# Whisper

## Robust Speech Recognition via Large-Scale Weak Supervision

Peter Sullivan
DL-NLP 10-13-22

# Motivation + High Level Overview

- DL models are ‘brittle’ → train on a massively diverse data
- How to get massive data? → internet video with existing captions
  - Assume OK quality on human transcriptions, detect and remove ASR captions

- *Weakly supervised* multitask training on
  - ASR (English audio + English captions)
  - Speech Translation (~~English~~ audio + English captions)
  - Non-English ASR (~~English~~ audio + *same language* captions)
  - Voice activity detection
- Use prefix tokens to indicate which task

# Data

*“audio that is paired with transcripts on the Internet”*

680,000 Hours Total

117,000 Non-English (covering 96 different languages)

125,000 X → English translated data

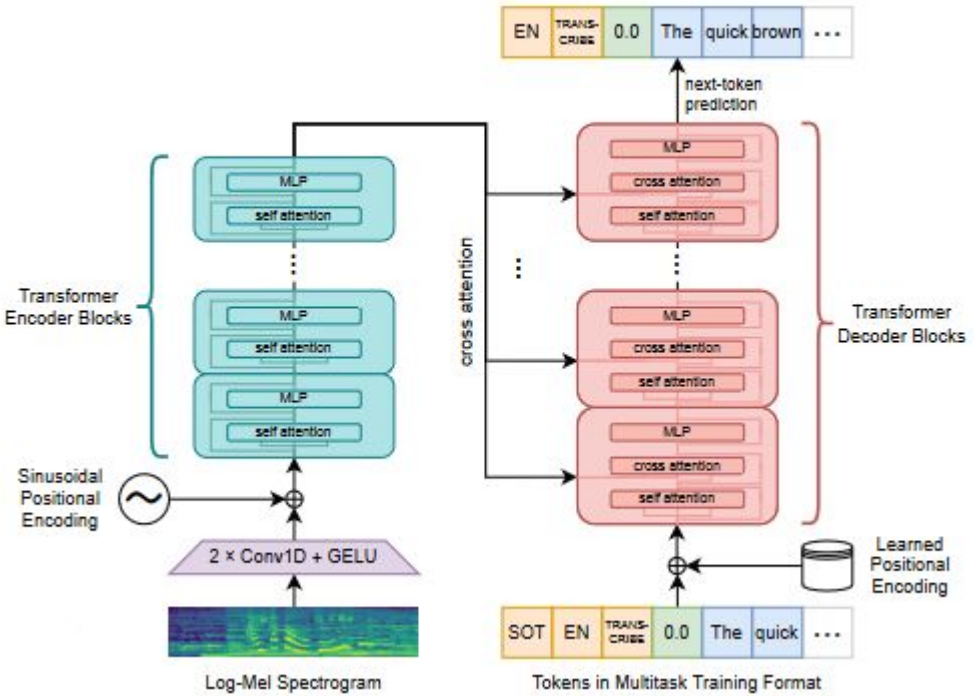
# Model

Observations:

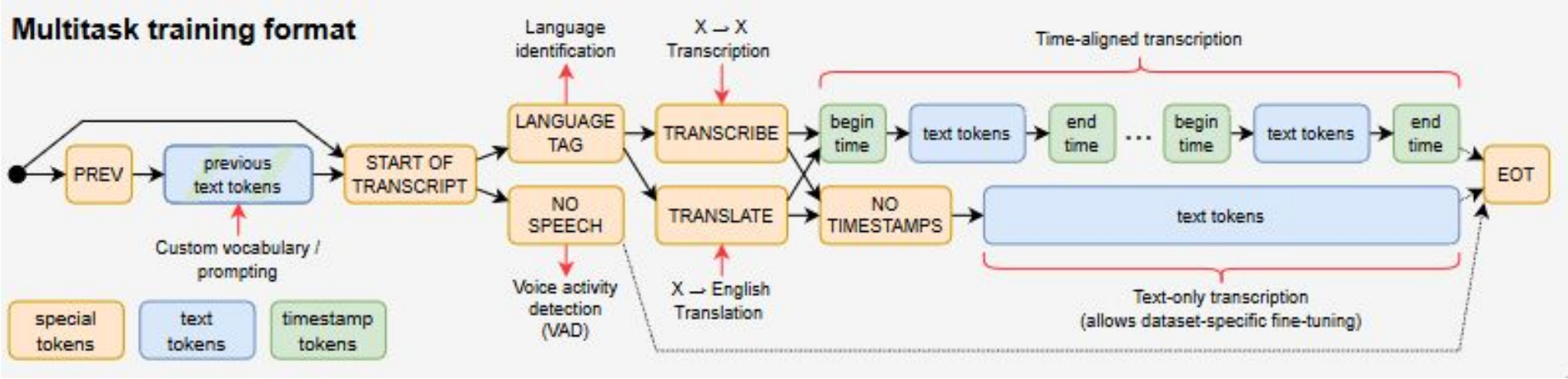
“It’s just a seq2seq transformer”

Novelty: prefix tokens + timestep

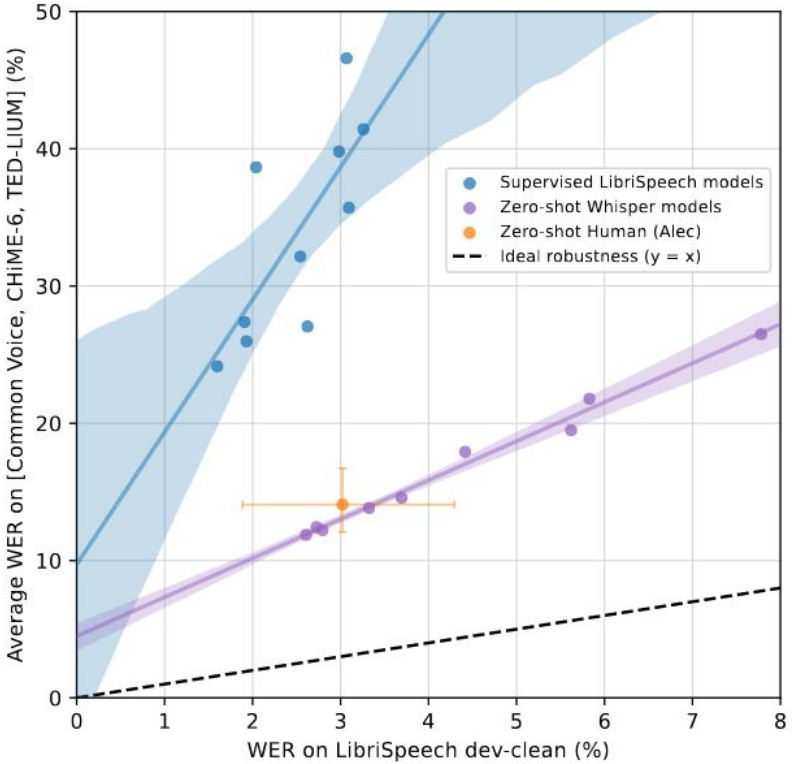
encoder sinusoidal encoding?



# Prefix tokens

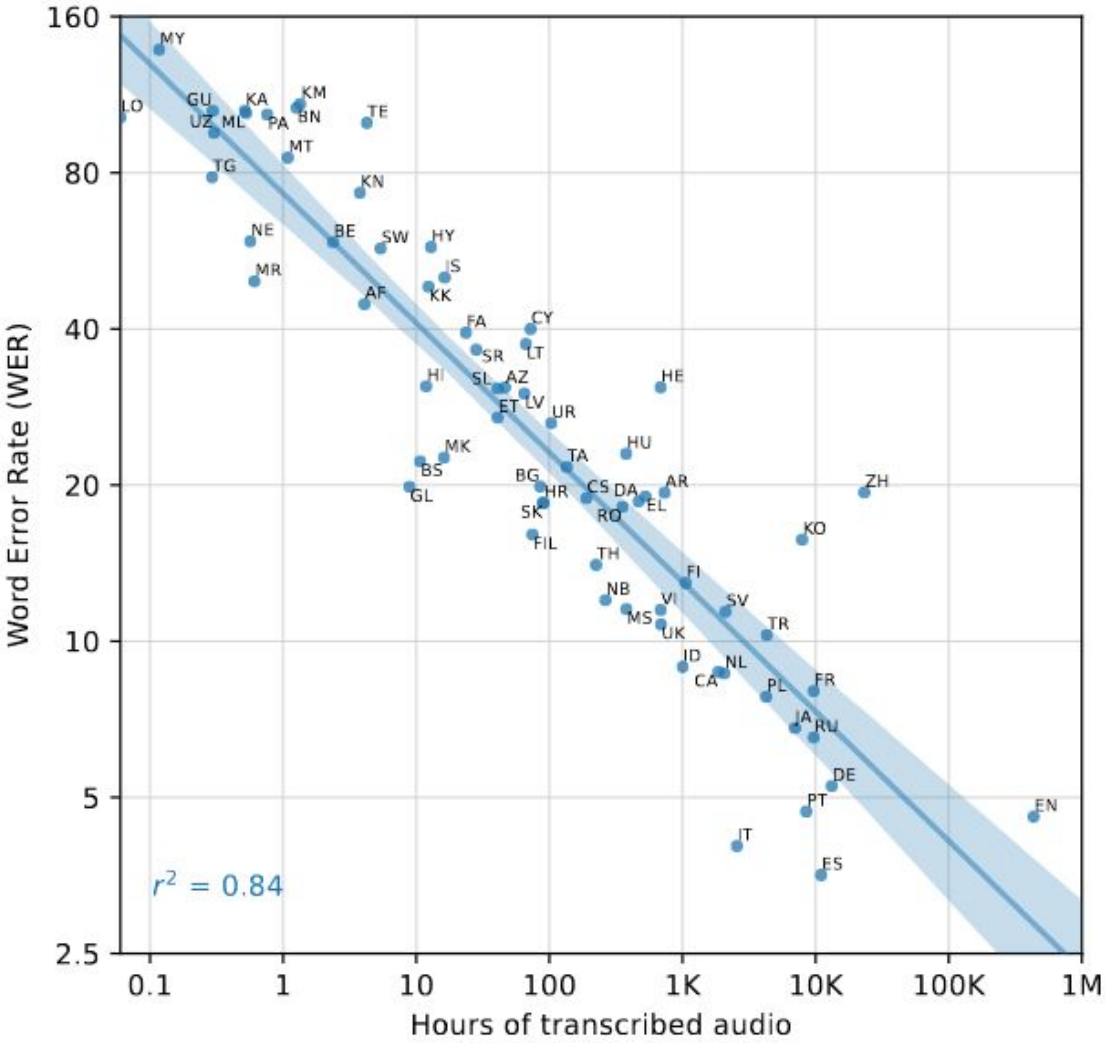


# Robustness?

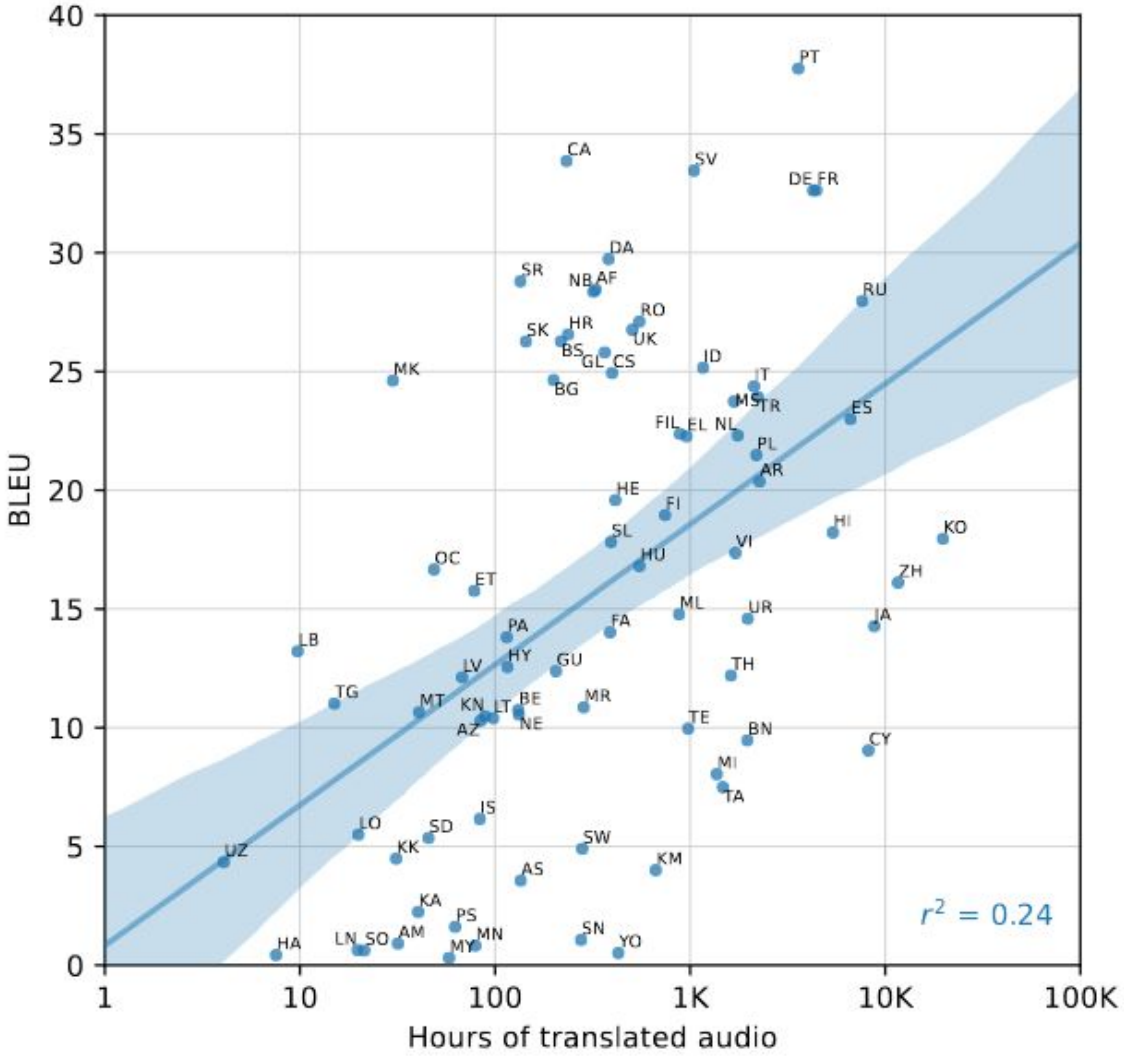



| Dataset | wav2vec 2.0 Large 960h | Whisper Large | RER (%) |
|---|---|---|---|
| LibriSpeech test-clean | **2.7** | **2.7** | 0.0 |
| Artie | 24.5 | **6.7** | 72.7 |
| Fleurs (English) | 14.6 | **4.6** | 68.5 |
| Common Voice | 29.9 | **9.5** | 68.2 |
| Tedlium | 10.5 | **4.0** | 61.9 |
| CHiME6 | 65.8 | **25.6** | 61.1 |
| WSJ | 7.7 | **3.1** | 59.7 |
| VoxPopuli (English) | 17.9 | **7.3** | 59.2 |
| AMI-IHM | 37.0 | **16.4** | 55.7 |
| CallHome | 34.8 | **15.8** | 54.6 |
| Switchboard | 28.3 | **13.1** | 53.7 |
| CORAAL | 38.3 | **19.4** | 49.3 |
| AMI-SDM1 | 67.6 | **36.9** | 45.4 |
| LibriSpeech test-other | 6.2 | **5.6** | 9.7 |
| Average | 29.5 | **12.9** | 55.4 |

# ASR (Fleurs)


Word Error Rate (WER)
Hours of transcribed audio
160
80
40
20
10
5
2.5
0.1
1
10
100
1K
10K
100K
1M
$r^2$ = 0.84
MY
LO
GU
UZ
ML
KA
PA
KM
BN
TE
MT
TG
KN
NE
BE
SW
HY
IS
MR
AF
KK
FA
CY
SR
LT
HI
SL
AZ
HE
ET
LV
UR
HU
MK
TA
BS
BG
GL
HR
CS
DA
AR
EL
ZH
SK
RO
KO
FIL
TH
FI
NB
VI
SV
MS
TR
UK
ID
NL
CA
PL
FR
JA
RU
DE
PT
EN
IT
ES

# SLT (Fleurs)


BLEU
Hours of translated audio
$r^2 = 0.24$
0
5
10
15
20
25
30
35
40
1
10
100
1K
10K
100K
PT
CA
SV
DE FR
DA
SR
NB AF
RU
RO
SK
HR
UK
BS
GL
CS
ID
MK
BG
IT
MS
TR
ES
FIL
EL
NL
PL
AR
HE
FI
SL
HU
VI
HI
KO
OC
ET
ZH
ML
UR
JA
PA
FA
LB
LV
HY
GU
TH
TG
MT
KN
LT
BE
MR
AZ
NE
TE
BN
CY
MI
TA
IS
LO
SD
SW
UZ
KK
KM
AS
KA
PS
SN
HA
LN
SO
AM
MN
MY
YO

# Aggregate Multilingual ASR

| Model | MLS | VoxPopuli |
|---|---|---|
| Supervised Baseline | - | 37.5 |
| VP-10K + FT | - | 15.3 |
| XLS-R (1B) | 10.9 | 10.6 |
| mSLAM-CTC (2B) | 9.7 | **9.1** |
| Zero-Shot Whisper | **8.1** | 15.2 |

MLS = Multilingual Librispeech

# Average SLT (Covost 2)

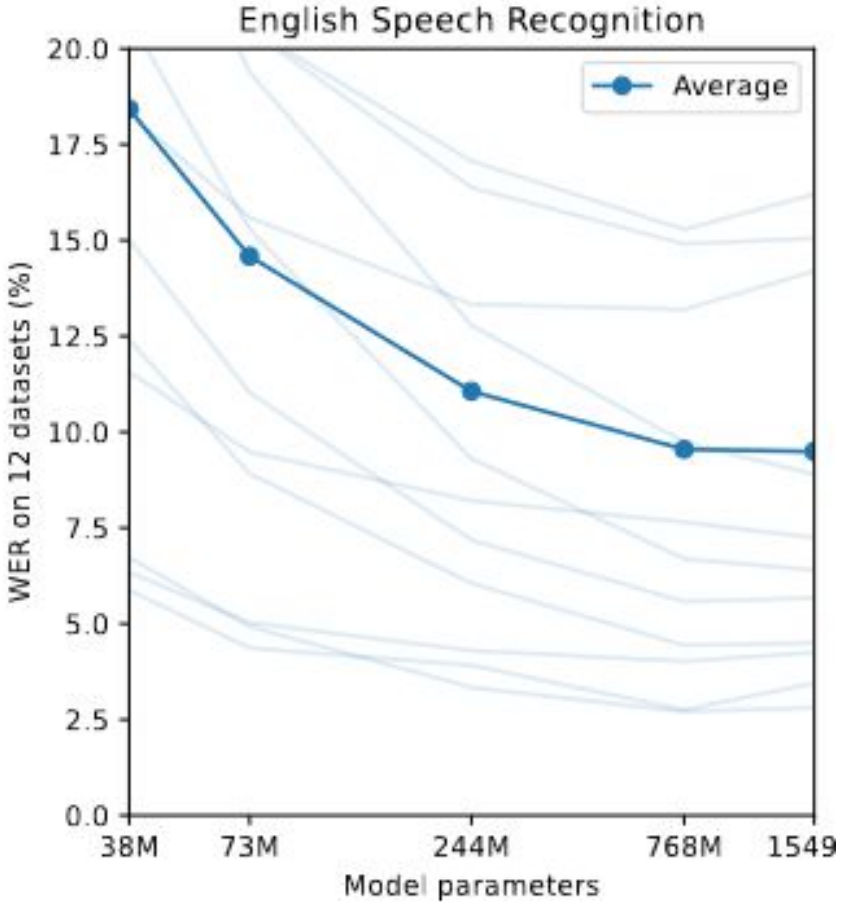
Average BLEU score performance, broken down by High / Mid / Low resource test settings

| X → English | High | Mid | Low | All |
|---|---|---|---|---|
| XMEF-X | 34.2 | 20.2 | 5.9 | 14.7 |
| XLS-R (2B) | 36.1 | 27.7 | 15.1 | 22.1 |
| mSLAM-CTC (2B) | **37.8** | 29.6 | 18.5 | 24.8 |
| Zero-Shot Whisper | 35.0 | **31.1** | **23.1** | **27.3** |

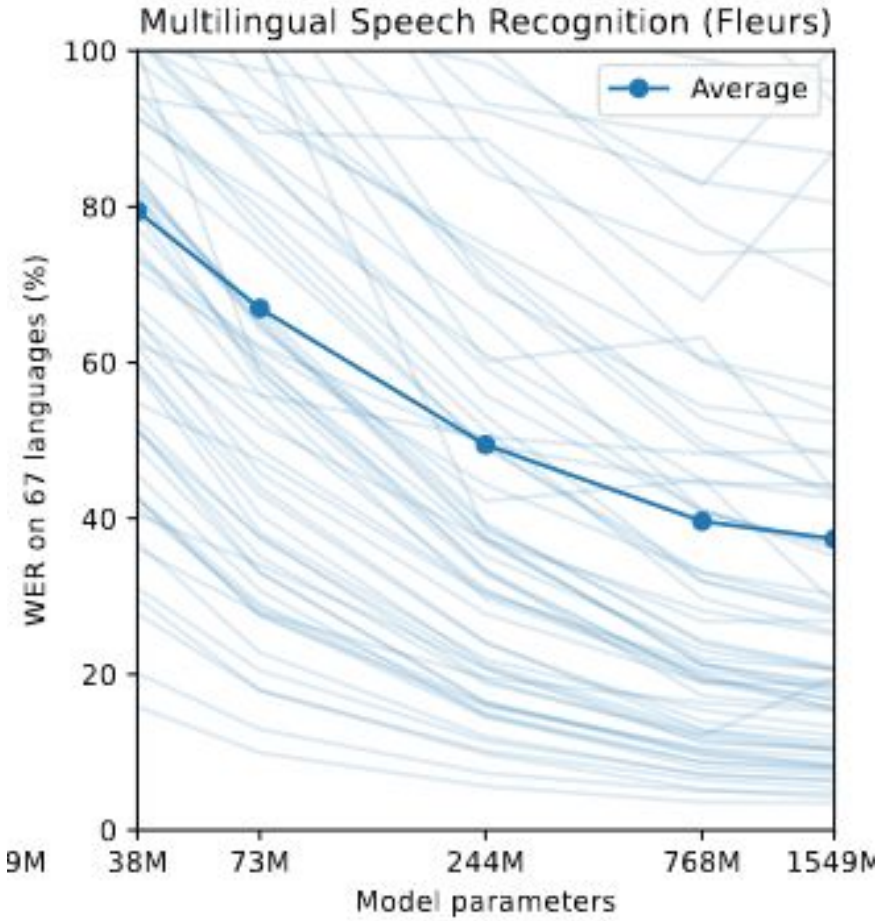
# Scaling (w/ dataset training hours)

| Dataset size | English WER (↓) | Multilingual WER (↓) | X→En BLEU (↑) |
|---|---|---|---|
| 3405 | 30.5 | 92.4 | 0.2 |
| 6811 | 19.6 | 72.7 | 1.7 |
| 13621 | 14.4 | 56.6 | 7.9 |
| 27243 | 12.3 | 45.0 | 13.9 |
| 54486 | 10.9 | 36.4 | 19.2 |
| 681070 | **9.9** | **29.2** | **24.8** |

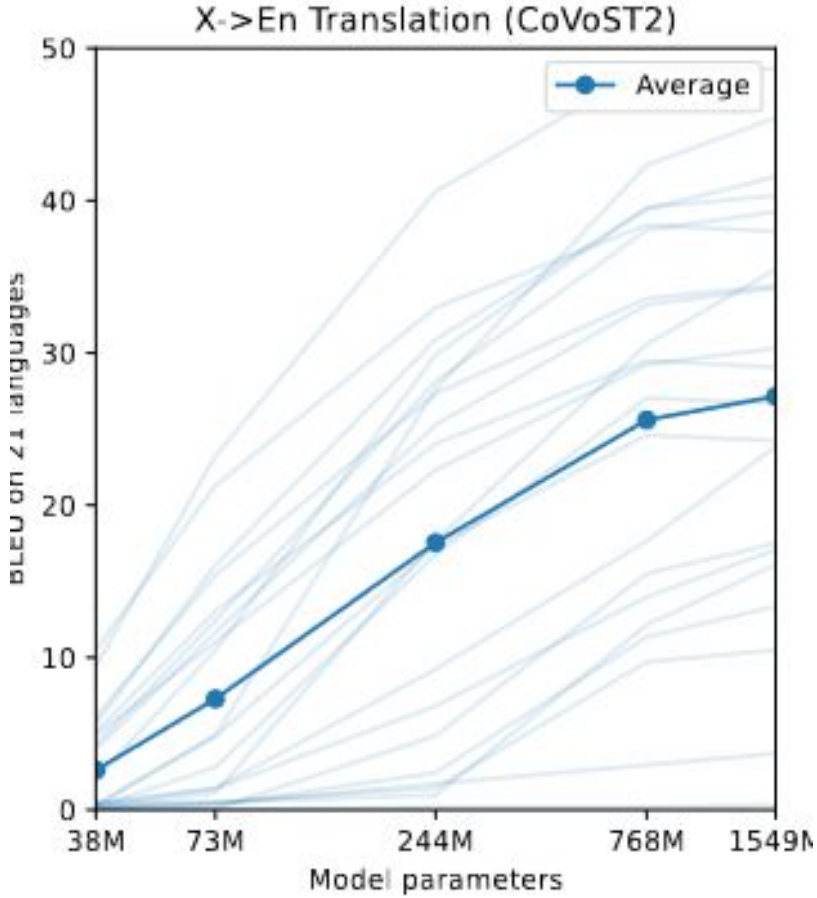
# Scaling cont.



# Scaling cont.



# Scaling cont.



# Scaling cont.

Weird