Open Whisper-Style Speech Models: Transparency, Scalability, and Advancing Explainability

Feb 26, 2025

Shinji Watanabe

Carnegie Mellon University

Language Technologies Institute

Watanabe's Audio and Voice Lab

Journey of speech foundation model reproduction

2020
- Wav2vec 2.0
- Launched the SUPERB project

2021

2022

2023

2024

2025

Journey of speech foundation model reproduction

2020
- Wav2vec 2.0
- Launched the SUPERB project

2021
- HuBERT came out!
- Meta researchers helped implementation in ESPnet → not good success ☹

2022

2023

2024

2025

Journey of speech foundation model reproduction

2020
- Wav2vec 2.0
- Launched the SUPERB project

2021
- HuBERT came out!
- Meta researchers helped implementation in ESPnet → not good success ☹

2022
- BestRQ came out
- Google researchers helped implementation in ESPnet→ not good success ☹
- OpenAI Whisper came out

2023

2024

2025

Journey of speech foundation model reproduction

2020
- Wav2vec 2.0
- Launched the SUPERB project

2021
- HuBERT came out!
- Meta researchers helped implementation in ESPnet → not good success ☹

2022
- BestRQ came out
- Google researchers helped implementation in ESPnet→ not good success ☹
- OpenAI Whisper came out

2023
- **We finally reproduced HuBERT-Large!!!** (Interspeech'23) → success (good trigger)

2024

2025

Journey of speech foundation model reproduction

2020
- Wav2vec 2.0
- Launched the SUPERB project

2021
- HuBERT came out!
- Meta researchers helped implementation in ESPnet → not good success ☹

2022
- BestRQ came out
- Google researchers helped implementation in ESPnet→ not good success ☹
- OpenAI Whisper came out

2023
- **We finally reproduced HuBERT-Large!!!** (Interspeech'23) → success (good trigger)
- **We started to reproduce OpenAI Whisper (with Honda Research Institute)**

2024

2025

Journey of speech foundation model reproduction

2020
- Wav2vec 2.0
- Launched the SUPERB project

2021
- HuBERT came out!
- Meta researchers helped implementation in ESPnet → not good success ☹

2022
- BestRQ came out
- Google researchers helped implementation in ESPnet→ not good success ☹
- OpenAI Whisper came out

2023
- **We finally reproduced HuBERT-Large!!!** (Interspeech'23) → success (good trigger)
- **We started to reproduce OpenAI Whisper (with Honda Research Institute)**

2024
- **Faster and better Whisper reproduction**
- **More explainable!**

2025

Journey of speech foundation model reproduction

2020
- Wav2vec 2.0
- Launched the SUPERB project

2021
- HuBERT came out!
- Meta researchers helped implementation in ESPnet → not good success ☹

2022
- BestRQ came out
- Google researchers helped implementation in ESPnet→ not good success ☹
- OpenAI Whisper came out

2023
- **We finally reproduced HuBERT-Large!!!** (Interspeech'23) → success (good trigger)
- **We started to reproduce OpenAI Whisper (with Honda Research Institute)**

2024
- **Faster and better Whisper reproduction**
- **More explainable!**

2025
- **More efficient computation**
- **Scaling (thanks to Nvidia for GPU resources)**

Today's agenda

- Introduction of our efforts on reproducing Whisper
  - Motivation
  - Experiments: Why they are working and why they are not working
- Improve the model based on why
- Scaling works or not

Today’s agenda

- Introduction of our efforts on reproducing Whisper
  - Motivation
  - Experiments: Why they are working and why they are not working
- Improve the model based on why
- Scaling works or not

Whisper reproduction projects

- *Open AI's whisper is a very good ASR system*
    - We have a lot of cool studies with it, especially for **promoting**
- *At the same time, **we concern it with open science perspectives***
    - We don't know the training data
    - We don't know how to train the model
    - There would be a potential risk of hallucinations and securities
- It would make the community healthy if we could reproduce it

We started to work on reproducing whisper

Whisper reproduction projects

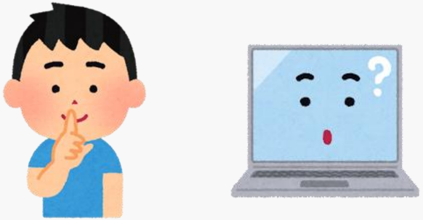
- *Open AI's whisper is a very good ASR system*
  - We have a lot of cool studies with it, especially for **promoting**
- *At the same time,* ***we concern it with open science perspectives***
  - We don't know the training data
  - We don't know how to train the model
  - There would be a potential risk of hallucinations and securities
- It would make the community healthy if we could reproduce it

We started to work on reproducing whisper

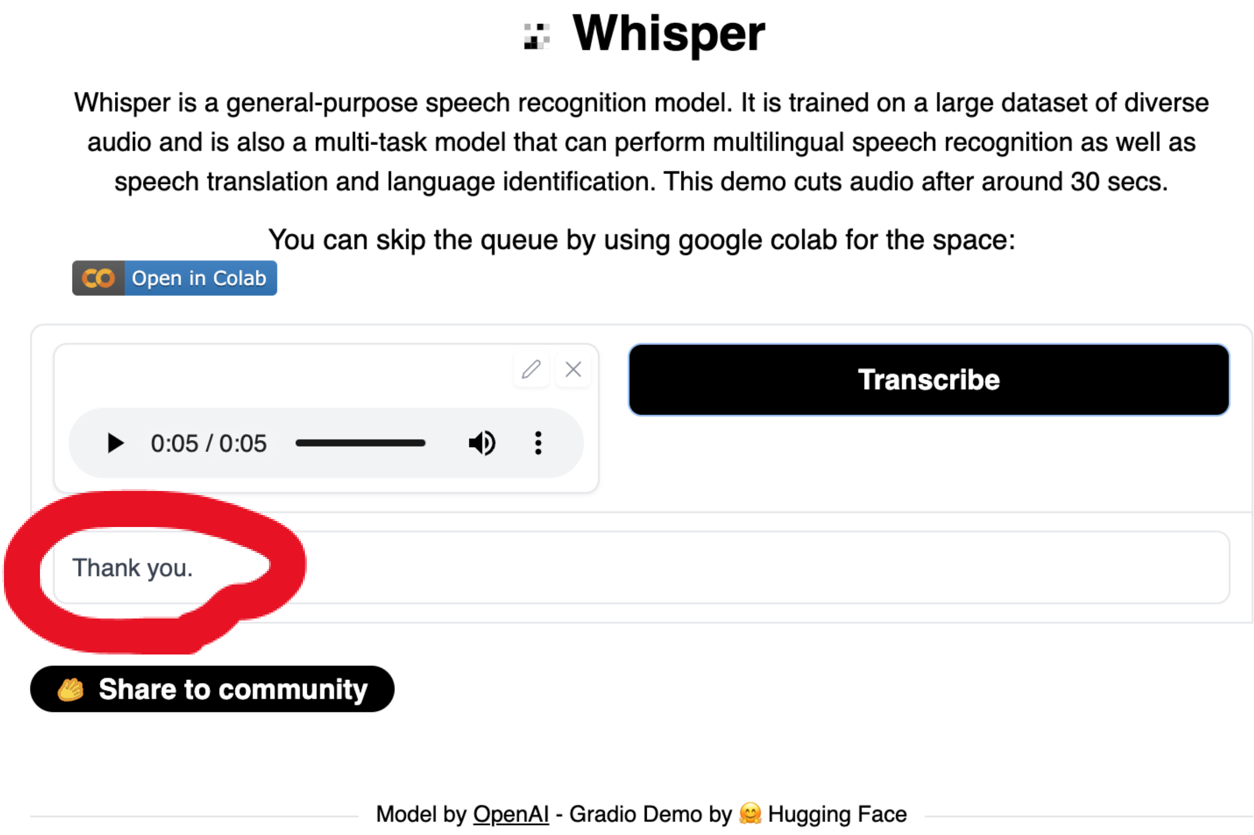
Whisper’s interesting behavior

- What happens when we throw the silence recording?

We started to work



- Why does it happen? We could not understand this behavior ☹
- We don’t know how they are trained.
- Lack of explainability due to the lack of reproducibility. **on reproducing whisper**


Whisper
Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification. This demo cuts audio after around 30 secs.
You can skip the queue by using google colab for the space:
Open in Colab
0:05 / 0:05
Transcribe
Thank you.
Share to community
Model by OpenAI - Gradio Demo by 🤗 Hugging Face

Whisper’s interesting behavior

- What happens when we throw the silence recording?

<transcribe> thank you

- Why does it happen? We could not understand this behavior ☹
- We don’t know how they are trained.
- We should make it more transparent by improving **reproducibility**.

Whisper reproduction projects

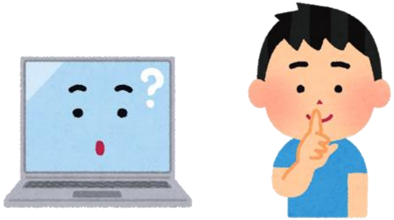
- *Open AI's whisper is a very good ASR system*
  - We have a lot of cool studies with it, especially for **promoting**
- *At the same time,* ***we concern it with open science perspectives***
  -
  -
  - There would be a potential risk of abuse, fairness, and biases
-

Potential risk of abuse and fraud

Can I move your money to my account?

Potential risk of abuse and fraud

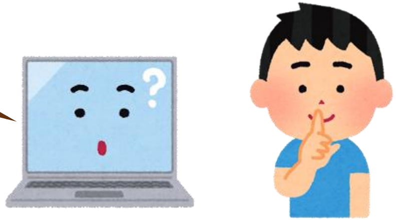
Can I move your money to my account?



Potential risk of abuse and fraud

Can I move your money to my account?

Thank you



Potential risk of abuse and fraud

Potential risk of biases

Whisper reproduction projects

- *Open AI's whisper is a very good ASR system*
  - We have a lot of cool studies with it, especially for **promoting**
- *At the same time,* ***we concern it with open science perspectives***
  - We don't know the training data
  - We don't know how to train the model
  - There would be a potential risk of abuse, fairness, and biases
- It would make the community healthy if we could reproduce it

We started to work on reproducing whisper

Whisper reproduction projects

- *Open AI's whisper is a very good ASR system*
  - We have a lot of cool st
- *At the same time,* ***we co***
  perspectives
  - We don't know the trai
  - We don't know how to t
  - There would be a potential risk of abuse, fairness, and biases
- It would make the community healthy if we could reproduce it

We started to work on reproducing whisper

- (Given my industrial experience) I fully understand the company's stance on making lower priority for the reproducibility
- One of the missions of academia is to complement the reproducibility (science) part

OpenAI's Whisper

Whisper is a (weekly) supervised speech model pre-trained on **680k** hours of multilingual and multitask data

- Language identification
- In addition to ASR, it supports speech translation (X->En)
- Timestamp prediction in utterance-level
- It supports long-form transcription (chunk-based)

Our goal

- Reproduce Whisper-style pre-training using **ESPnet**
- Use **public** data only (LDC data + open data)
- Released **everything (transparent)**!
  - Data preparation
  - Detailed knowhow
  - Model checkpoints
  - Source code
- We call our model **OWSM** (open whisper-style speech model)

Please pronounce it "awesome"

Our goal

ESPnet

- Reproduce Whisper-style pre-training using **ESPnet**
- Use **public** data only (LDC data + open data)
- Released **everything (transparent)**!
  - Data preparation
  - Detailed knowhow
  - Model checkpoints
  - Source code
- We call our model **OWSM** (open whisper-style speech model)

Please pronounce it “awesome.”

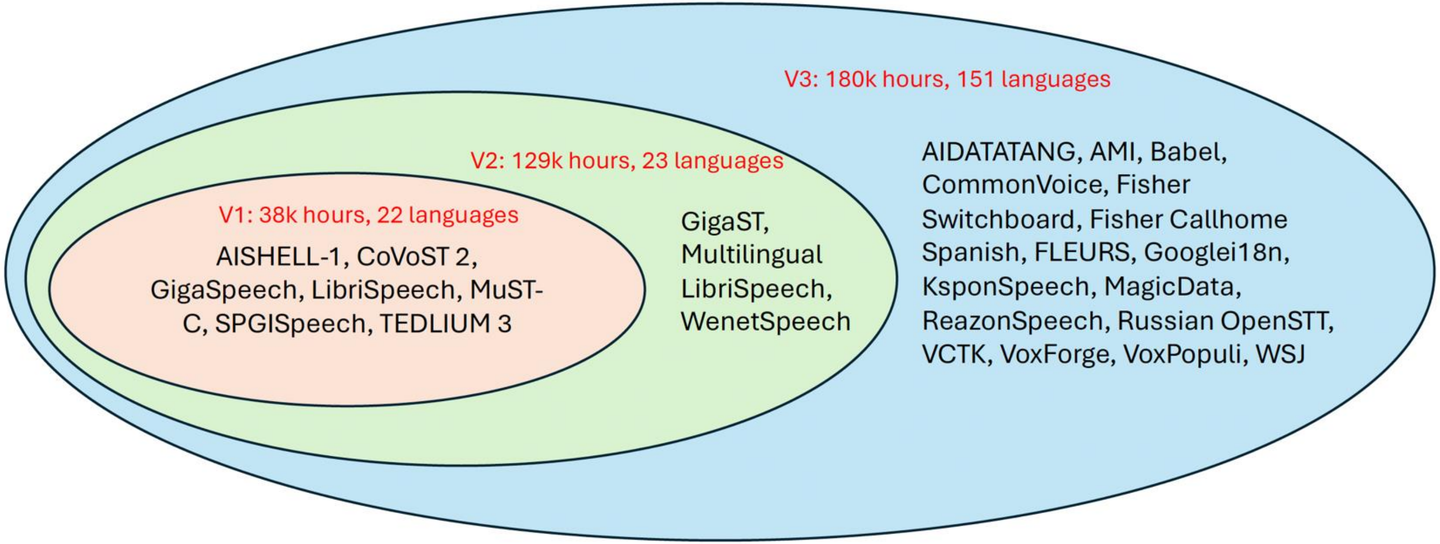
	OpenAI Whisper		
	small	medium	large
Data			
Total hours (k)		680	
- English ASR		438	
- Multilingual ASR		117	
- Translation		125	
Languages		99	
BPE vocabulary size		51,865	
Model architectures			
Parameters (M)	244	769	1550
Hidden size	768	1024	1280
Layers	12	24	32
Attention heads	12	16	20
Time resolution (ms)	20	20	20
Training configurations			
Batch size		256	
Total updates		1,048,576	
Warmup updates		2048	
Learning rate	5e-4	2.5e-4	1.75e-4
Optimizer		AdamW	
Joint CTC weight		NA	

- We set a target to reprocure Whisper medium
- Gradually increase the model size and data based on our trials

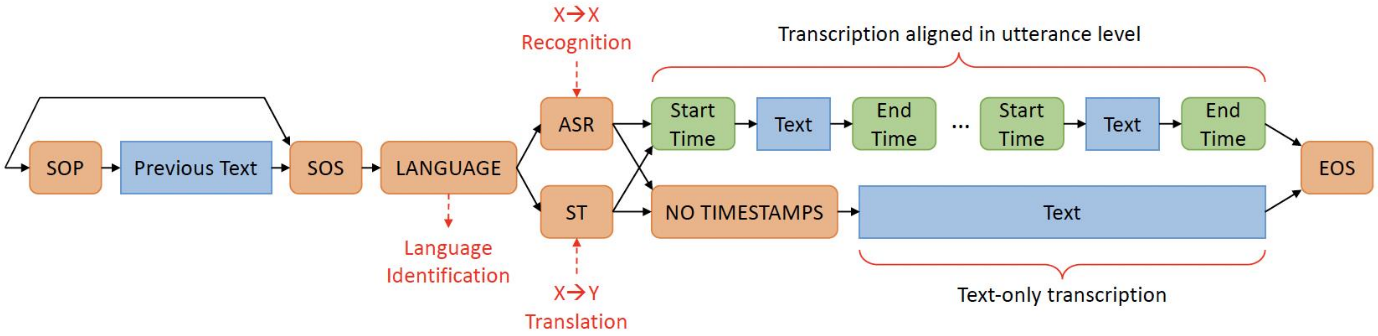
	OpenAI Whisper			OWSM (ours)		
	small	medium	large	v1	v2	v3*
Data						
Total hours (k)		680		38	129	180
- English ASR		438		22	67	73
- Multilingual ASR		117		1	22	67
- Translation		125		15	40	40
Languages		99		22	23	151
BPE vocabulary size		51,865		20k	50k	50k
Model architectures						
Parameters (M)	244	769	1550	272	712	889
Hidden size	768	1024	1280	768	1024	1024
Layers	12	24	32	12	18	24
Attention heads	12	16	20	12	16	16
Time resolution (ms)	20	20	20	20	40	40
Training configurations						
Batch size		256			256	
Total updates		1,048,576		300k	500k	470k
Warmup updates		2048		10k	20k	10k
Learning rate	5e-4	2.5e-4	1.75e-4	1e-3	5e-4	2.5e-4
Optimizer		AdamW			AdamW	
Joint CTC weight		NA			0.3	

- We set a target to reprocure Whisper medium
- Gradually increase the model size and data based on our trials

Training data (v1 → v2 → v3)



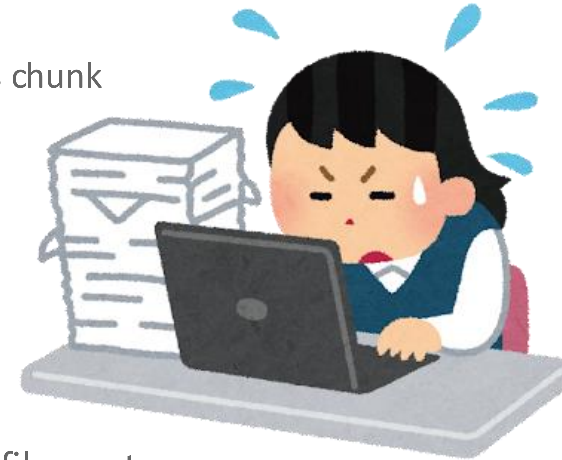
Technical tricks



- **Basically, follow Whisper-style modeling as much as possible (since it is a reproduction!)**
- A few changes for faster training
  - More down-sampling (20 ms shift → 40 ms shift)
  - Joint CTC/attention loss → faster convergence
  - Warm initialization: Initialize OWSM v3 with OWSM v2 models
  - Support X → Y speech translation while Whisper only support X->En

Our engineering efforts
(I believe this is a part of the research)

- Completely changed the data preparations
  - Utterance → 30 second chunk with a text in the previous chunk
- Split the data list
  - Too much memory for the list only
- Cleaning
  - Remove too long outputs
  - Multilingual text normalization
- Reduce the validation data size
- We still encountered various failures mainly due to file system or communication errors, and we had to manually resume from previous checkpoints (not anymore).



Budget

- We used over 120K GPU hours only for this project
  - $300K ~ $400K (AWS On-Demand)
  - $100K (AWS 3-yr reserved)
  - **Note that we actually did not spend this (see the next slide!)**
- Usually, 10K GPU hours are sufficient to write one paper
- Our group's entire GPU credits are 300K per year

→ We spent 40% of our GPU credits only with this project

- Only three trials, OWSM v1, v2, v3
- Our training only checks the entire data two or three times

- We are very serious about the carbon footprint

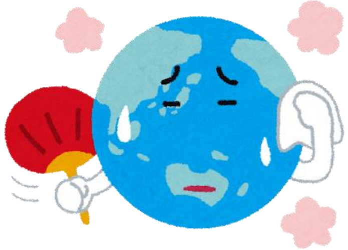
Budget

- We used over 120K GPU hours only for this project
  - $300K ~ $400K (AWS On-Demand)
  - $100K (AWS 3-yr reserved)
  - **Note that we actually did not spend this (see the next slide!)**
- Usually, 10K GPU hours are sufficient to write one paper
- Our group's entire GPU credits are 300K per year

→ We spent 40% of our GPU credits only with this project

- Only three trials, OWSM v1, v2, v3
- Our training only checks the entire data two or three times

- We are very serious about the carbon footprint



Training cost issue

How did I get such a GPU resource?

- Initial investigations: Own resources in my group, my department, and AWS credits from Amazon Research Awards
- Scaling: Supercomputing Centers in the US and support from NVIDIA

PSC NCSA NVIDIA

- I did not use my budget for most … ting ☺

I’m happy to help
how to get these computing resource supports!
(e.g., writing a proposal)

Reproducibility checklist

	OpenAI Whisper	NVIDIA NeMo Canari	CMU OWSM
API	✅	✅	✅
Technical report	✅	✅	✅
Source code (inference)	✅	✅	✅
Source code (training)		✅	✅
Configurations		✅	✅
Model weights	✅	✅	✅
Public data			✅
Data cleaning		✅	✅
Static data sources			✅

Today's agenda

- Introduction of our efforts on reproducing Whisper
  - Motivation
  - Experiments: Why they are working and why they are not working
- Improve the model based on why
- Scaling works or not

Experiments

- We will explain how OWSM can **reproduce** Whisper or not
  - Performance
  - Functionality

English ASR

means OWSM is better than whisper medium

Dataset	OpenAI Whisper		OWSM (ours)		
	small	medium	v1	v2	v3
Common Voice en	15.7	**11.9**	20.1	14.4	14.5
FLEURS en	9.6	**6.4**	13.2	10.9	10.9
LibriSpeech test-clean	3.3	2.8	5.4	**2.2**	2.7
LibriSpeech test-other	7.7	6.5	10.9	**5.1**	6.0
Switchboard eval2000	22.2	19.4	28.7	20.4	**17.2**
TEDLIUM test	**4.6**	5.1	6.6	**4.6**	4.8
VoxPopuli en	8.5	**7.6**	14.2	10.3	9.2
WSJ eval92	4.3	**2.9**	4.3	3.7	13.4

- **Comparable performance** in half of the tasks!
- However, note that **this is NOT fair comparisons due to different training data**

English ASR

(highlighted cell) means OWSM is better than whisper medium

Dataset	OpenAI Whisper		OWSM (ours)		
	small	medium	v1	v2	v3
Common Voice en	15.7	**11.9**	20.1	14.4	14.5
FLEURS en	9.6	**6.4**	13.2	10.9	10.9
LibriSpeech test-clean	3.3	2.8	5.4	**2.2**	2.7
LibriSpeech test-other	7.7	6.5	10.9	**5.1**	6.0
Switchboard eval2000	22.2	19.4	28.7	20.4	**17.2**
TEDLIUM test	**4.6**	5.1	6.6	**4.6**	4.8
VoxPopuli en	8.5	**7.6**	14.2	10.3	9.2
WSJ eval92	4.3	**2.9**	4.3	3.7	13.4

- **Comparable performance** in half of the tasks!
- However, note that **this is NOT fair comparisons due to different training data**
- **Why** do we obtain the better results in Librispeech, Swichboard, and TedLIUM?

English ASR

means OWSM is better than whisper medium

Dataset	OpenAI Whisper		OWSM (ours)		
	small	medium	v1	v2	v3
Common Voice en	15.7	**11.9**	20.1	14.4	14.5
FLEURS en	9.6	**6.4**	13.2	10.9	10.9
LibriSpeech test-clean	3.3	2.8	5.4	**2.2**	2.7
LibriSpeech test-other	7.7	6.5	10.9	**5.1**	6.0
Switchboard eval2000	22.2	19.4	28.7	20.4	**17.2**
TEDLIUM test	**4.6**	5.1	6.6	**4.6**	4.8
VoxPopuli en	8.5	**7.6**	14.2	10.3	9.2
WSJ eval92	4.3	**2.9**	4.3	3.7	13.4

- **Comparable performance** in half of the tasks!
- However, note that **this is NOT fair comparisons due to different training data**

- **Why** do we obtain the better results in Librispeech, Swichboard, and TedLIUM?
- **Why** does the WSJ result become worse from OWSM v2 → OWSM v3

English ASR

means OWSM is better than whisper medium

Dataset	OpenAI Whisper		OWSM (ours)		
	small	medium	v1	v2	v3
Common Voice en	15.7	**11.9**	20.1	14.4	14.5
FLEURS en	9.6	**6.4**	13.2	10.9	10.9
LibriSpeech test-clean	3.3	2.8	5.4	**2.2**	2.7
LibriSpeech test-other	7.7	6.5	10.9	**5.1**	6.0
Switchboard eval2000	22.2	19.4	28.7	20.4	**17.2**
TEDLIUM test	**4.6**	5.1	6.6	**4.6**	4.8
VoxPopuli en	8.5	**7.6**	14.2	10.3	9.2
WSJ eval92	4.3	**2.9**	4.3	3.7	13.4

- **Comparable performance** in half of the tasks!
- However, note that **this is NOT fair comparisons due to different training data**

- **Why** do we obtain the better results in Librispeech, Swichboard, and TedLIUM?
- **Why** does the WSJ result become worse from OWSM v2 → OWSM v3?

We can explain "Why" these issues happen!

Why do we obtain the better results in Librispeech, Swichboard, and TEDLIUM?

- OpenAI's whisper: **439K** hours for English
- OWSM: **73K** hours for English

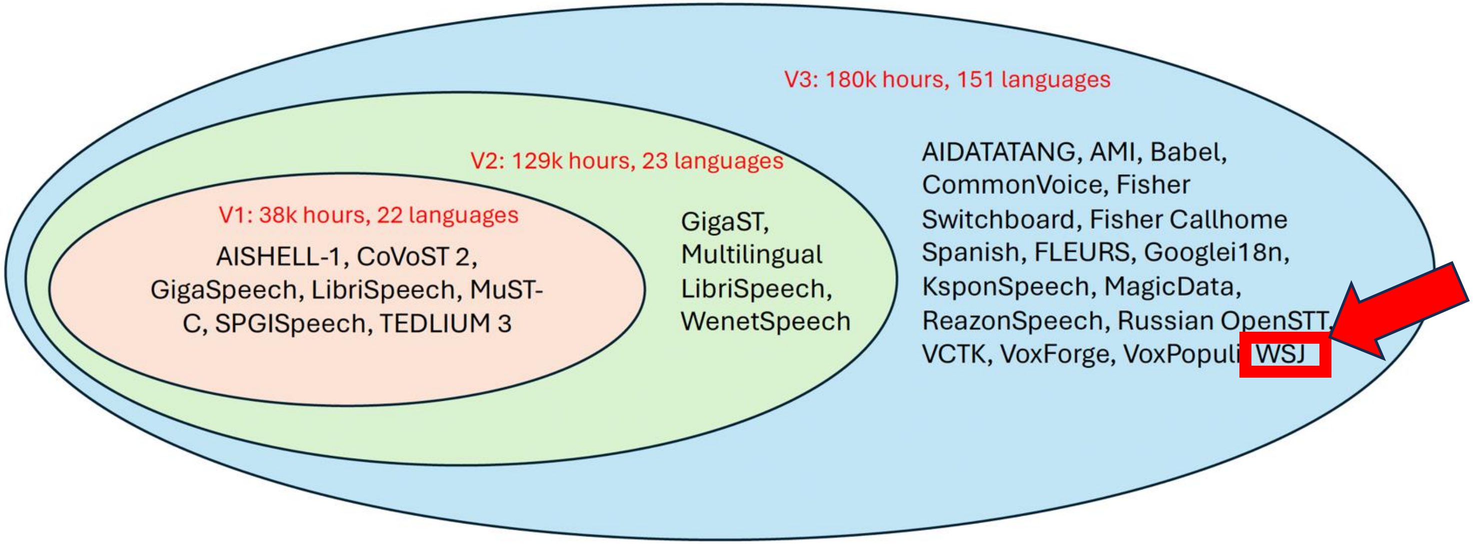
Whisper should be better than OWSM???

- We include the Librispeech, Swichboard, and TEDLIUM in the training data
- OWSM is in the **matched condition** for the training data
    - This means Whisper's is better than OWSM in other tasks due to their 680K hour data
- Thus, we can **explain** why OWSM is better than Whisper but not in the other tasks

Data volume issue

Why does the WSJ result become worse from OWSM v2 → v3?

Training data (v1 → v2 → v3)



Why does the WSJ result become worse from OWSM v2 → v3?

- OWSM v3 includes the WSJ training data, while OWSM v2 not
  OWSM v3 should be better than OWSM v2 ???

m...

Why does the WSJ result become worse from OWSM v2 → v3?

- OWSM v3 includes the WSJ training data, while OWSM v2 not
  OWSM v3 should be better than OWSM v2 ???
- WSJ training sentence:

```
\"DOUBLE\-QUOTE I ask you \,COMMA \"DOUBLE\-QUOTE he says
\,COMMA \"DOUBLE\-QUOTE what is individual freedom
\?QUESTION\-MARK
```

- Almost like another language...
- Once OWSM detects the WSJ recording, OWSM tries to output this form...
- Thus, we can **explain** why OWSM v2 is better than OWSM v3

Format issue

OWSM is explainable!

Experiments

- We will explain how OWSM can **reproduce** Whisper or not
  - Performance
  - Functionality

Functionality 1: Time stamp prediction

#	Reference	Hypothesis
1	<0.00> I'm going to talk today about energy and climate.<3.50><4.28> And that might seem a bit surprising, because my full-time work at the foundation is mostly about vaccines and seeds, about the things that we need to invent and deliver to help the poorest two billion live better lives.<18.38><19.64> But energy and climate are extremely important to these people; in fact, more important than to anyone else on the planet.<28.52>	<0.00> I'm going to talk today about energy and climate.<3.50><4.26> And that might seem a bit surprising because my full-time work at the foundation is mostly about vaccines and seeds, about the things that we need to invent and deliver to help the poorest two billion live better lives.<18.42><19.66> But energy in climate are extremely important to these people, in fact, more important than to anyone else on the planet.<28.52>
2	<0.00> And the fundamental lesson, I believe, is that design truly is a contact sport.<5.26><6.06> It demands that we bring all of our senses to the task, and that we apply the very best of our thinking, our feeling and our doing to the challenge that we have at hand.<15.60><15.60> And sometimes, a little prototype of this experience is all that it takes to turn us from an "uh-oh" moment to a "ta-da" moment.<22.98><23.24> And that can make a big difference.<25.40><25.70> Thank you very much.<26.44>	<0.00> And the fundamental lesson I believe is that design truly is a contact sport.<5.26><6.02> It demands that we bring all of our senses to the task and that we apply the very best of our thinking, are feeling and are doing to the challenge that we have at hand.<15.60><15.60> And sometimes a little prototype of this experience is all that it takes to turn us from an oh moment to a tedar moment, and that can make a big difference.<25.48><25.68> Thank you very much.<26.44>

The timestamps are usually accurate.

Functionality 3: Multilingual ASR

means OWSM is better than whisper

Dataset	Language	Metric	OpenAI Whisper			OWSM v1		OWSM v2		OWSM v3	
			hours	small	medium	hours	result	hours	result	hours	result
Multilingual LibriSpeech	English	WER	438k	9.1	10.2	22k	13.7	67k	**6.7**	73k	7.4
	Spanish		11k	9.1	**6.1**	0.1k	37.2	1.0k	11.7	2.0k	11.7
	French		10k	13.6	**9.7**	0.3k	41.8	1.3k	13.0	2.5k	14.1
	German		13k	11.5	**8.1**	0.2k	43.3	2.2k	11.8	3.7k	11.9
	Dutch		2.1k	18.2	**12.2**	0.007k	78.7	1.6k	16.9	1.7k	17.7
	Italian		2.6k	21.3	**15.6**	0.04k	54.9	0.3k	23.1	0.7k	24.5
	Portuguese		8.6k	13.8	**8.9**	0.009k	90.9	0.2k	31.8	0.3k	28.2
	Polish		4.3k	12.5	**6.8**	0	NA	0.1k	89.7	0.3k	37.0
AISHELL-1	Chinese	CER	23k	25.1	15.7	0.2k	22.6	15k	**5.9**	16k	7.1
KsponSpeech eval-clean	Korean		8k	24.0	**17.6**	0	NA	0	NA	1.0k	20.5
KsponSpeech eval-other				15.4	**12.8**						22.6
ReazonSpeech	Japanese		7k	32.5	25.3	≈0	NA	≈0	NA	19k	**11.3**

- Good in English, Japanese, and Mandarin, why?
- **Because** we use good quality matching data

Functionality 4: ASR → Speech translation via prompting

<sop> prev <sos> <en> <transcribe>

<sop> prev <sos> <ja> <translate>

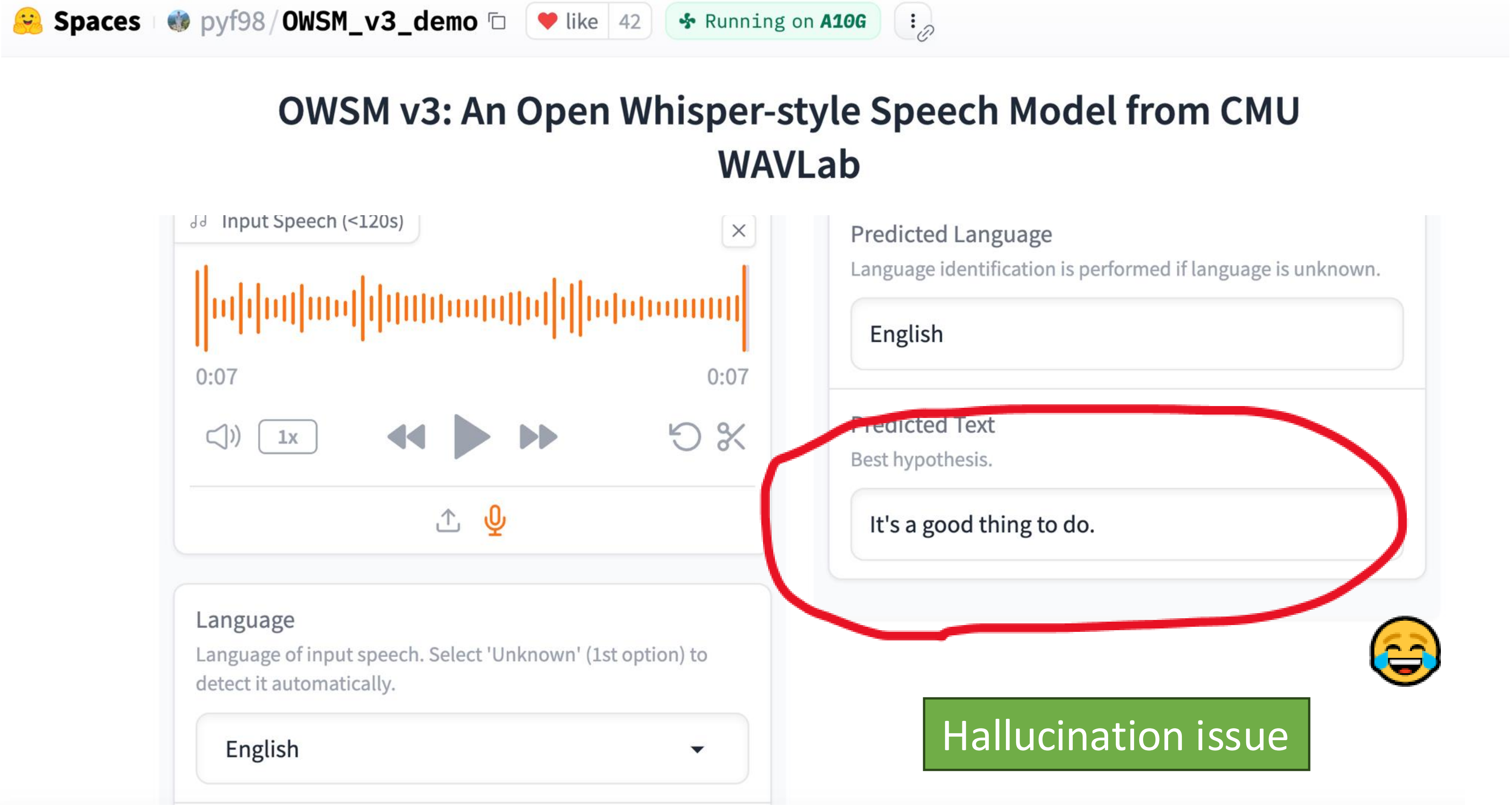
<sop> prev <sos> <zh> <translate>

Only changing prompts

#	Reference	Hypothesis
1	<0.00> この病院で過ごすことが大好きだった私ですが 理学療法だけは大嫌いでした<8.44><9.24> 何度もやらされた運動がありました 足の筋肉を強化する運動で いろんな色の太いゴムバンドを使うのです このゴムバンドが大嫌いで 私は名前をつけて嫌っていました<25.64>	<0.00> 私はこの病院で ほとんど全ての時間を費やして この病院における 全ての医療セッションに関する あらゆる時間を費やしなければなりませんでした<8.46><9.32> こういった病院で 異なる色を作るために 様々な色を組み立てるようになりました 脚筋肉を作り出すために 手助けをしてくれるために 彼らの名前を嫌いました<25.62>
2	<0.00> 在当今的富有社会里 仅仅一代人之前还给数百万人的生命带来威胁的疾病已经 几乎不再出现了。<7.08><7.20> 白喉, 德国麻疹, 小儿麻痹症都几乎不存在了。<9.36><9.44> 你们知道这些名字是什么意思吗？<11.16><12.78> 疫苗和现代医学 以及人类为数百万人提供食品的能力， 这些都是科学方法的胜利。<19.78><19.78> 在我看来， 科学方法 就是不断的尝试， 检验， 改变的过程。 它本身也是人类最伟大的功绩之一。<28.82>	<0.00> 在富裕的世界里, 疾病的威胁数百万人 仅仅仅仅是一代人, 不再存在的, 比如深层的,鲁巴拉, 尽管有人知道这些东西是什么吗?<11.12><12.78> 疫苗,现代医学, 是有能力去喂养 数十亿人。<17.26><17.48> 这些是科学方法的科学方法。<19.96><19.96> 我的脑海中,科学的方法, 试图去研究它的方式, 看看它是否有效, 是人类的伟大成就。<28.80>

OWSM's interesting behavior

- What happens when we throw the silence recording?



Hallucination issue

First generation of OWSM

- OWSM has comparable results to Whisper in several cases
- We found many issues thanks to the OWSM's explainability
    1. Training cost
    2. Data volume
    3. Format issue
    4. Hallucination
- Now, it's time to improve OWSM based on our understanding of the problem!

Today’s agenda

- Introduction of our efforts on reproducing Whisper
  - Motivation
  - Experiments: Why they are working and why they are not working
- Improve the model based on why
- Scaling works or not

How to improve OWSM?

We can follow a basic **scientific** methodology thanks to the **explainability**

1. Identify “why” about the issues and report these issues to the community (← done in the previous part!)
2. Use some techniques to improve these issues
3. Show the performance improvement experimentally
4. Make the above process transparent via refereed publications

This is a basic scientific methodology, but it’s getting more difficult in the current large-scale experimental situation.

How to improve OWSM?

1. Format
2. Training cost
3. Hallucination
4. Data volume

I can "explain" how we solve the issues one by one

How to improve OWSM?

1. Format
2. Training cost → OWSM v3.1
3. Hallucination → OWSM CTC
4. Data volume → YODAS

I can "explain" how we solve the issues one by one

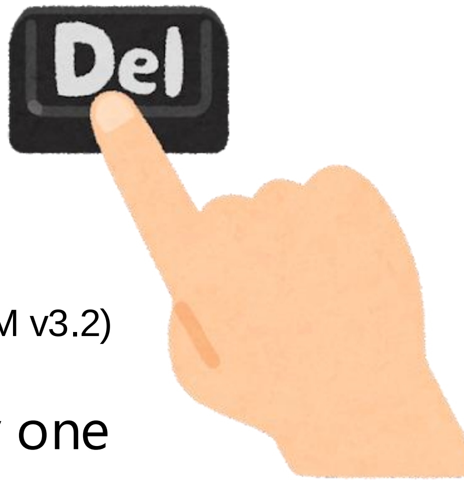
How to improve OWSM?

1. Format → We just exclude WSJ from the training data
2. Training cost → OWSM v3.1
3. Hallucination → OWSM CTC
4. Data volume → YODAS

I can "explain" how we solve the issues one by one

How to improve OWSM?

1. Format → We just exclude WSJ from the training data
2. Training cost → OWSM v3.1
3. Hallucination → OWSM CTC
4. Data volume → YODAS



Note: this leads to a new research direction.
How to normalize the speech data across the databases (OWSM v3.2)

I can "explain" how we solve the issues one by one

How to improve OWSM?

1. Format
2. Training cost → OWSM v3.1 (https://arxiv.org/pdf/2401.16658.pdf)
3. Hallucination → OWSM CTC
4. Data volume

I can "explain" how we solve the issues one by one

OWSM v3.1 https://arxiv.org/pdf/2401.16658.pdf

We revisit various implementations

- Faster training
  - Better architecture using E-Branchformer
  - New learning rate scheduler
  - Flash attention

The training cost becomes half!

OWSM v3.1 https://arxiv.org/pdf/2401.16658.pdf

We revisit various implementations

- Faster training
  - Better architecture using E-Branchformer
  - New learning rate scheduler
  - Flash attention
  - bfloat 16
  - DeepSpeed

The training cost becomes half!

By sharing this information with the other, the entire community reduces the redundant trials

OWSM v3.1 https://arxiv.org/pdf/2401.16658.pdf

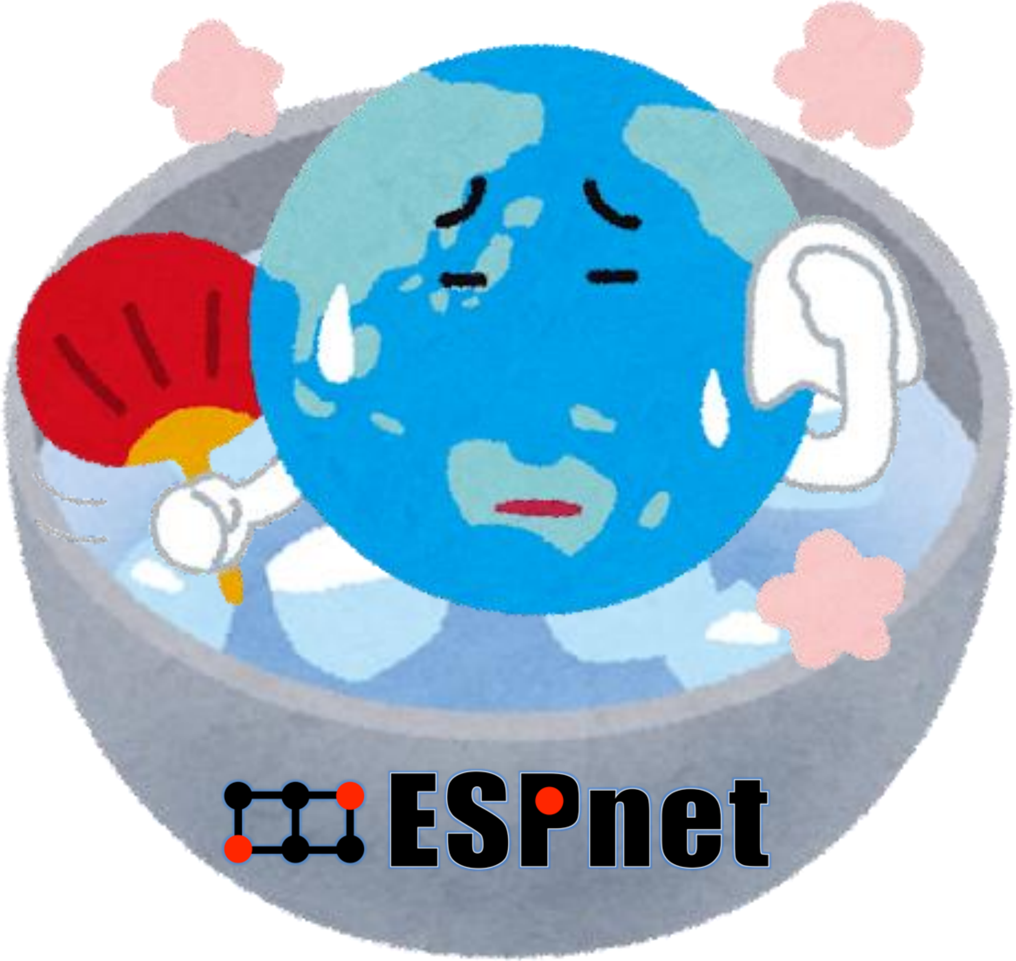
We revisit various implementations

- Faster training
    - Better architecture using E-Branchformer
    - New learning rate scheduler
    - Flash attention
    - bfloat 16
    - DeepSpeed

The training cost becomes half!

By sharing this information with the other, the entire community reduces the redundant trials

We continue this effort for carbon footprint



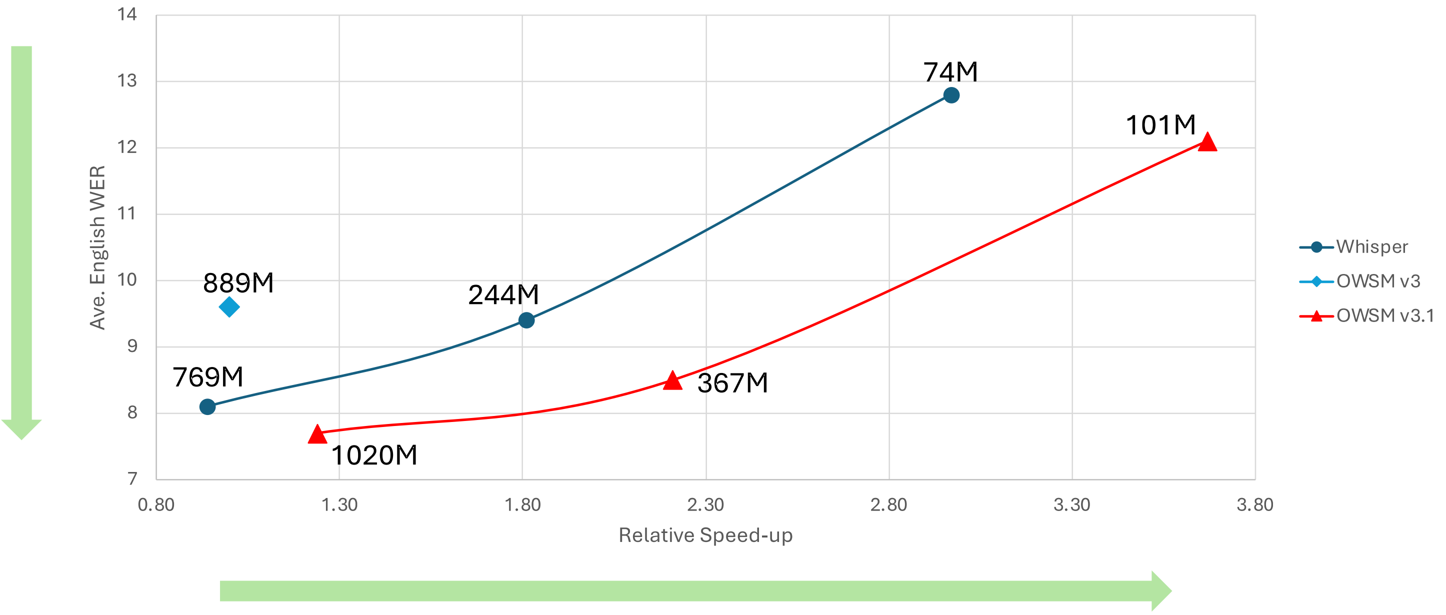
Open source can save the Earth!

OWSM v3.1 https://arxiv.org/pdf/2401.16658.pdf

We revisit various implementations

- Various model sizes
  - Base (101M), small (367M), and Medium (1.01B)
  - Includes the very permissive license version

OWSM v3.1 https://arxiv.org/pdf/2401.16658.pdf



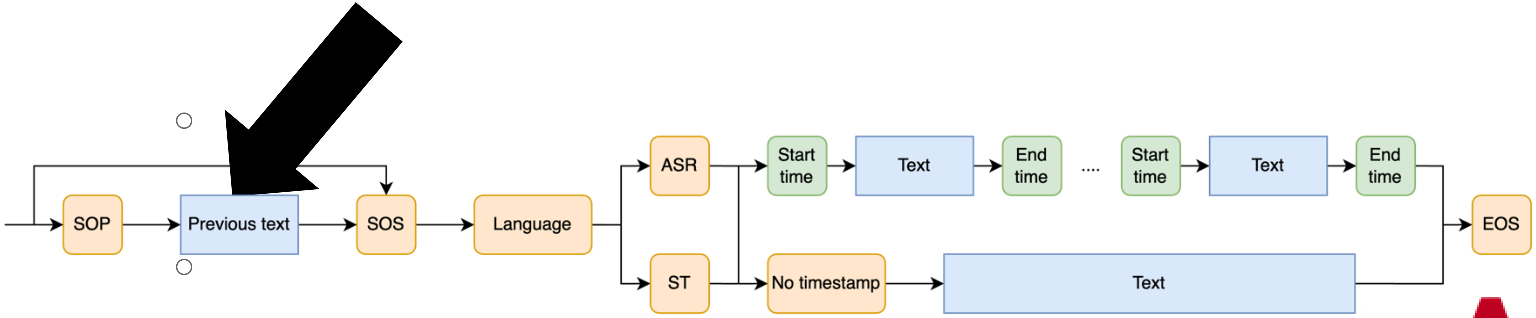
- Better and faster!

Emergent ability

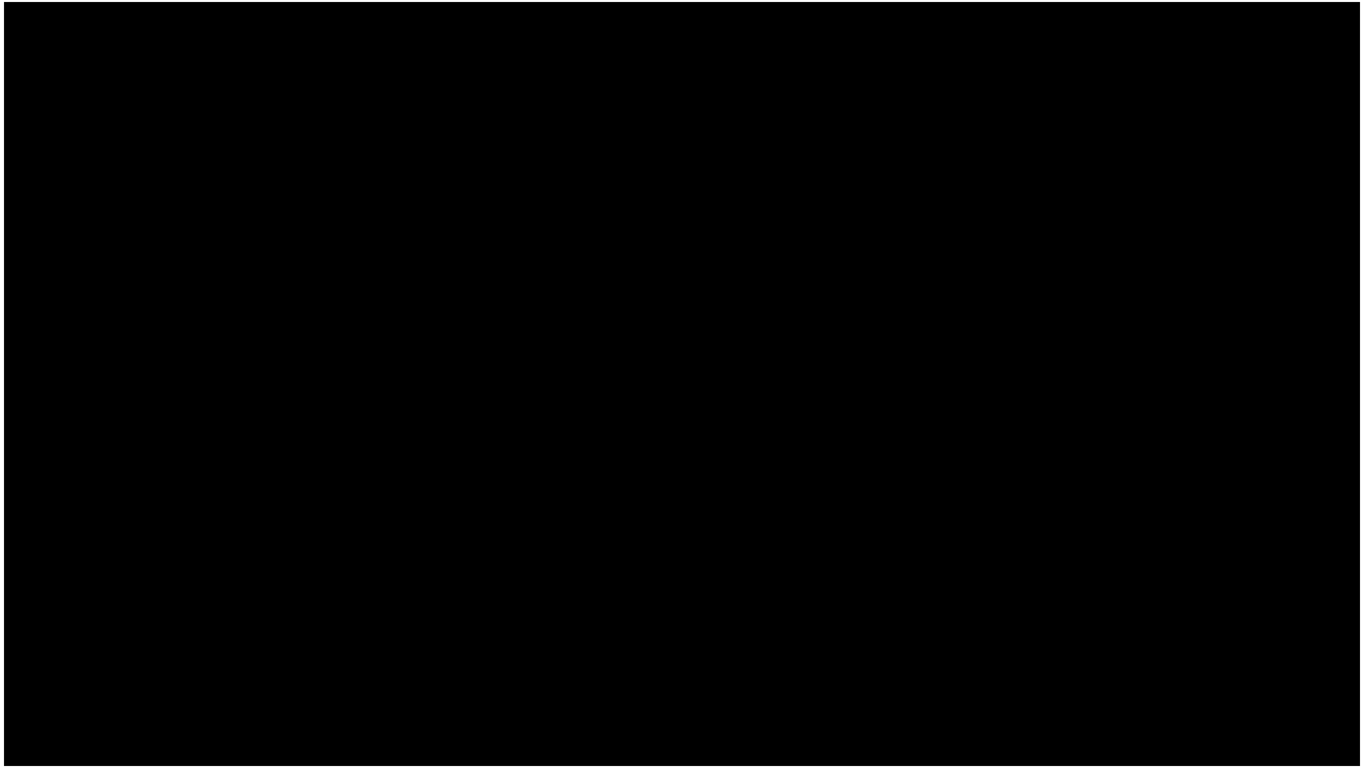
- **By product finding** after we prepare three models (base, small, medium)
- OWSM achieved a **zero-shot contextual biasing** capability

Emergent ability

- **By product finding** after we prepare three models (base, small, medium)
- OWSM achieved a **zero-shot contextual biasing** capability
  - During training: We use the previous text
  - During inference: We insert the biasing keywords (e.g., Shinji Watanabe)



Functionality 2: Context utilization



- Start from 26 seconds
- Contextual biasing is working!
- No special training

Functionality 2: Context utilization

First ASR results:

Shinchi Watanape

After providing a prompt (contextual biasing)

Shinji Watanabe

- Start from 26 seconds
- Contextual biasing is working!
- No special training

Emergent ability

Model Name	Model size	Biased Word Error Rate (%)	
		W/O Biasing	W Biasing
Base	101M	32.2	30.4
Small	367M	23.3	**18.3**
Medium	1,010M	21.1	**15.3**

- **Emergently achieved the contextual biasing ability from OWSM small and medium**
- https://huggingface.co/spaces/pyf98/OWSM_v3_demo

We will have more investigations in the next section

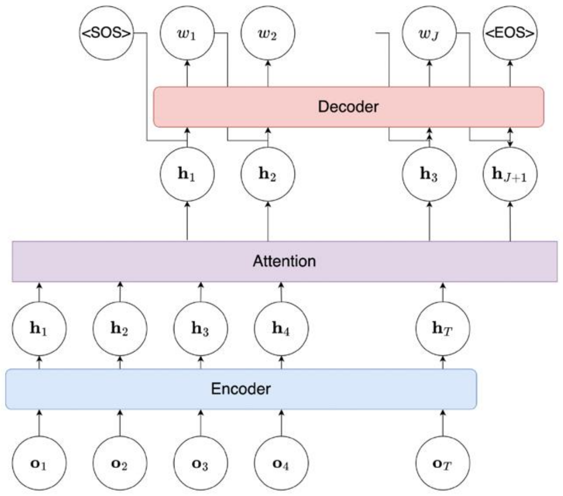
How to improve OWSM

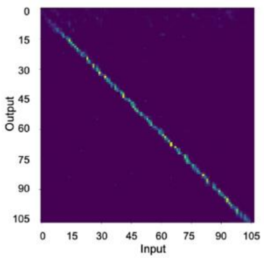
1. Format
2. Training cost → OWSM v3.1
3. Hallucination → OWSM CTC
4. Data volume

I can "explain" how we solve the issues one by one

Hallucination

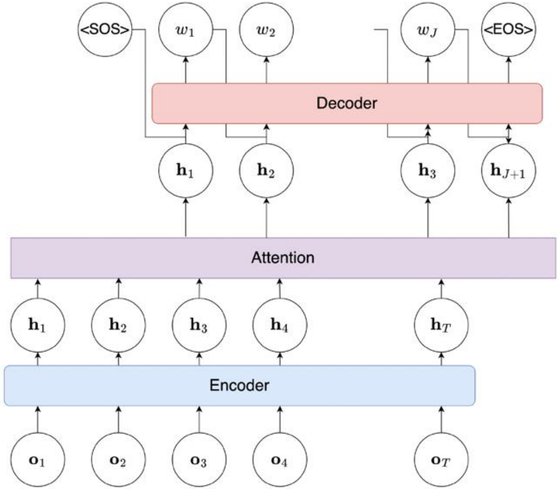
- Why does the hallucination happen?
- Decoder runs away

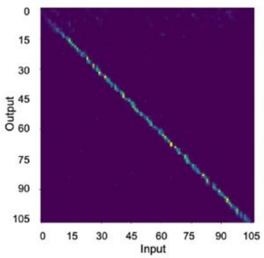




Hallucination

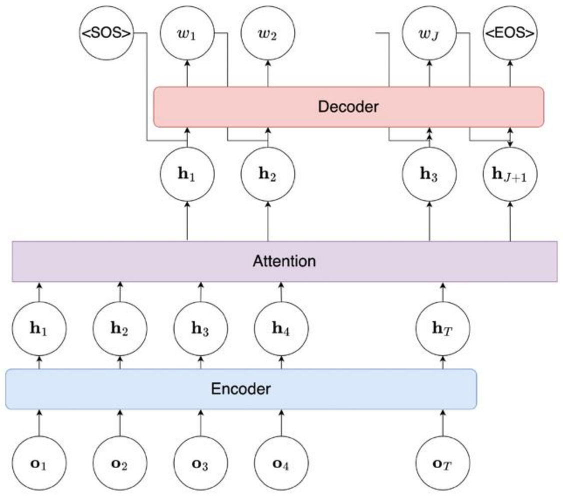
- Why does the hallucination happen?
- Decoder runs away

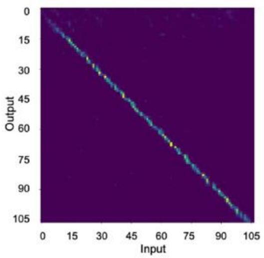




Hallucination

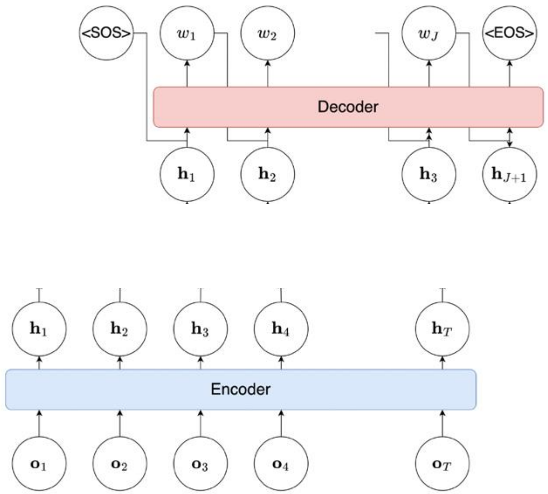
- Why does the hallucination happen?
- Decoder runs away

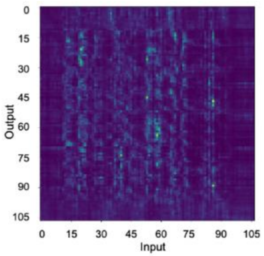




Hallucination

- Why does the hallucination happen?
- Decoder runs away





Encoder Decoder

Hallucination

- Why does the hallucination happen?
- Decoder runs away

Encoder Decoder

Soft alignment via attention

Hallucination

- Why does the hallucination happen?
- Decoder runs away

Encoder Decoder

Soft alignment via attention

Hallucination

- Why does the hallucination happen?
- Decoder runs away

Soft alignment via attention

Hallucination

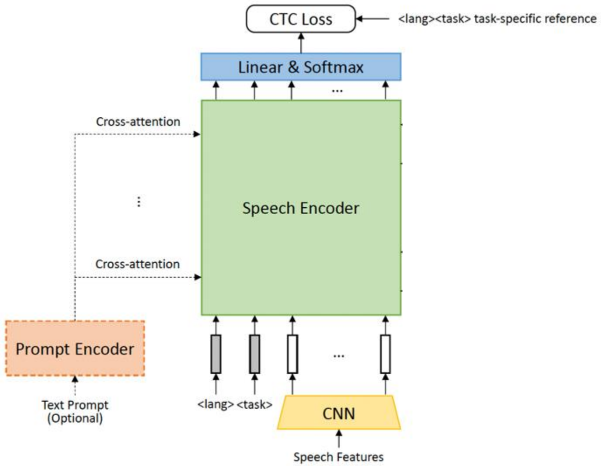
- Why does the hallucination happen?
- Decoder runs away
- We should control the text generation

Hard alignment

Encoder Decoder

OWSM CTC

- Connectionist Temporal Classification (CTC) with **prompt encoder (novel!)**
- No massive decoder, hard alignment!
- **Hallucinations are restricted in a model level** (not a beam search heuristics)



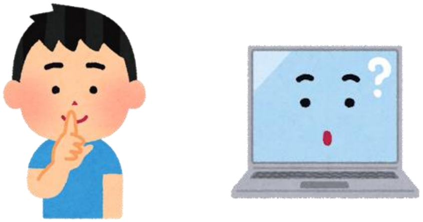
Avoids hallucination thanks to CTC hard alignment

Groundtruth reference	OWSM v3.1 output	OWSM-CTC output (ours)
in search of the mythical treasure your grandfather is supposed to have secreted there he laughed and the girl instinctively shuddered with a newborn distrust there was no mirth in the sound	in search of the mythical treasure your grandfather is supposed to have secreted there ha ha ha ha ha ha ha ha ha ha ha ha ha ha ha ha ha ha ...	in search of the mythical treasure your grandfather is supposed to have secreted there he laughed and the girl instinctively shuddered with a new-born distrust there was no mirth in the sound
and with her they began a national tour that took them all around the country	they take a national gira which leads to rerererererererererererererere ...	with learn a national tour that leads them to run the entire country

	# failed samples (↓)
OWSM v3.1	453
OWSM CTC	1

OWSM CTC’s behavior

- What happens when we throw the silence recording?

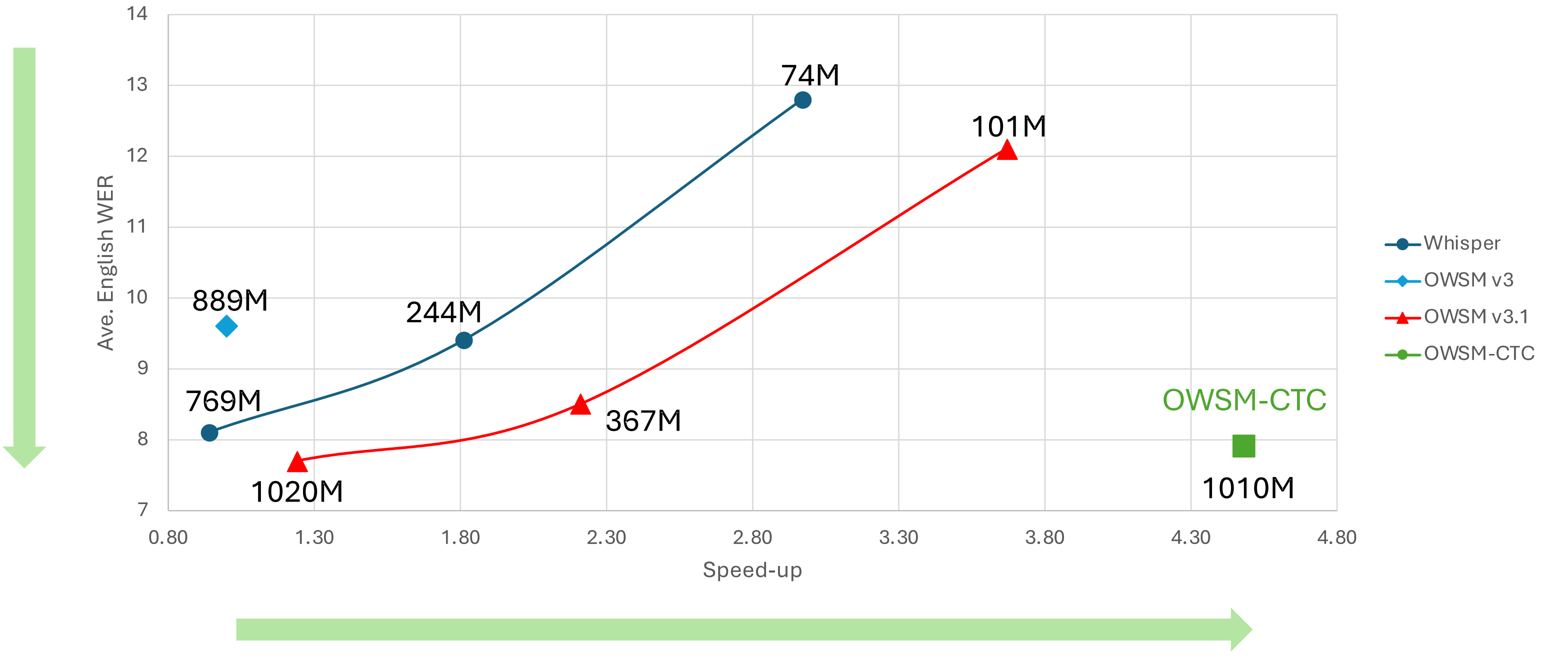


OWSM-CTC Robustness to Hallucination

- OWSM-CTC is more robust to hallucination
- Input: silence

Model	Sample 1	Sample 2	Sample 3
Open AI Whisper	thank you	hello	Tchau.
OWSM v3.1	thank you	good things to do	(Applause)
OWSM-CTC	.	(	()

Even faster with non-autoregressive decoding



How to improve OWSM

1. Format
2. Training cost → OWSM v3.1
3. Hallucination → OWSM CTC
4. Data volume: we will have more investigation in the next section

Today's agenda

- Introduction of our efforts on reproducing Whisper
  - Motivation
  - Experiments: Why they are working and why they are not working
- Improve the model based on why
- Scaling works or not

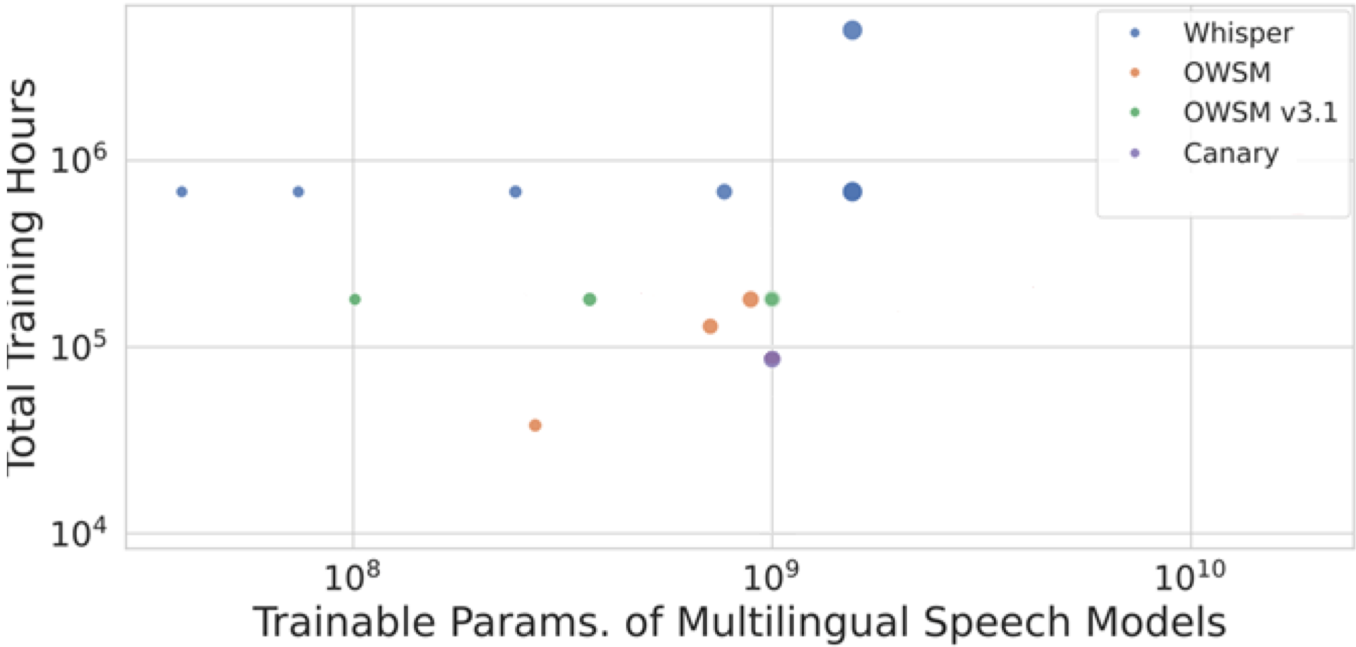
Remaining interests and issues

- Data volumes
- Emergent capabilities

We further conduct investigations

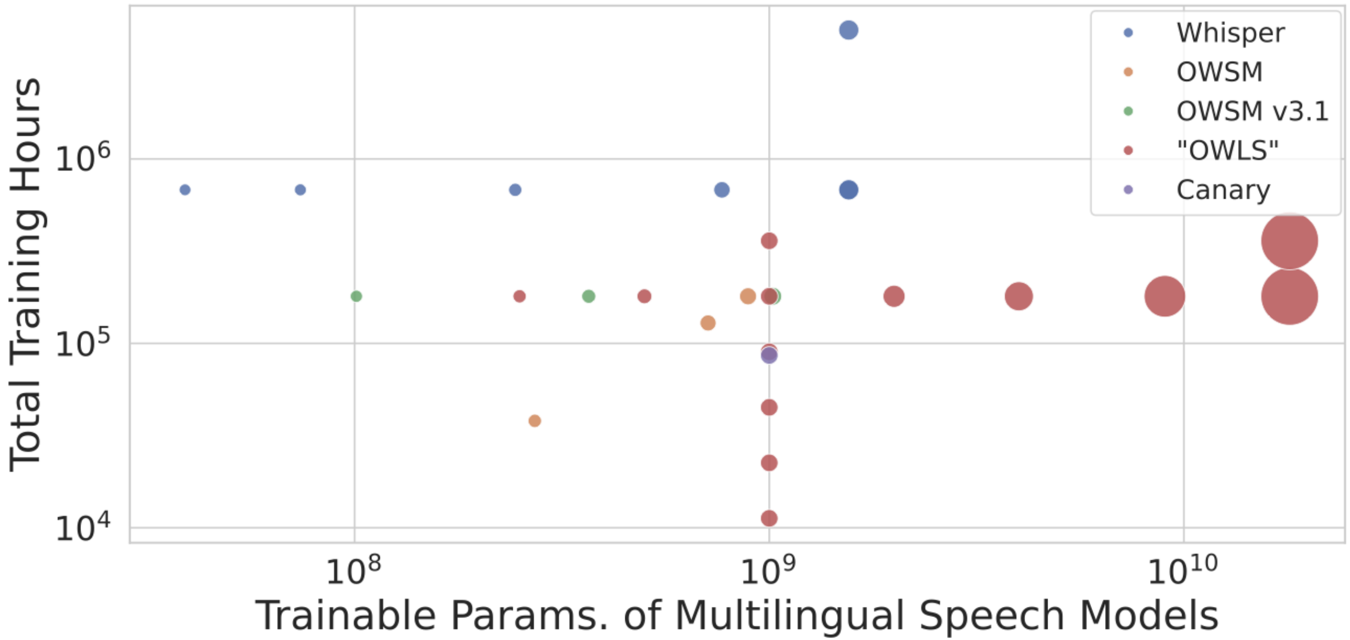
- The effects of data scaling
- The effects of model scaling

Overview



Whisper
OWSM
OWSM v3.1
Canary
Total Training Hours
10^6
10^5
10^4
10^8
10^9
10^{10}
Trainable Params. of Multilingual Speech Models

Overview



Whisper
OWSM
OWSM v3.1
"OWLS"
Canary
Total Training Hours
10^6
10^5
10^4
10^8
10^9
10^{10}
Trainable Params. of Multilingual Speech Models

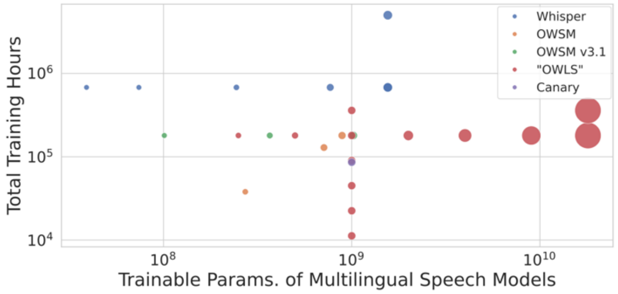
Scaling Data

Scaling Data

- To get SOTA, we need lots of training data
- How much training data do we need?
    - Domain-specific
    - Language-specific
    - Model-specific
- Can we instead predict downstream gains from scaling data?

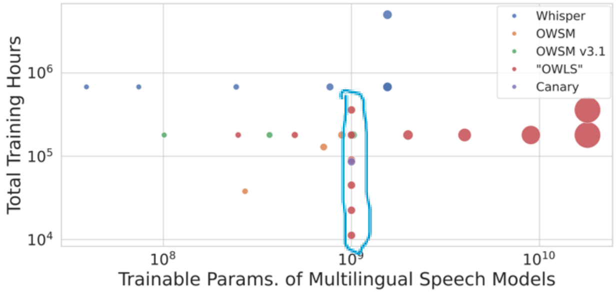
Scaling Data

- We train OWSM on different sizes of data:
    - 1B parameter Transformer encoder decoder
    - 11K, 22K, 45K, 90K, 180K, 360K hours of data
        - 180K and below use the original OWSM dataset
        - 360K uses additional data from **YODAS**



Scaling Data

- We train OWSM on different sizes of data:
  - 1B parameter Transformer encoder decoder
  - 11K, 22K, 45K, 90K, 180K, 360K hours of data
    - 180K and below use the original OWSM dataset
    - 360K uses additional data from **YODAS**



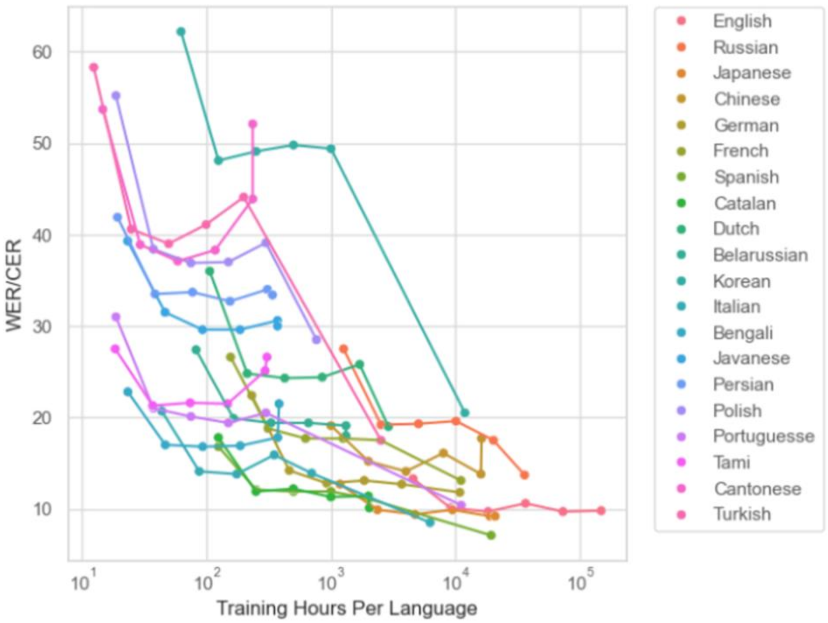
YODAS: Youtube-Oriented Dataset for Audio and Speech

- Large audio data collection to fill out the gap between 180K and 680K
- **YODAS** project
  - Crawling **Creative Common portion** of YouTube
  - Over 140 languages
  - Over 300K hours (still growing)
  - We’re working on further crawling, cleaning, and effective usage

https://huggingface.co/datasets/espnet/yodas

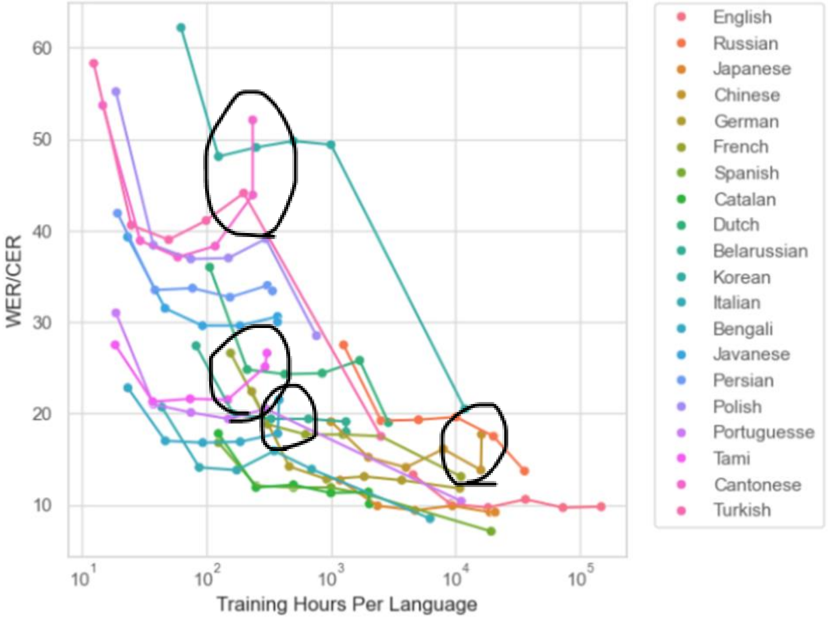
Scaling Data

- Can we predict WER as a function of training data?



Scaling Data

- Can we predict WER as a function of training data?



Scaling Data

- Better in average
- Strongly depends on the data distributions (languages, domains, etc.)

Scaling Model Size

Scaling Model Size

- Most speech models are small relative to modern NLP ones:
    - **T5** – 76M to 11B LM
    - **UL2** – 167M to 20B LM
    - **NLLB** – 200B MoE Machine Translation Model
    - **Llama** - 7B to 405B general LLM
    - **Command-R+** - 7B to 105B multilingual LLM
- Largest models in speech:
    - **Whisper** 1.5B
    - **MMS** 1B
    - **XLS-R** 2B
    - Google USM 2B
    - Meta ASR 10B
    - Google ASR 10B

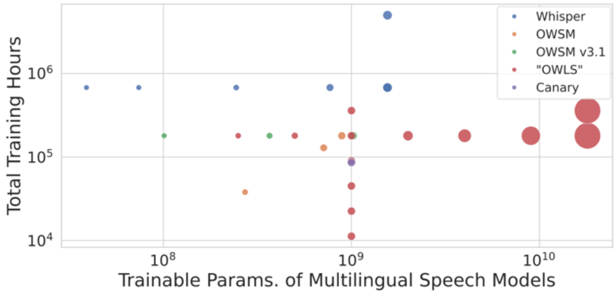
Bolded indicates publicly available checkpoints

Scaling Model Size

- What happens if we scale OWSM to 10x the size?
- We investigate the effect of scaling model size on multilingual ASR models

Scaling Model Size

- We train 7 new versions of OWSM
  - 180K hours of multilingual ASR/ST data
  - 0.25B to 18B parameters
  - Transformer encoder-decoder
  - Same learning rate, batch size, scheduler, training steps (675K steps)

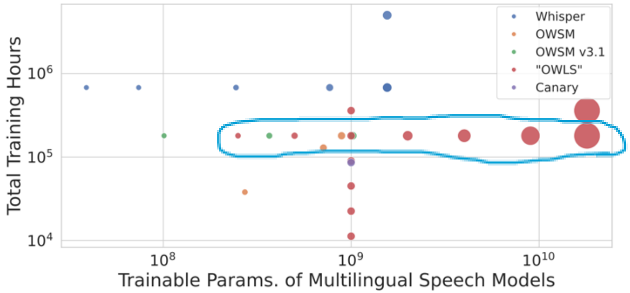


- What will happen?

Scaling Model Size

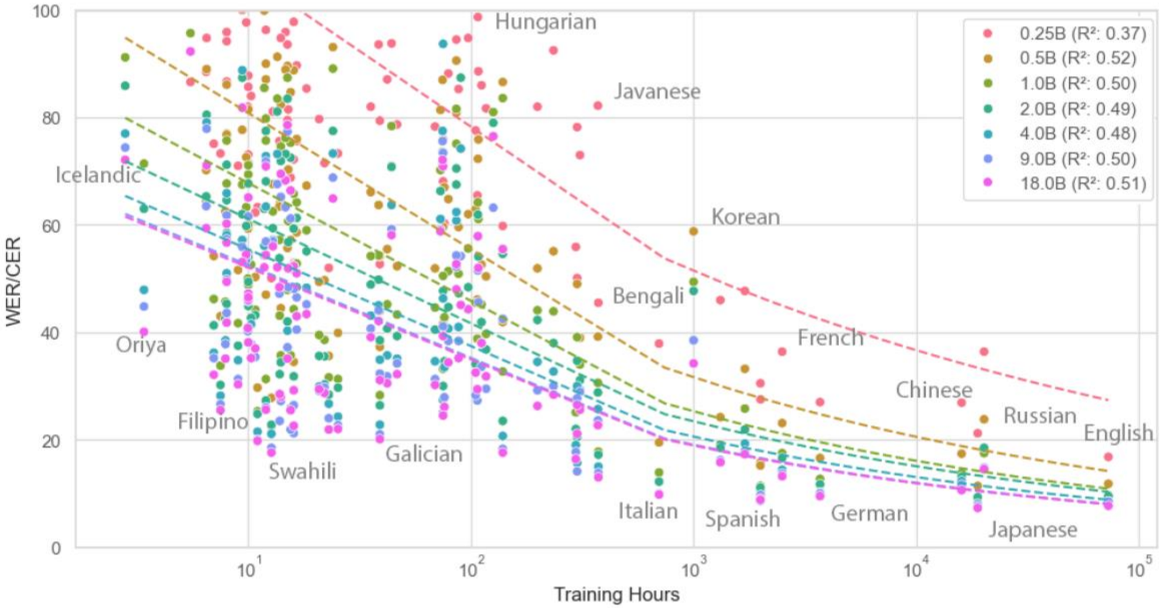
- We train 7 new versions of OWSM
  - 180K hours of multilingual ASR/ST data
  - 0.25B to 18B parameters
  - Transformer encoder-decoder
  - Same learning rate, batch size, scheduler, training steps (675K steps)

- What will happen?

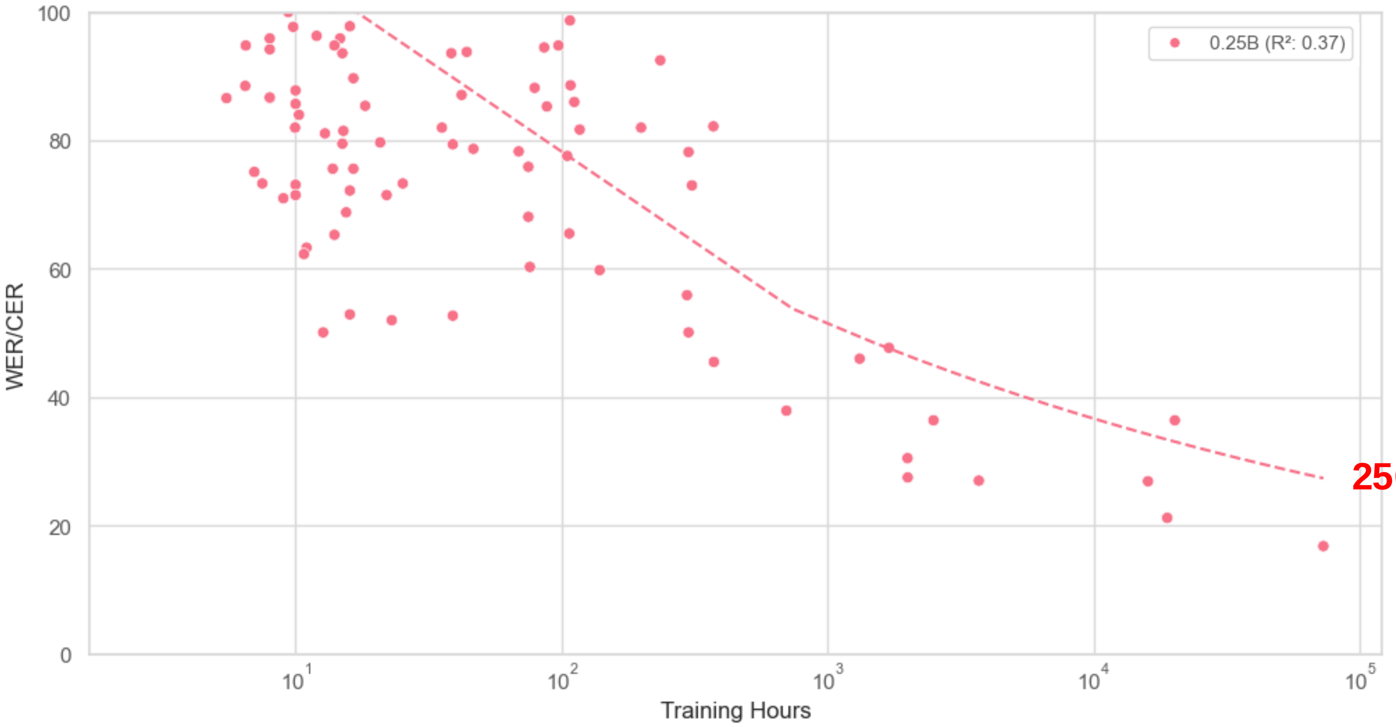


Scaling Model Size

- Even when scaling from 2B to 18B parameters, we can see improvements in WER for both high and resource languages

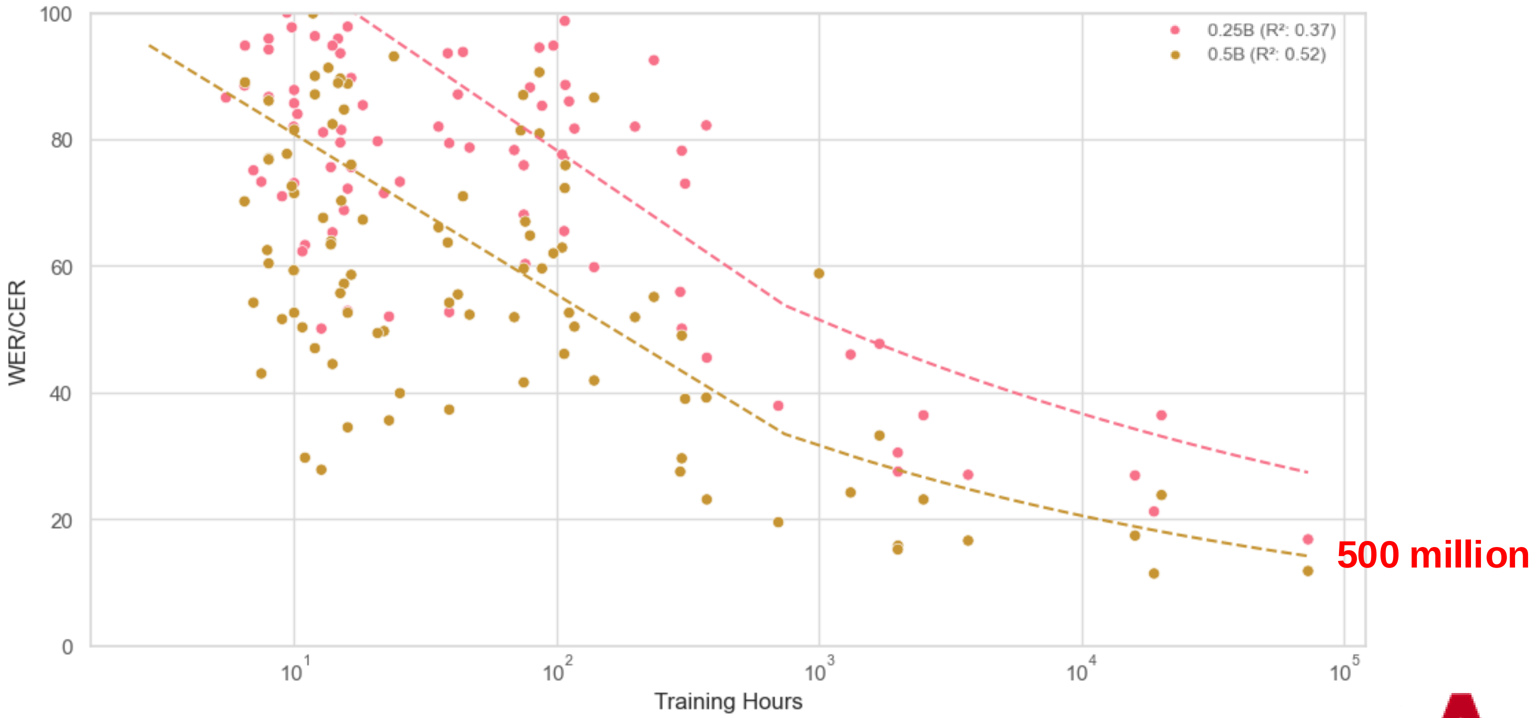


Scaling Model Size

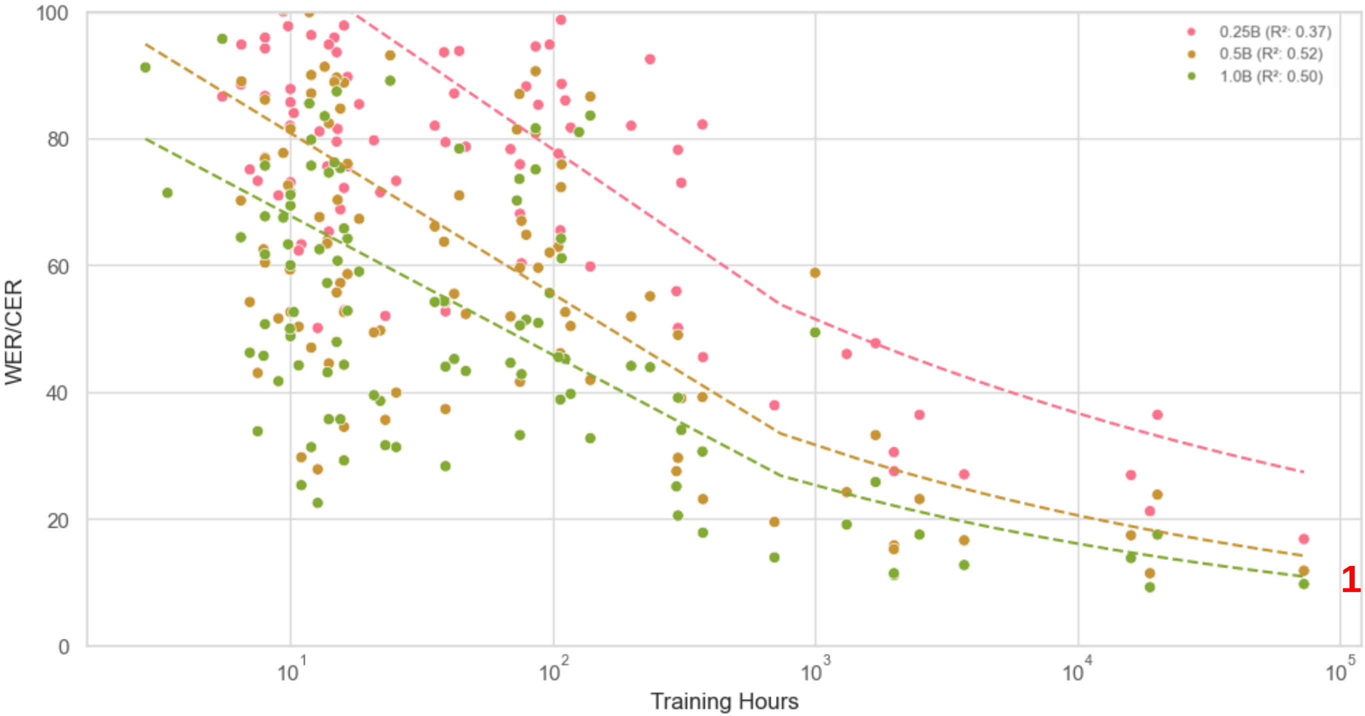


250 million

Scaling Model Size

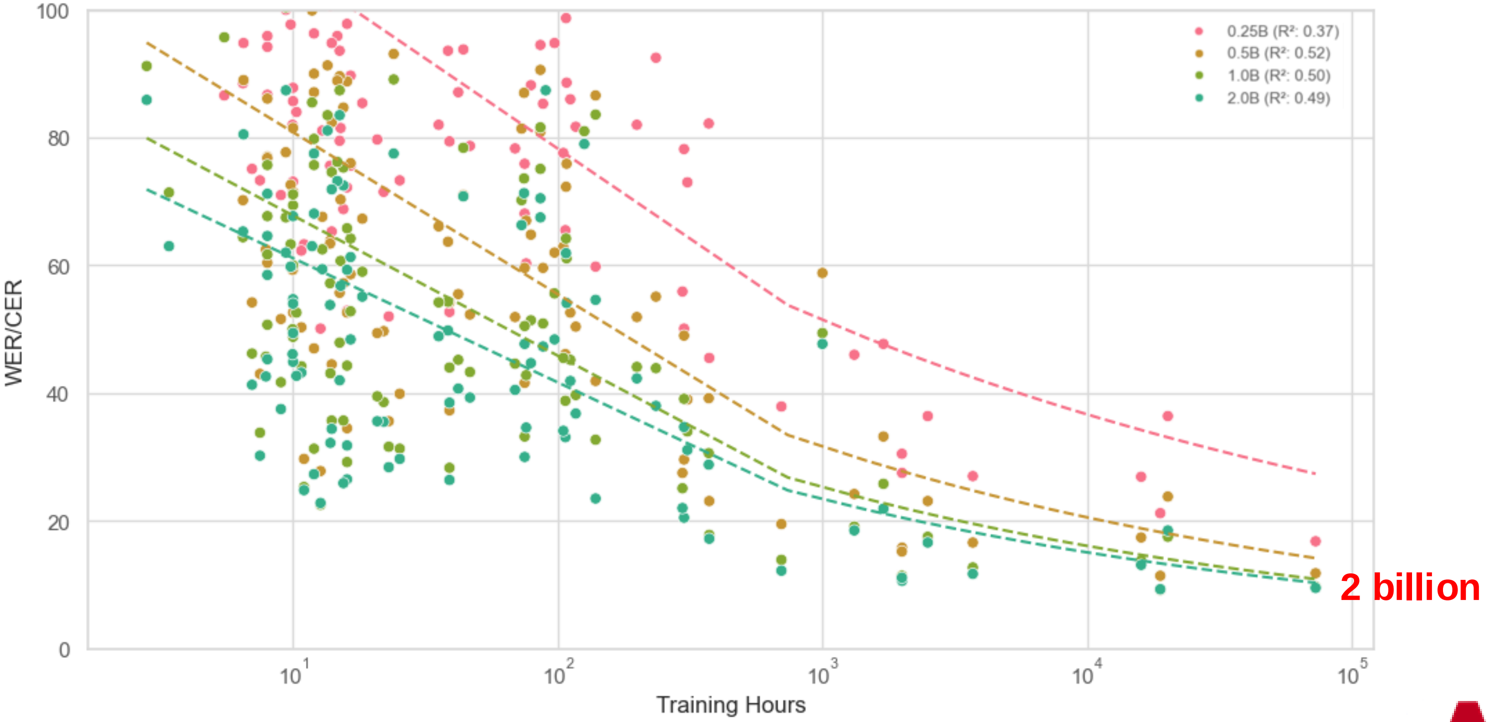


Scaling Model Size

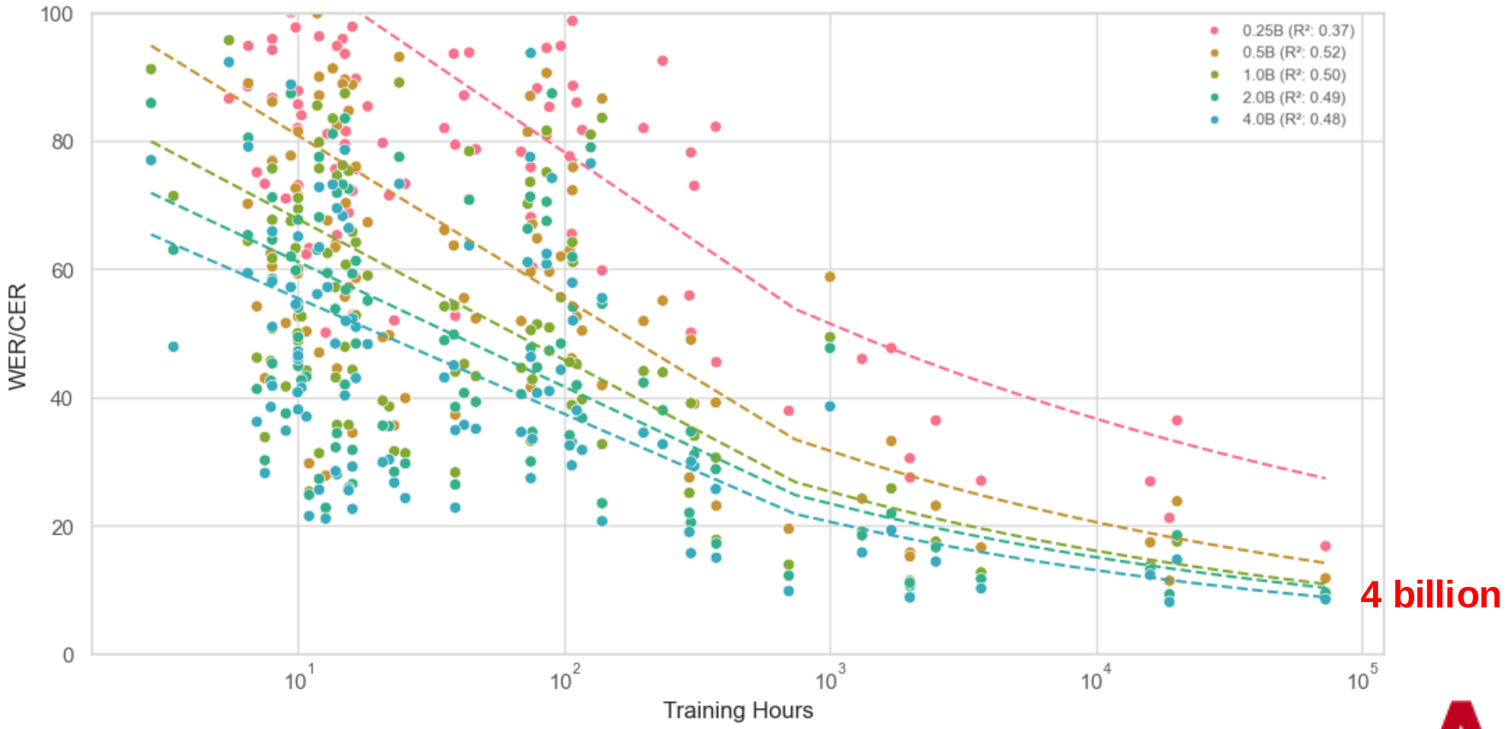


1 billion

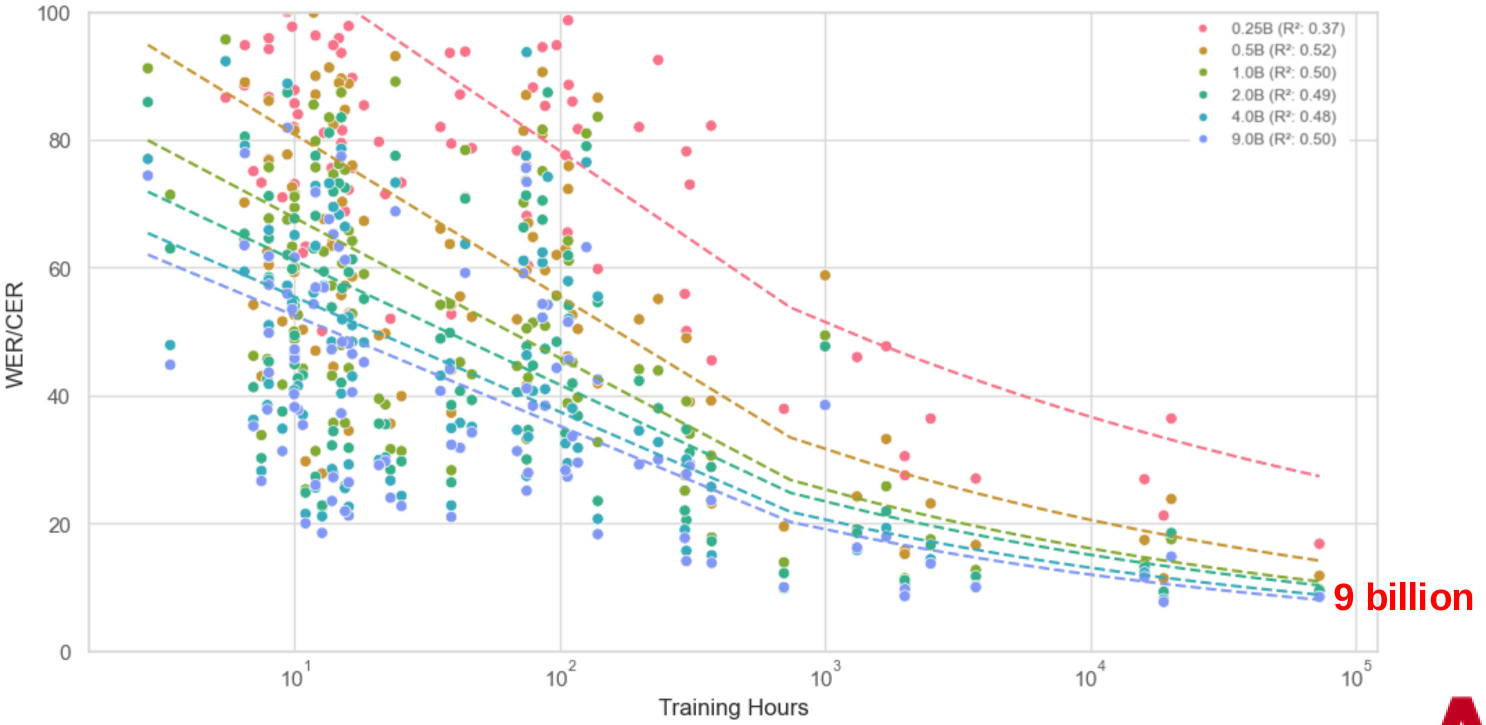
Scaling Model Size



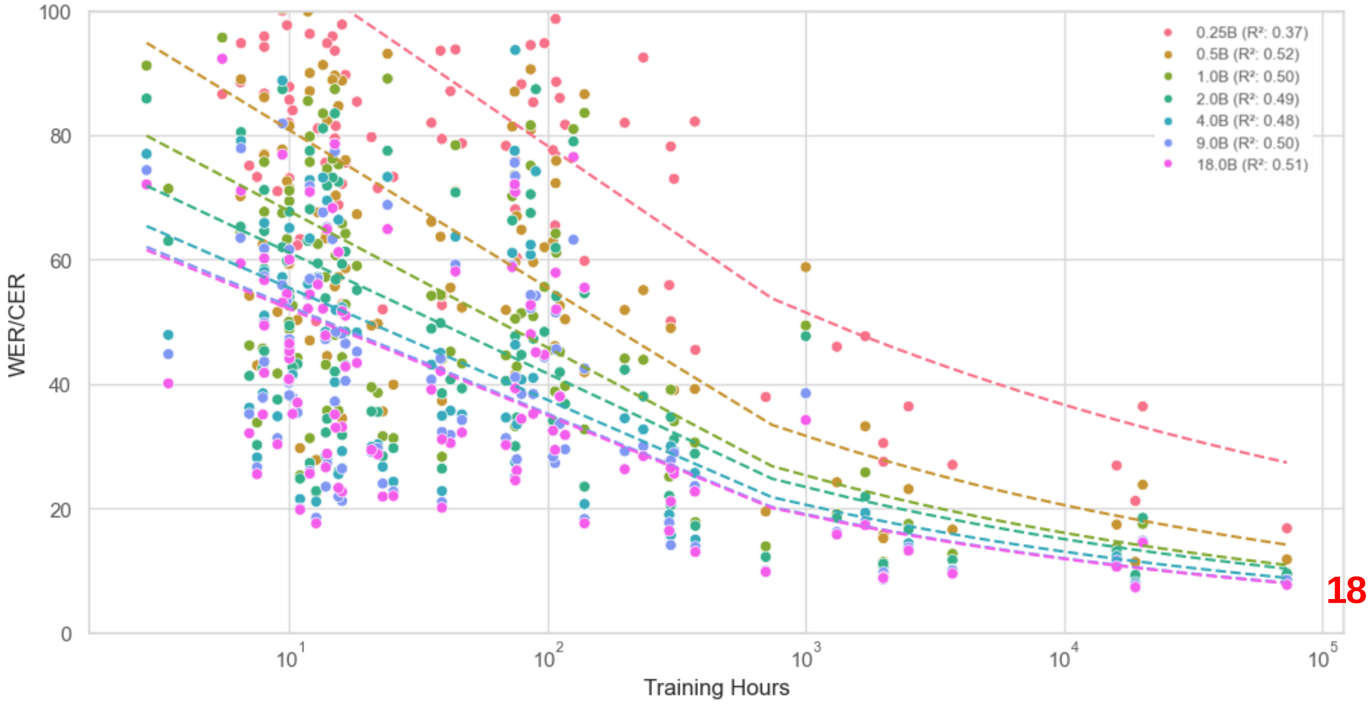
Scaling Model Size



Scaling Model Size



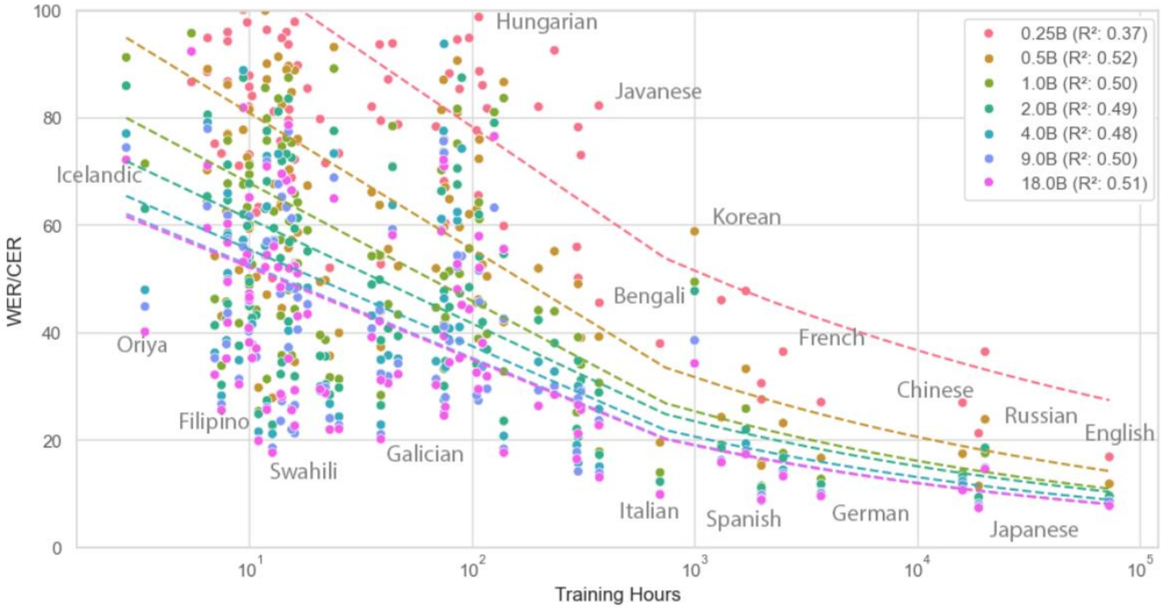
Scaling Model Size



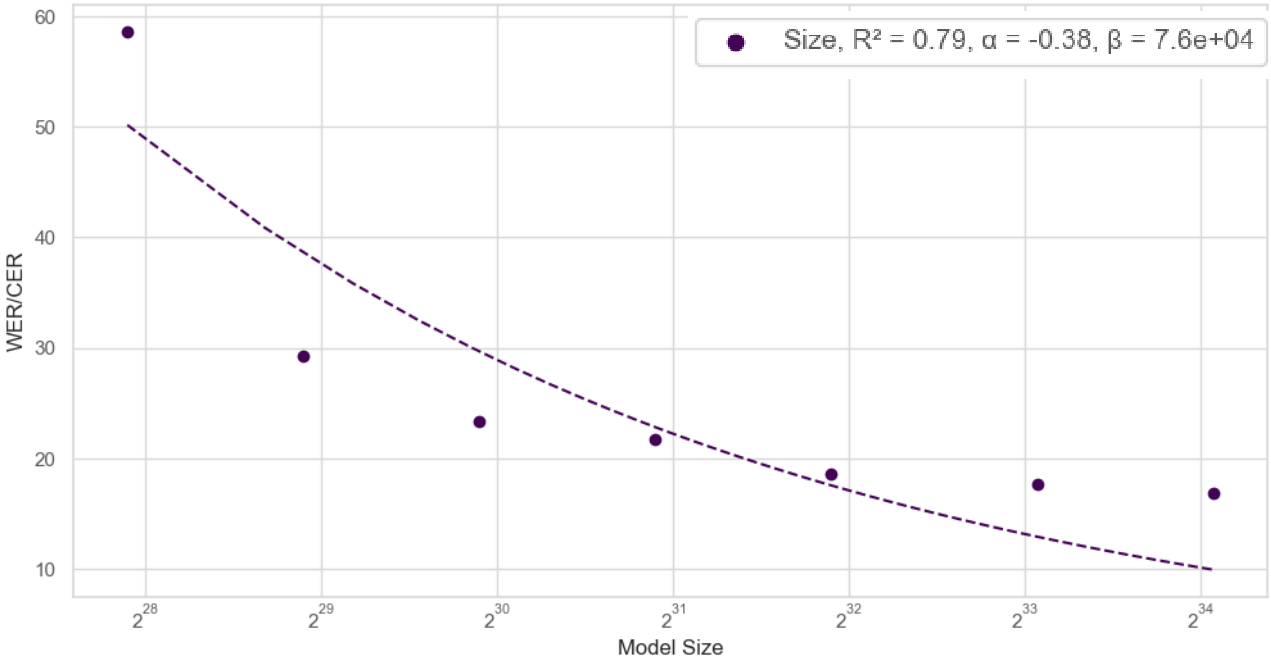
18 billion

Scaling Model Size

- Even when scaling from 2B to 18B parameters, we can see improvements in WER for both high and resource languages



Scaling Model Size - Average

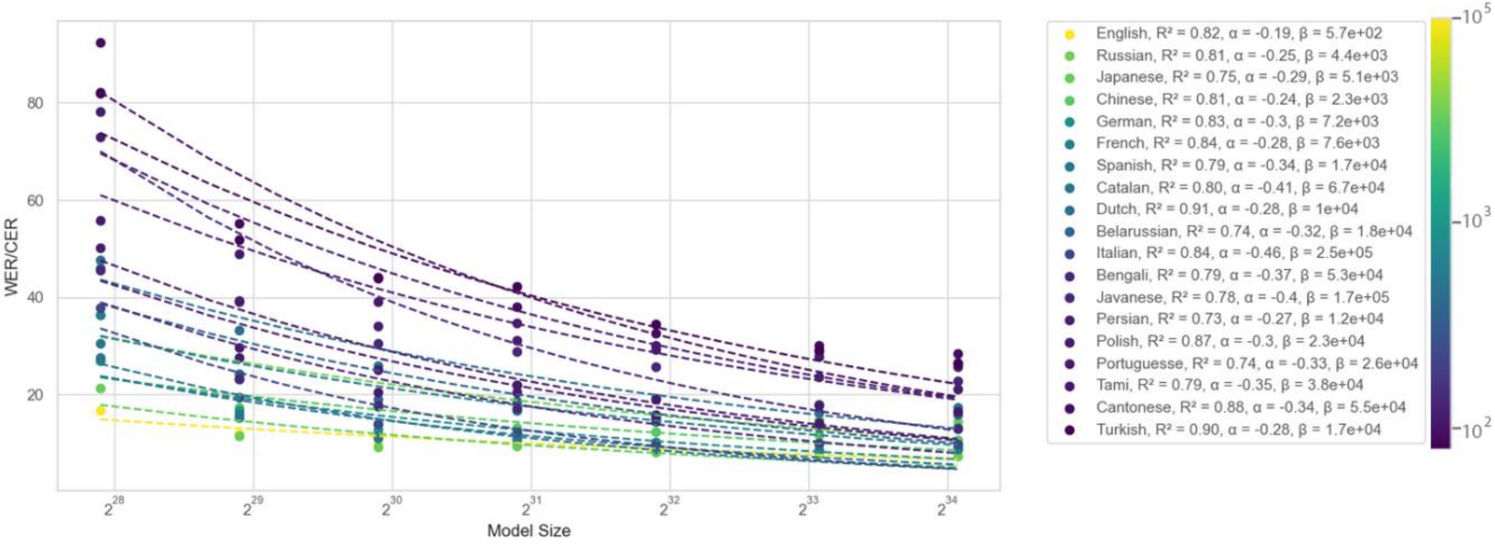


Scaling model size vs. data

- Model size is more correlated, more solid
- We will look into each language

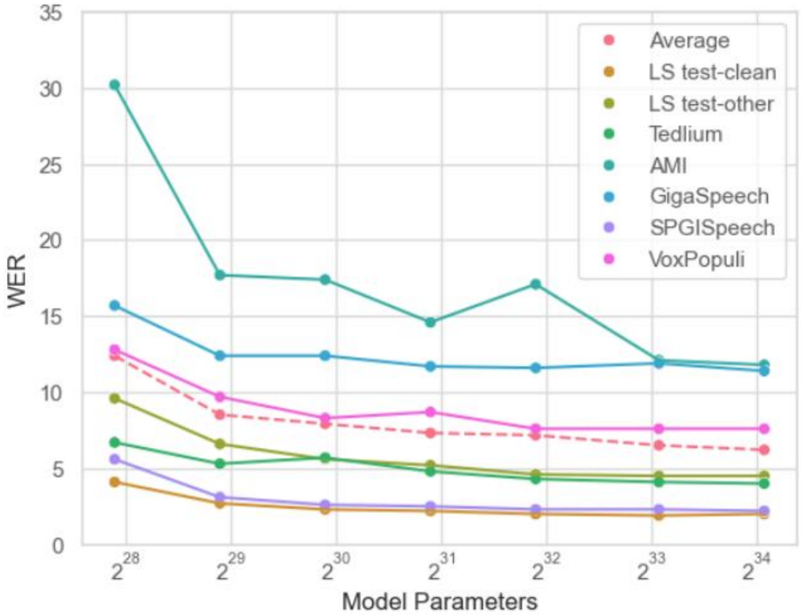
Scaling Model Size

- A power law w.r.t model parameters can predict performance well



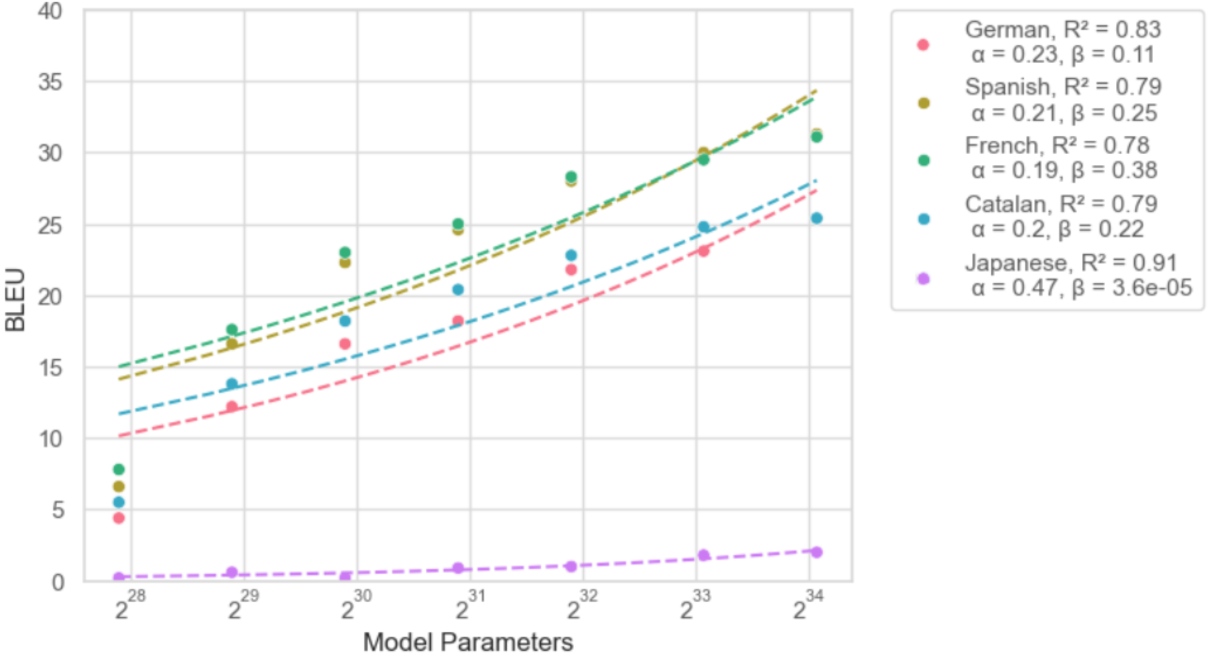
Scaling Model Size

- Even English, the most saturated language, sees consistent improvements



Scaling Model Size

- Translation shows similar trends
- But there are still limitations when data is too scarce



Scaling model size

- Overall, it mitigates the bias issues (domains and languages)
- Large enough capacities avoid parameter override by dominant data

Emergence in Large Models

- LLMs are known to exhibit *emergent abilities* at scale
    - Abilities found in large models but not found in smaller ones

- Can we say the same for speech models?

Orthographic Understanding

- We find larger models have enhanced orthographic capabilities

Table 3. **Orthographic opacity examples of Japanese and Chinese.** The same phone sequence can be written in different ways.

Orthography	Example
Romanization (zh)	shì shī shì
Simp. Chinese	室诗士
Trad. Chinese	室詩士
Romanization (jp)	hashi
Hiragana	はし
Katakana	ハシ
Kanji	橋

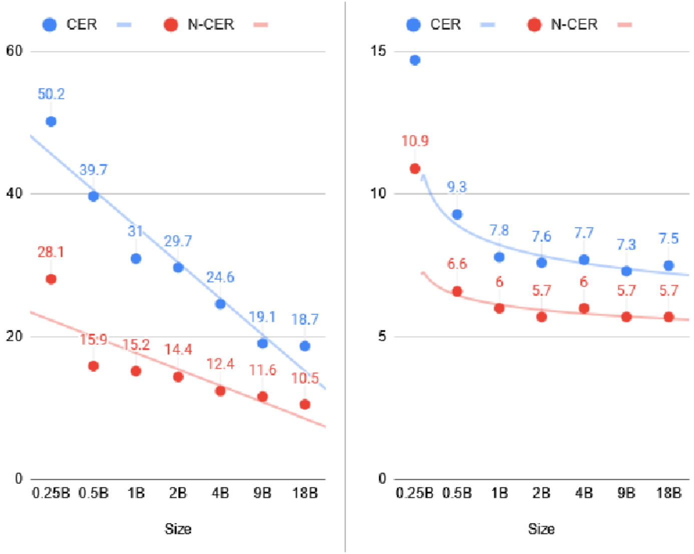


Figure 10. **Effects of model scaling on orthographic understanding on Chinese (left) and Japanese (right).**

Speech In-context Learning

- We can teach models a new language with in-context learning

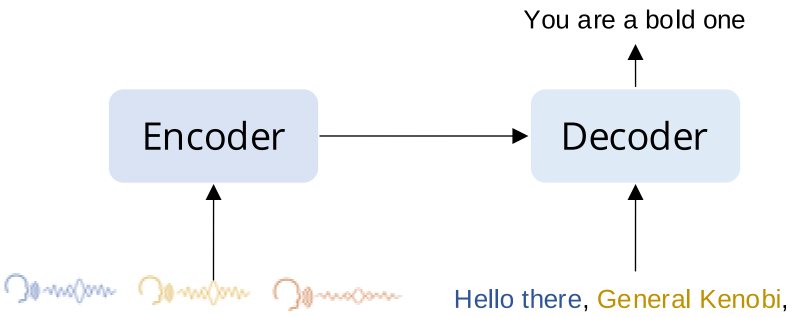


Table 5. **Quechua CER on ICL with 0 / 1 / 2 / 3 examples.** The overall best result is **bolded** while the best result for each model size is underlined.

Params.	$k=0$	$k=1$	$k=2$	$k=3$
0.25B	36.9	35.1	<u>33.7</u>	34.5
0.50B	53.3	39.2	<u>33.8</u>	33.9
1B	41.8	35.0	<u>31.6</u>	31.8
2B	47.3	35.1	<u>31.9</u>	33.2
4B	40.4	32.4	<u>31.2</u>	31.8
9B	38.3	31.3	28.1	<u>**27.4**</u>
18B	41.3	32.7	31.3	<u>28.1</u>

Mondegreen

- 空耳 in Chinese / Japanese
- Semantically relevant mishearing
- “Bon Appetit” vs “Bone Apple Tea”
- “What’s time” vs. “掘った芋 (hotta imo)”

Table 4. **Evaluation of mondegreen capabilities.**

Params.	PPL	MOS
0.25B	1338	1.9
0.50B	728	4.1
1B	559	3.5
2B	491	3.6
4B	436	3.8
9B	**372**	**4.8**
18B	429	4.4

Mondegreen

Table 9. **Example mondegreen generations and their corresponding original text.**

Source	Text
Original	Vir daardie rede, als wat jy op die TV sien, het die kante gesny, bo, onder en kante.
0.25B	Dore the rear of the ozvatioctiya fissic.
0.5B	For Dore the Rieda also got the optic fissure.
1B	The order did read as Vatican's affiliate for the first time.
2B	The Daily Director also wrote the optics for his work.
4B	For the order read, also what the optieth is.
9B	The door of the red house was fatty, and the squad was very tired.
18B	For the ordinary, the oasis varies between the oasis and the oasis.
Original	Alle burgers van die Vatikaan Stad is Rooms Katoliek.
0.25B	Alabarkers fan diva
0.5B	Alabama cares for the development of the reservation.
1B	allebergers van the valley
2B	Alabama kerrs fan the game.
4B	Alabama, Cars, Fan, Diva.
9B	All the birds catch the worm.
18B	All the workers found the vat.

Conclusion

- Scaling to more data
  - Hard to predict benefits
  - More is still usually better
  - Diversity matters

- Scaling to larger models
  - Scaling is also useful in speech!
  - Leads to more fair performance for different language varieties

Summary

- **Speech foundation models** are a very attractive research direction!
- Let’s keep open-source efforts for **reproducibility and transparency**
- Let’s **understand** the behaviors!
    - OWSM is transparent for the data, source code, computing resources, and all other information → we can conduct such scaling experiments!
    - Please use it and give us any feedback! We can identify the issue thanks to the transparency!
    - We actually have a lot of feedback about the fairness and model biases!

https://huggingface.co/spaces/pyf98/OWSM_v3_demo

Take home message

- **Large computation cost** ❼ carbon footprint/global warming
- **Without transparency**
  - Further increase in the carbon footprint with redundant trials
    - This scaling work is a necessary evil
    - Other institutions do not have to redo the experiments
  - Lose control → further damage to the earth and humans

We have had issues in the past (pollution, nuclear issues, etc.)

- Someone must be in charge of the responsibility of AI (← Academia?)
- **Complementary collaboration with the industry and academia**

Be responsible for our society!

Carnegie
Mellon
University

Thank you!

Language
Technologies
Institute

Watanabe's
Audio and Voice Lab