (either transcription or translation) with an or token. After this, we specify whether to predict timestamps or not by including a token for that case.” At this point, the task and desired format is fully specified, and the output begins.

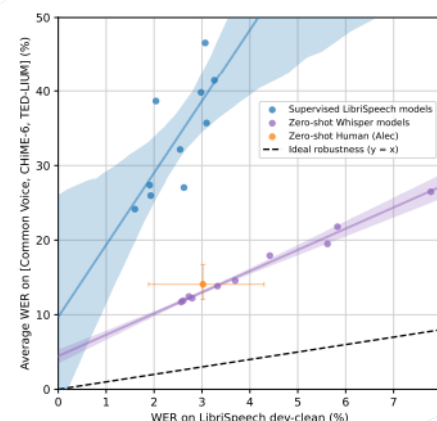
Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

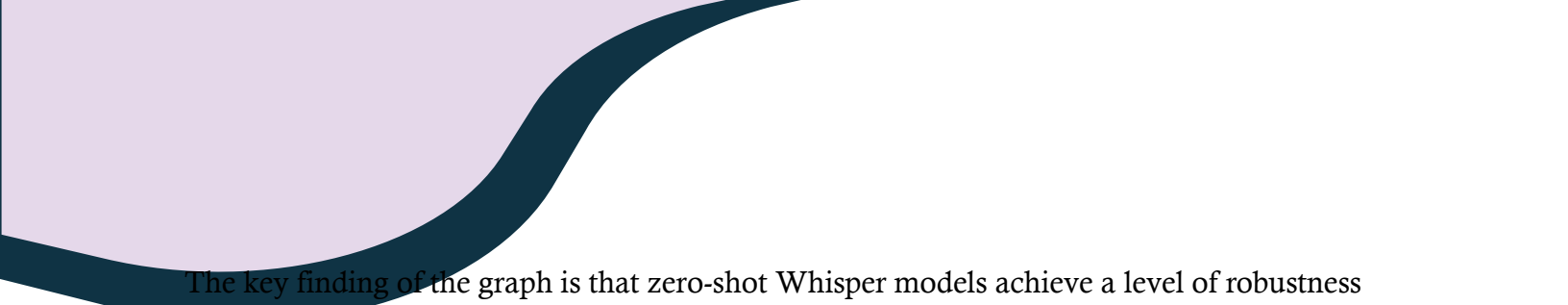
Table 1. Architecture details of the Whisper model family.

## Evaluating Whisper's Generalizability Through Zero-Shot Performance

This section investigates Whisper's ability to generalize across diverse domains, tasks, and languages without requiring dataset-specific fine-tuning. To assess this capability, a comprehensive evaluation is conducted on a broad range of existing speech processing datasets.

In a **zero-shot setting**, Whisper is applied to these datasets without leveraging any of their training data. This approach isolates Whisper's inherent generalizability, independent of dataset-specific biases or artifacts.





The key finding of the graph is that zero-shot Whisper models achieve a level of robustness that is closer to human performance than supervised LibriSpeech models. LibriSpeech dev-clean: Both supervised and zero-shot Whisper models perform well, with Whisper either matching or even outperforming a human.

**Combined dataset:** Supervised LibriSpeech models make significantly more errors (roughly double) compared to a human on this dataset. This indicates that these models are not robust and struggle to generalize to unseen data.

**Zero-shot Whisper models:** In contrast, zero-shot Whisper models perform much better on the combined dataset. Their WER falls within the 95% confidence interval of human performance, which suggests a high degree of robustness and generalizability.

Overall, the graph demonstrates that by training on a massive dataset of diverse audio data, Whisper models are able to achieve human-level robustness without the need for fine-tuning on specific tasks or datasets.

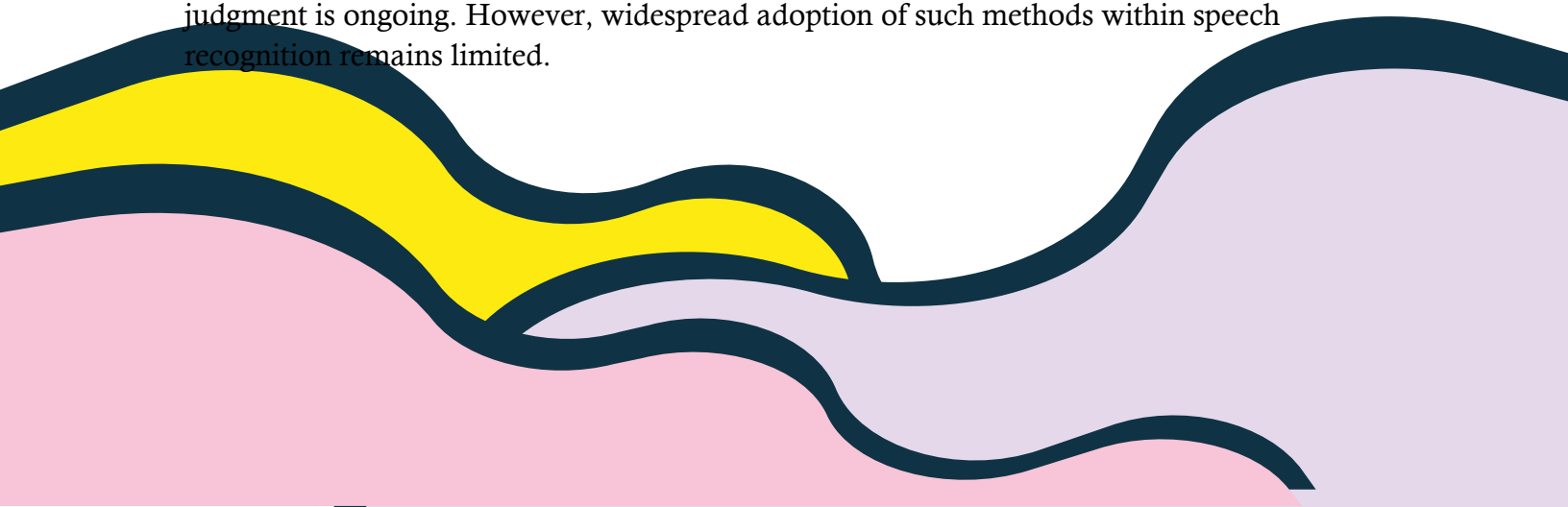
### **Mitigating the Impact of WER Limitations in Zero-Shot Speech Recognition**

#### **Evaluation:**

The below section discusses the limitations of Word Error Rate (WER) as the primary evaluation metric for speech recognition, particularly for zero-shot models like Whisper.

WER, based on string edit distance, penalizes all deviations between a model's output and the reference transcript, including inconsequential stylistic differences. Consequently, systems producing human-perceived correct transcripts can still incur high WER due to minor formatting variations. While this poses a challenge for all automated transcription, it's especially critical for zero-shot models that haven't been exposed to specific dataset formatting styles.

This is a recognized issue, and research into evaluation metrics that better align with human judgment is ongoing. However, widespread adoption of such methods within speech recognition remains limited.

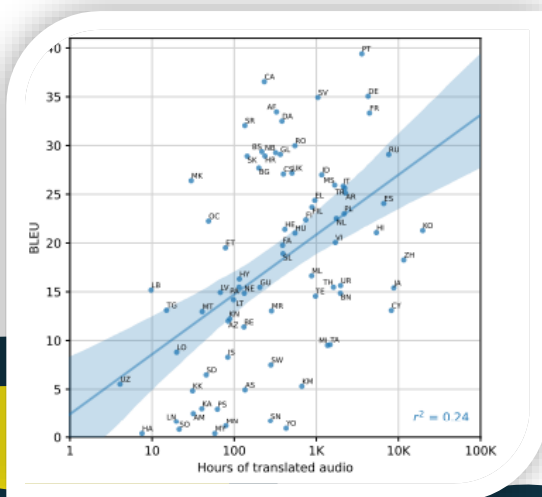
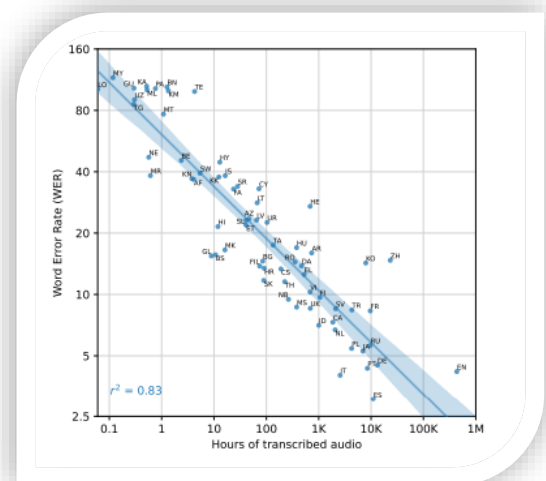




To address this challenge, they implemented a comprehensive text normalization process prior to WER calculation. This normalization minimizes penalties for non-semantic discrepancies. The text normalizer itself was developed through a process of iterative manual inspection, allowing us to identify common patterns where a naive WER approach would penalize Whisper models for inconsequential variations.

## Impact of Pre-training Data on Performance

**Zero-Shot Speech Recognition:** Analyses the correlation between the amount of pre-training data and a model's zero-shot speech recognition performance on the Fleurs benchmark. The findings show a strong positive correlation, indicating that models trained on larger datasets tend to perform better in zero-shot settings where they haven't been specifically fine-tuned for the task.



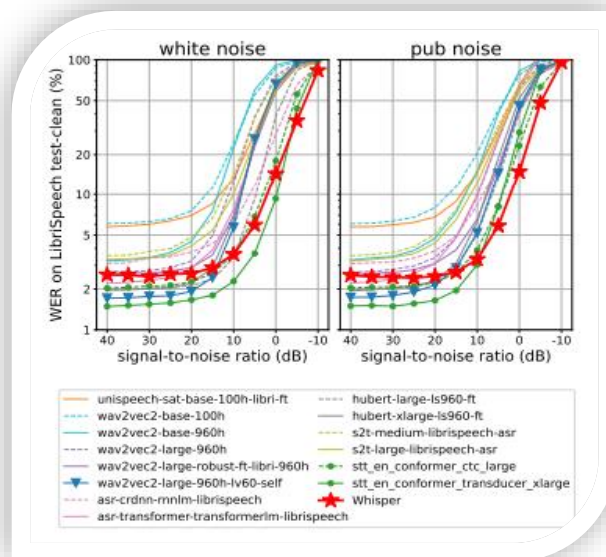
## Zero-Shot Translation Performance:

Here, the focus shifts to the impact of pre-training data on Whisper's zero-shot translation performance in Fleurs. In contrast to speech recognition, the correlation is found to be moderate. This suggests that the amount of pre-training translation data may not be as critical a factor for Whisper's zero-shot translation capabilities.



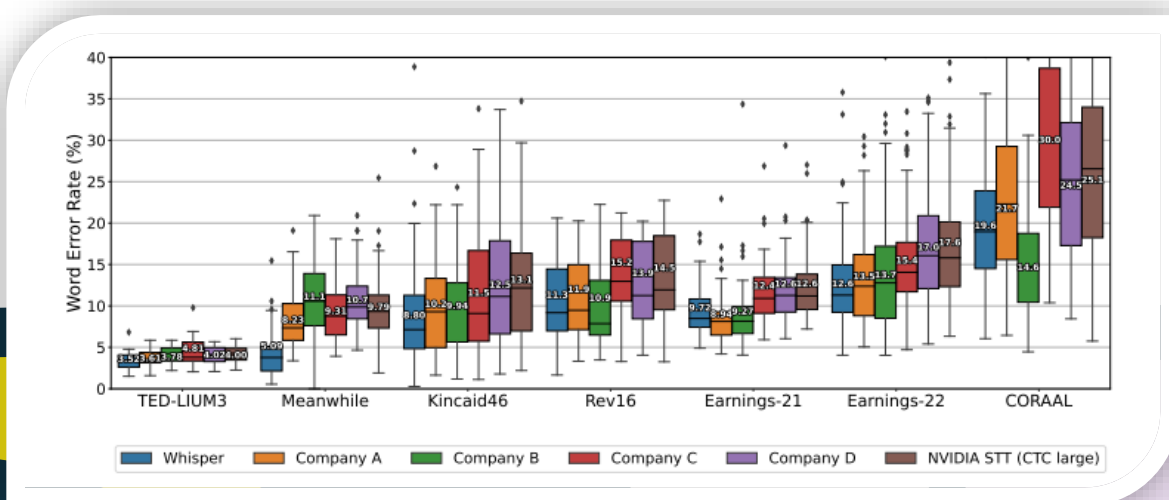
## Noise Robustness Compared to Other Models

**LibriSpeech Performance under Noise:** This section compares Whisper's performance to other Automatic Speech Recognition (ASR) systems under varying noise conditions (additive white noise and pub noise) using the LibriSpeech test-clean dataset. The results demonstrate that Whisper's accuracy degrades slower than other models, including state-of-the-art commercial and open-source systems like NVIDIA STT. This indicates Whisper's superior robustness to noise.



## Performance in Long-Form Transcription

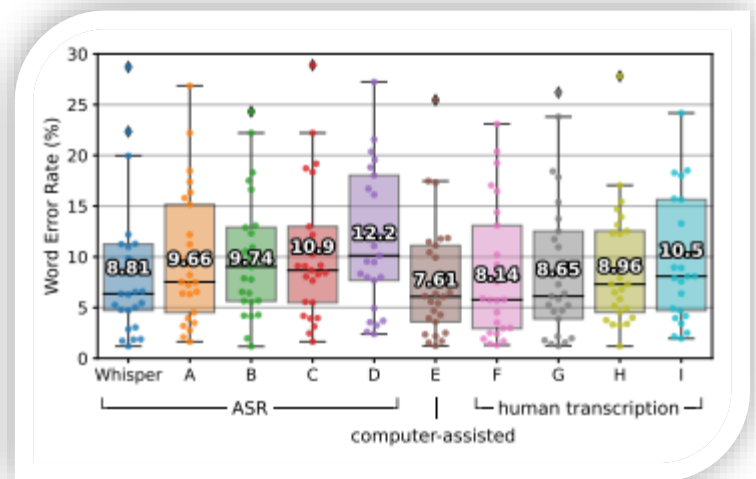
**Comparison with Other ASR Systems:** Here, the focus is on Whisper's performance in transcribing long-form audio (minutes to hours) compared to other ASR systems. The analysis involves comparing the distribution of word error rates (WER) on seven long-form



datasets. Whisper outperforms the best open-source model (NVIDIA STT) on all datasets and often surpasses commercial ASR systems as well.

### Comparison with Human Transcribers

**Human-Level Performance:** This section delves into Whisper's accuracy compared to professional human transcribers. The analysis utilizes the Kincaid46 dataset and compares WER distributions. The results show that Whisper's performance is close to that of human transcribers, including both computer-assisted and non-assisted services.



Despite the absence of further fine-tuning, Whisper achieves surprisingly strong performance, rivaling fully-supervised systems in some instances. The authors propose a system trained on a massive amount of weakly supervised data, and argue that this approach is more effective than supervised learning. Their system achieves high accuracy on standard benchmarks without any fine-tuning, and can also perform speech translation in multiple languages. The paper finds that the amount of training data significantly improves performance. In specific scenarios, its accuracy and robustness in handling diverse audio conditions even approach human-level capabilities.

Reference: <https://openai.com/blog/whisper>, <https://github.com/openai/whisper>,  
<https://huggingface.co/openai/whisper-large-v3>,  
<https://huggingface.co/spaces/openai/whisper>.