OWLS: Scaling Laws for Multilingual Speech Recognition and Translation Models

William Chen[1], Jinchuan Tian[1], Yifan Peng[1], Brian Yan[1], Chao-Han Huck Yang[2], Shinji Watanabe[1]

Carnegie Mellon University[1] Nvidia[2]

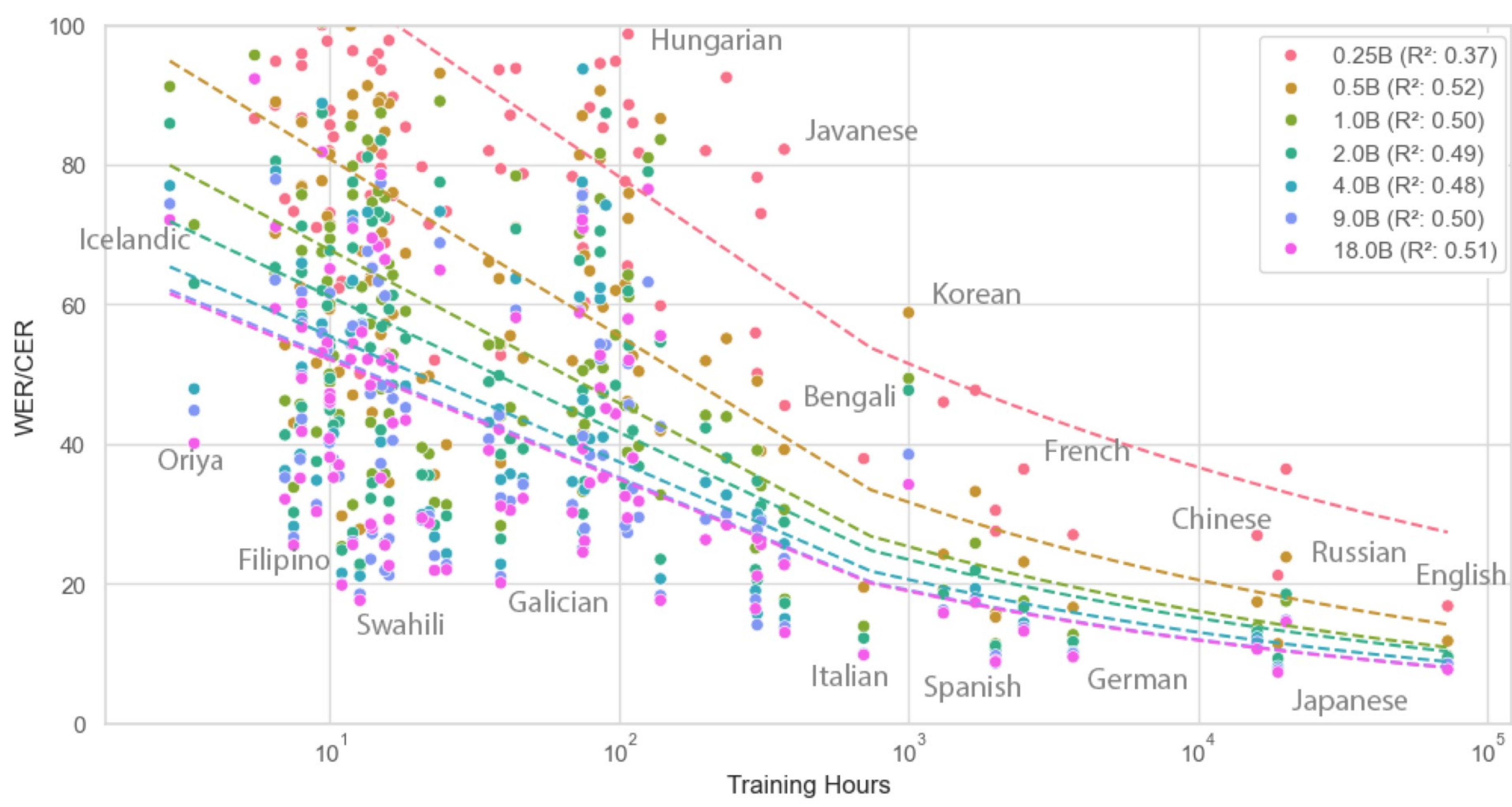
Introduction and TLDR

- LLMs exhibit profound capabilities due to their scale in terms of both size and training data.

- What happens if you scale speech recognition models to that level?

- We train Whisper-style models on 180K hours of data from 250M to 18B parameters to find out.

- We also experiment with scaling training data from 11K to 360K hours, when model size is fixed at 1B parameters.



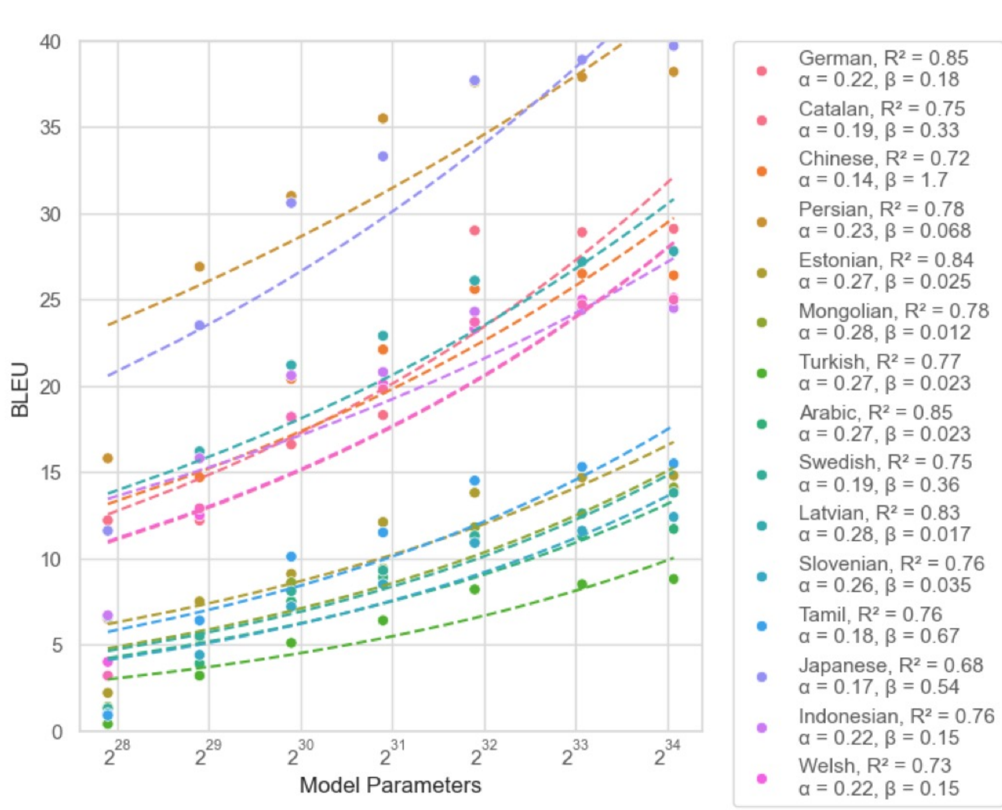
- ASR gets predictably better as model size increases. Improves low-resource WER without degrading high-resource.

- ST BLEU can nearly double from 250M to 18B parameters!

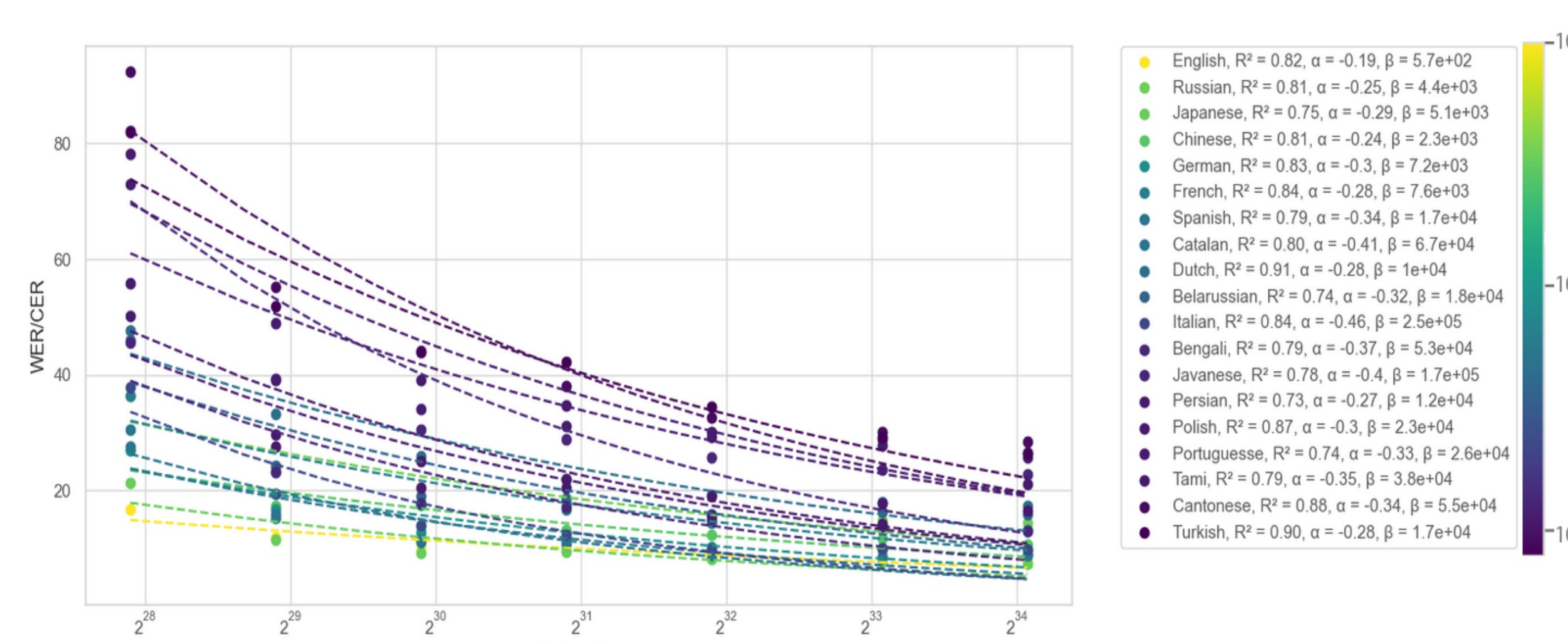
- Scaling training data is less conclusive; diversity matters more than sheer quantity.

- Larger models exhibit interesting zero-shot abilities, such as in-context learning and orthographic understanding.

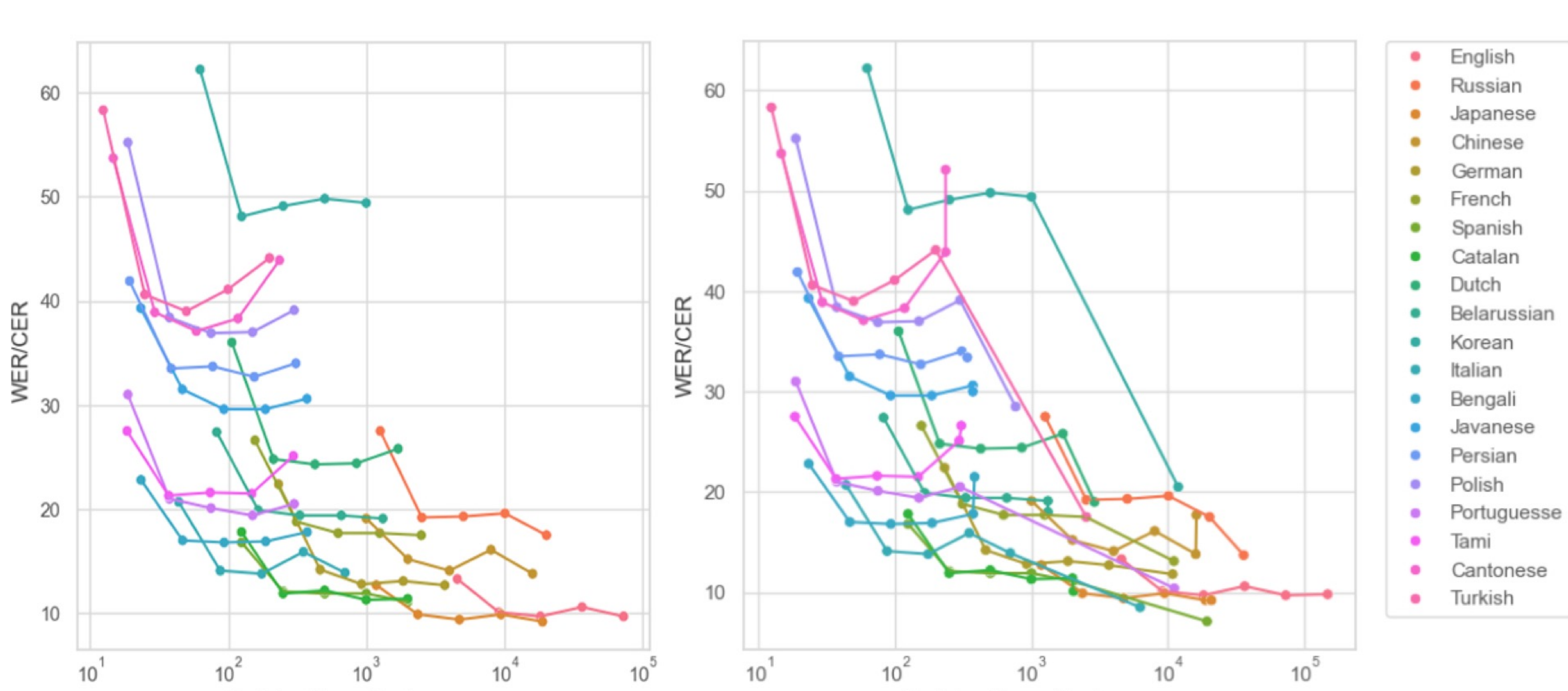
Scaling from 250M to 18B Parameters



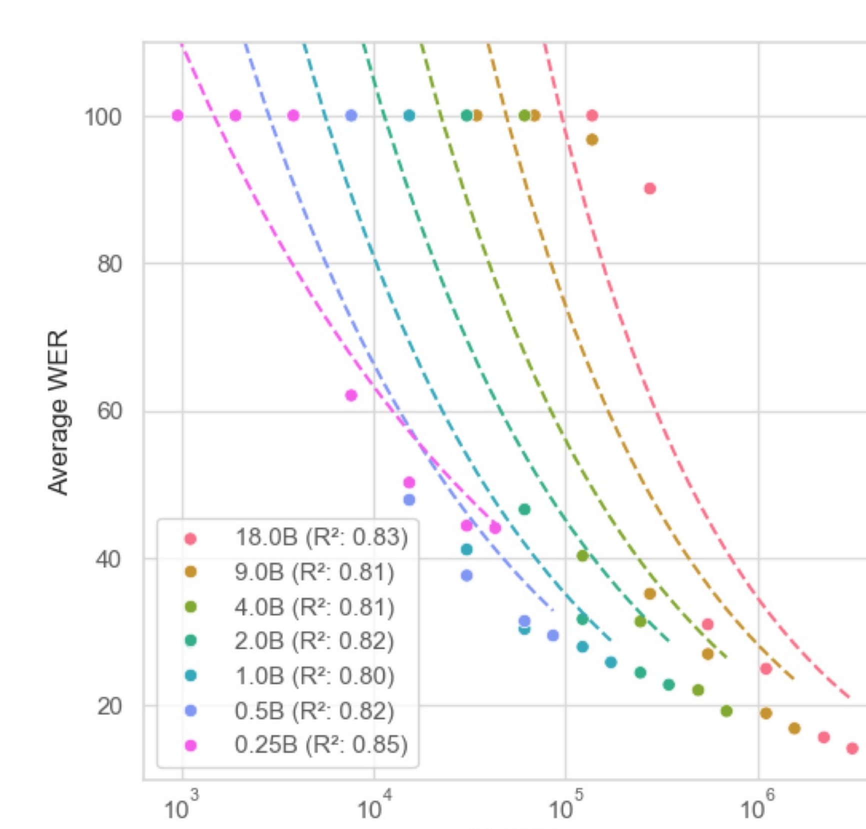
- ST performance improves from < 10 on 15 languages to > 10 on 17 languages
- ASR WER can be directly modeled with a power law function
- Substantial improvements from 2B parameters to 18B parameters! Current models may be too small.



Scaling Data and Compute



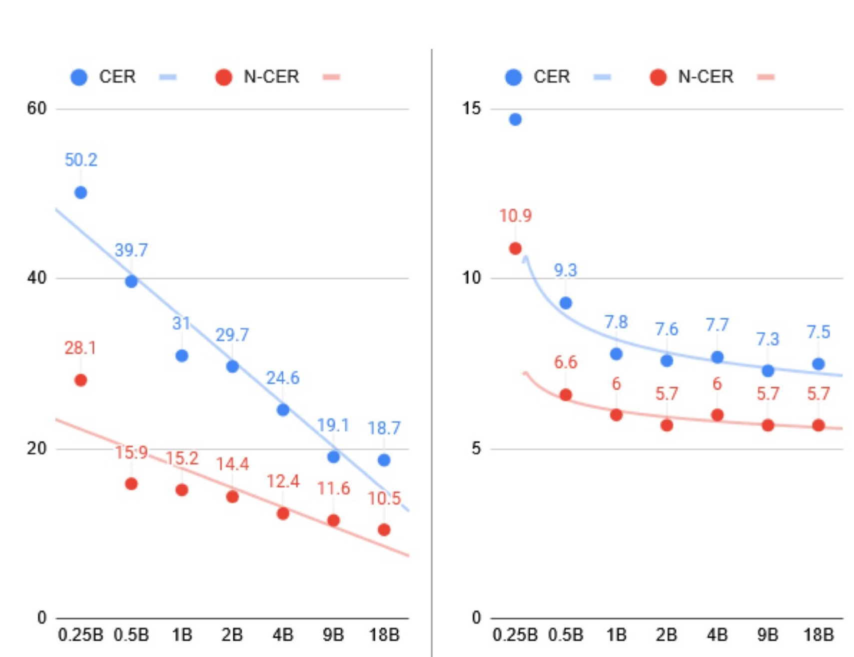
- Left figure shows the effect of increasing from 11K hours of mostly read speech to 360K hours of read + conversational speech.
- General performance doesn't really improve if you keep adding in data from the same domain.
- Right figure compares training TFLOPS with WER. Models around 1-2B parameters are more compute optimal.



Zero-shot Behaviors and Emergent Abilities

Orthographic Understanding

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



- Larger models can be better at choosing what writing system to use. Left shows Chinese, right is Japanese.
- We calculate CER at the character and phone level. Phone level performance stagnates, but character-level improves!

In-Context Learning

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

- We teach the model to perform ASR on Quechua using ICL.
- Concatenate prompt audio and text along time dim.
- Larger models are better at using more context!

Semantic Mishearing

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

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.

- Humans can mishear audio into semantically informative sentences. "Bon Appetit vs Bone Apple Tea"
- Larger models are more likely to do the same, while smaller models are more phonetic.